Enhancing Cross-modal Completion and Alignment for Unsupervised Incomplete Text-to-Image Person Retrieval

Tiantian Gong[1] , Junsheng Wang[2] , Liyan Zhang[1]*
[1]Nanjing University of Aeronautics and Astronautics
[2]Nanjing University of Science and Technology
{tiantian_gong, zhangliyan}@nuaa.edu.cn, jswang@njust.edu.cn

Abstract

Traditional text-image person retrieval methods heavily rely on fully matched and identity-annotated multimodal data, representing an ideal yet limited scenario. The issues of handling incomplete multimodal data and the complexities of labeling multimodal data are common challenges encountered in real-world applications. In response to these challenges encountered, we consider a more robust and pragmatic setting termed unsupervised incomplete text-image person retrieval, where person images and text descriptions are not fully matched and lack the supervision of identity labels. To tackle these two problems, we propose the **Enhancing Cross-modal Completion and Alignment** (ECCA) method. Specifically, we propose a feature-level cross-modal completion strategy for incomplete data. This approach leverages the available cross-modal high semantic similarity features to construct relational graphs for missing modal data, which can generate more reliable completion features. Additionally, to address the cross-modal matching ambiguity, we propose weighted inter-instance granularity alignment as well as enhanced prototype-wise granularity alignment modules that can map semantically similar image-text pairs more compact in the common embedding space. Extensive experiments on public datasets, fully demonstrate the consistent superiority of our method over SOTA text-image person retrieval methods.

1 Introduction

The goal of the person re-identification (ReID) task is to match images of individuals who share the same person identity across various camera viewpoints. Categorized based on the query object's data type, ReID can be segmented into three primary categories: image-based ReID [Sun *et al.*, 2018; Xuan and Zhang, 2021; Yu *et al.*, 2019; Gong *et al.*, 2023a], text-based person search [Liu *et al.*, 2019; Chen *et al.*, 2018; Jing *et al.*, 2020b; Gong *et al.*, 2023b] and video-based ReID [Hou *et al.*, 2021; Bai *et al.*, 2022; Hou *et al.*, 2019]. Text-based person ReID is a cross-modal fine-grained retrieval task, and its objective is to explore the fine-grained information shared between visual and linguistic domains while simultaneously establishing their tighter fine-grained alignment. In recent years, numerous effective text-image representation learning methods [Ding *et al.*, 2021; Shao *et al.*, 2022] have made remarkable advancements. These studies adhere to a similar scheme: 1) They employ the cross-modal alignment loss to align visual and textual representations into a shared embedding space. 2) Text-based person ReID models are trained on fully matched and labeled image-text pairs. These approaches heavily depend on completely matched and labeled image-text pairs, as depicted in Figure 1 (a). Indeed, this assumption is idealistic and constrained by an array of inevitable practical factors, e.g., privacy protection [Zhang *et al.*, 2022; Dou *et al.*, 2022;

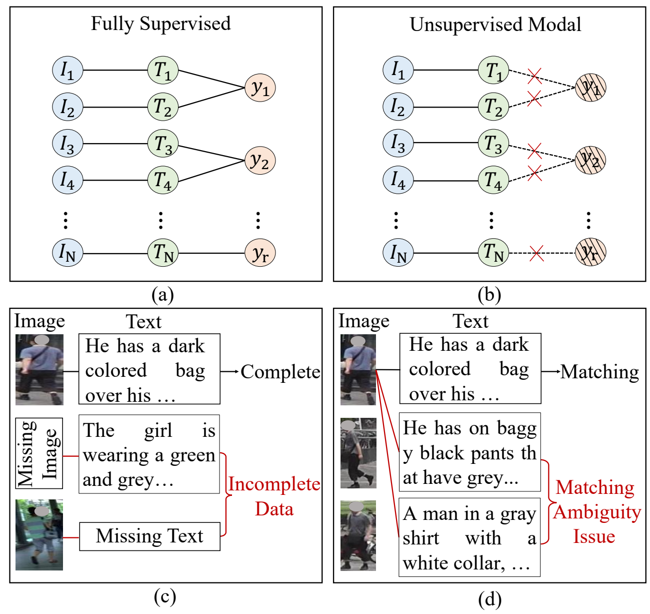

Figure 1: (a) Conventional text-image person retrieval methods. (b) Unsupervised text-image person retrieval. (c) Incomplete multi-modal data problem is a common challenge in real-world applications. (d) Unsupervised text-image person retrieval approaches encounter cross-modal matching ambiguity.

*Corresponding Author

Tang *et al.*, 2016; Li *et al.*, 2018], data missing [Xiang *et al.*, 2023], and data corruption [Xian *et al.*, 2023]. Therefore, previous approaches that relied on complete and labeled modality data to construct ranking loss for exploring text and image alignment do not perform effectively in these scenarios. As depicted in Figures 1 (b) and (c), real-world applications frequently confront challenges involving unlabeled and incomplete multimodal data. In this work, we first propose a more robust and practical setting referred to as unsupervised incomplete text-based person ReID, where person images and text descriptions are not fully matched and lack the supervision of identity labels during the training phase.

Certainly, unsupervised incomplete text-image ReID must address two key challenges: (1) How to effectively handle incomplete multimodal training data? (2) How to establish alignment between distinct fine-grained features across images and texts in the absence of true label supervision, and devise cross-modal alignment loss functions? To tackle the aforementioned concerns, we propose a novel Enhancing Cross-modal Completion and Alignment (ECCA) method, as shown in Figure 2, which comprises four key modules: cross-modal nearest neighbors construction with high semantic similarity, cross-modal feature-level completion for missing data, weight inter-instance granularity alignment, and enhanced prototype-wise granularity alignment. Specifically, we propose the high semantic similarity neighbor generation method, in which a new Jaccard distance metric is proposed to calculate the distance between two nearest neighbor samples and select the most reliably k-reciprocal nearest neighbors from cross-modality and self-modality. Relational graphs for missing features are then built using the nearest neighbors with high semantic similarity to the incomplete modality data that is reconstructed by weighting the neighbors. In addition, to address the challenge of cross-modal matching ambiguity as shown in Figure 1 (d), we propose weighted inter-instance granularity alignment as well as enhanced prototype-wise granularity alignment modules that can encourage the model to map semantically similar image-text pairs more compact in the common embedding space.

Our key contributions can be summarized three-fold: **(1)** We pioneer a new unsupervised incomplete text-image ReID task, aiming to improve the robustness and generalization of text-based ReID. **(2)** We put forward cross-modal neighbor construction with high semantic similarity and feature-level missing modality completion modeling to achieve reliable missing modal feature completion. **(3)** We propose the weighted inter-instance granularity alignment and enhanced prototype-wise granularity alignment modules, which can reduce the effect of cross-modal matching ambiguities.

2 Related Work

2.1 Text-Based Person Re-Identification

The existing text-based person ReID methods can essentially be classified into two categories: cross-modal interaction-based and cross-modal interaction-free methods. The former [Niu *et al.*, 2020; Gao *et al.*, 2021; Ding *et al.*, 2021; Wang *et al.*, 2020] mainly utilizes various attention schemes to establish word-patch [Ding *et al.*, 2021; Chen *et al.*, 2018; Li *et al.*, 2017a; Li *et al.*, 2017b] or phrase-region [Jing *et al.*, 2020b; Niu *et al.*, 2020] multi-granularity alignment relations and predict the matching score for image-text pairs. The latter [Gao *et al.*, 2021; Niu *et al.*, 2020] primarily focuses on learning global features without interactive attention mechanisms for global alignment. Such methods often employ different model structures and optimizing functions [Zhang and Lu, 2018] to align the image and text embeddings in a shared latent feature space. Recently, some works applied image and text modal pre-training of CLIP [Li *et al.*, 2022; Shao *et al.*, 2022] and achieved significant improvement.

2.2 Unsupervised Text-Image Retrieval

Research on unsupervised text-based person ReID tasks is scarce. There are only a few studies on unsupervised image-text cross-modal retrieval. Patel et al. [Patel *et al.*, 2019] propose an unsupervised cross-modal retrieval framework that leverages a latent Dirichlet allocation topic modeling framework to supervise the training of deep CNN. Liu et al. [Liu *et al.*, 2022] propose an unsupervised deep cross-modal method that exploits unsupervised contrastive learning to model the relationship among intra- and inter-modality instances. Different from general unsupervised text-image retrieval, the text-based ReID task explores more fine-grained cross-modal semantic alignment. Therefore, we utilize text-IoU guided weights to facilitate cross-modal instance discriminate learning, and leverage unified prototypes to predict soft prototype assignments to minimize intra-class variations and maximize inter-class variations between different modalities.

2.3 Incomplete Cross-modal Retrieval

There is currently no work on the unsupervised incomplete text-based ReID research. Most related to our unsupervised incomplete text-based person ReID task is the traditional incomplete image-text retrieval task. Guo et al. [Guo and Zhu, 2019] propose a collective affinity learning method (CLAM) to recover the missing adjacency information. Jiang et al. [Jing *et al.*, 2020a] exploit the dual-aligned variational autoencoders (DAVAE) to generate completion features. Zeng et al. [Zeng *et al.*, 2021] investigate a prototype-based adaptive network (PAN) to reconstruct the completion samples by prototype propagation scheme. Our unsupervised incomplete text-based ReID method is fundamentally different from them in the following aspects: (1) CLAM is based on hashing to cope with partial cross-modal problems in hash space. Our method focuses more on improving accuracy in text-based ReID. (2) DAVAE and PAN are supervised incomplete or imbalanced image-text retrieval methods, which cannot effectively learn modality alignment representations without labels to generate complete representations.

3 Methodology

In the unsupervised incomplete text-based person ReID task, the fully matched training dataset is defined as $\mathcal{X} = \{(I_i, T_i)\}_{i=1}^{K_1}$, where I_i represents the i-th image instance, T_i is the i-th corresponding text description for that image I_i, and K_1 denotes the total number of fully matched image-text pairs. Incomplete multi-view data comprises missing vi-

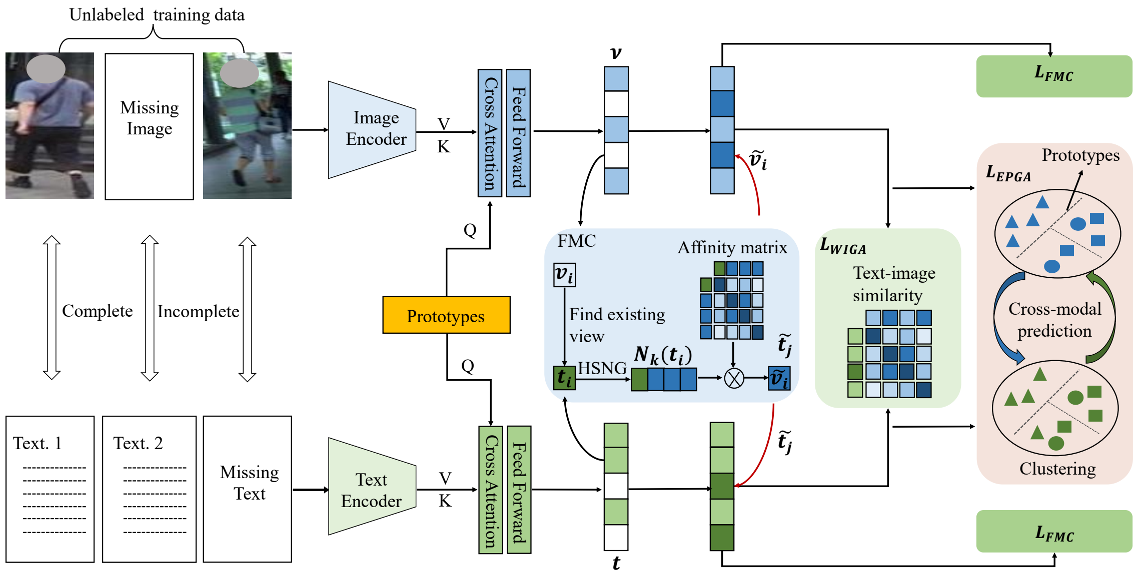

Figure 2: Illustration of the enhancing cross-modal completion and alignment (ECCA) framework for unsupervised incomplete text-image person retrieval. FMC, HSNG, WIGA and EPGA represent feature-level missing modality completion, high semantic similarity neighbor generation, weighted inter-instance granularity alignment and enhanced prototype-wise granularity alignment. The orange is shared prototypes across the image and text modalities for modal interaction and semantic alignment.

sual modality data $\mathcal{X}_t$ only with text modality data and missing text modality data $\mathcal{X}_v$ only with image modality data, where $\mathcal{X}_t = \{\tilde{I}_m, T_m\}_{m=1}^{K_2}$, $\mathcal{X}_v = \{I_n, \tilde{T}_n\}_{n=1}^{K_3}$, $\tilde{I}_m$ and $\tilde{T}_n$ is missing (unavailable, inaccessible, incomplete) data, and T_m, I_n as well as (I_i, T_i) are available (accessible) data during training. Here, K_2 and K_3 represent the total number of missing visual data and missing text data, respectively. $K_1 + K_2 + K_3 = N$ denotes the total number of samples.

3.1 Feature-level Cross-modal Completion

Firstly, we introduce the feature extraction networks for both the visual and textual modalities. For each image instance I_i and text instance T_i, the initial visual embedding z_i^v and the initial textual embedding z_i^t can be generated using the visual encoder $f^v(I_i, \theta^v)$ with trainable parameters θ^v, as well as textual encoder $f^t(T_i, \theta^t)$ with trainable parameters θ^t. To project features from distinct modalities into a joint embedding space that ensures substantial modality interaction and semantic alignment at the feature level, we adopt the shared prototypes across images and texts for local fine-grained implicit alignment. In specific terms, we define the shared prototypes as $D \in \mathbb{R}^{s\times d}$ across the image and text modalities. Here, s signifies the number of prototypes, while d denotes the dimension of features. The prototypes are randomly initialized, and the prototypes and common representations are jointly learned in the subsequent training process. For the fusion of image (text) representation, the shared prototype D serves as the query Q, while the original image (text) representation z_i^v (z_i^t) is employed as the key K and value V in the transformer's cross-attention operation. Hence, the fused visual and textual feature representations by,

$$v_i = MHA(D, z_i^v, z_i^v), \tag{1}$$

$$t_i = MHA(D, z_i^t, z_i^t), \tag{2}$$

here, v_i and t_i represent the reconstructed visual and textual contextualized features. The operation $MHA(.)$ refers to a transformer block, which comprises multi-head cross-attention and a feed-forward network [Vaswani *et al.*, 2017].

High Semantic Similarity Neighbor Generation. We propose the high semantic similarity neighbor generation (HSNG) method, in which a new Jaccard distance metric is proposed to calculate the distance between two nearest neighbor samples and selects the most reliably k-reciprocal nearest neighbors from cross-modality and self-modality. Specifically, for missing image feature $\tilde{v}_m$ shown in Figure 3, we can acquire the corresponding textual embedding feature t_m. Additionally, we calculate cross-modal cosine similarity between t_m and all existing image features $\{v_i\}_{i=1}^{K_1+K_3}$. By utilizing the k-nearest neighbor algorithm, we rank and identify the k most similar image embeddings to the textual representation t_m, denoted by,

$$N_k(t_m) = \{v_1, v_2, ..., v_k\}. \tag{3}$$

Further, for $v_l \in N_k(t_m)$, we calculate cross-modal cosine similarity between v_l and all existing text representations, and can obtain the k-nearest neighbor set for v_l as $N_k(v_l) = \{t_1, t_2, ..., t_k\}$. Accordingly, the cross-modality k-reciprocal nearest neighbors $\mathcal{R}_k(t_m)$ for t_m are formulated,

$$\mathcal{R}_k(t_m) = \{v_l | (v_l \in N_k(t_m)) \cap (t_m \in N_k(v_l))\}. \tag{4}$$

Furthermore, for $v_i, v_j \in N_k(t_m)$, we can calculate intra-modal cosine similarity between v_i, v_j and all existing image representations, and can obtain the k-nearest neighbor set for v_i and v_j defined as $N_k(v_i)$ and $N_k(v_j)$. Accordingly, we define the self-modality k-reciprocal nearest neighbor as,

$$\mathcal{R}_k(v_i) = \{v_j | (v_j \in N_k(v_i)) \cap (v_i \in N_k(v_j))\}. \quad (5)$$

Considering both cross-modality and self-modality k-reciprocal nearest neighbors, a new Jaccard distance metric is given by,

$$d(t_m, v_i) = 1 - \frac{|\mathcal{R}_k(t_m) \cap \mathcal{R}_k(v_i)|}{|\mathcal{R}_k(t_m) \cup \mathcal{R}_k(v_i)|}. \quad (6)$$

Under such constraints, we can find reliable cross-modal k-reciprocal nearest neighbors to improve the reliability of the nearest neighbor generation. Finally, we can obtain the high semantic similarity neighbor generation set formulated as,

$$N_{k'}(t_m) = \{v_1, v_2, ..., v_{k'}\}. \quad (7)$$

The same applies to the missing text features as well.

Feature-level Missing Modal Completion. To efficiently complete missing modality embeddings, we introduce a feature-level missing modality completion (FMC) method. Specifically, for missing image feature $\tilde{v}_m$ and missing text feature $\tilde{t}_n$, we can build the most relevant nearest neighbor sets of cross-modal features by the aforementioned neighbor generation method with high semantic similarity, defined as $N_{k'}(t_m) = \{v_1, v_2, ..., v_{k'}\}$ and $N_{k'}(v_n) = \{t_1, t_2, ..., t_{k'}\}$. The reconstructed visual representation $\tilde{v}_m$ of $\tilde{I}_m$ and textual representation $\tilde{t}_n$ of $\tilde{T}_n$ are formulated as,

$$\tilde{v}_m = A_v \cdot [t_m, N_{k'}(t_m)], \quad \tilde{t}_n = A_t \cdot [v_n, N_{k'}(v_n)], \quad (8)$$

where A_v denotes the affinity matrix of $[t_m, N_{k'}(t_m)] = [t_m, v_1, v_2, ..., v_{k'}] = [g_1, g_2, ..., g_{k'+1}]$, and A_t denotes the affinity matrix of $[v_n, N_{k'}(v_n)] = [v_n, t_1, t_2, ..., t_{k'}]$. Each value of the affinity matrixes A_v and A_t represents the degree of semantic similarity between two instances, formulated as,

$$A_v = Z^{-1} \cdot S, \quad (9)$$

where Z^{-1} represents the normalized Laplacian matrix of S, and each element $S_{ij} \in S$ is calculated by,

$$S_{ij} = \exp(\langle g_i, g_j \rangle), \quad (10)$$

here $\langle , \rangle$ denotes the cosine similarity between two instances. Similarly, this calculation process also applies to A_t. This approach is essentially equivalent to constructing graph relationships, enabling information to be transmitted across different samples based on the graph, thus enhancing the features of completion. Here, the affinity matrix A_v is the edges and the feature $[t_m, N_{k'}(t_m)]$ is the nodes. To mitigate the modal discrepancy between the generated representations and the original corresponding representations, the feature-level missing modality completion (FMC) is formulated as,

$$L_{FMC} = \frac{1}{K_2} \sum_{m=1}^{K_2} \|\tilde{v}_m - t_m\|_2^2 + \frac{1}{K_3} \sum_{n=1}^{K_3} \|\tilde{t}_n - v_n\|_2^2. \quad (11)$$

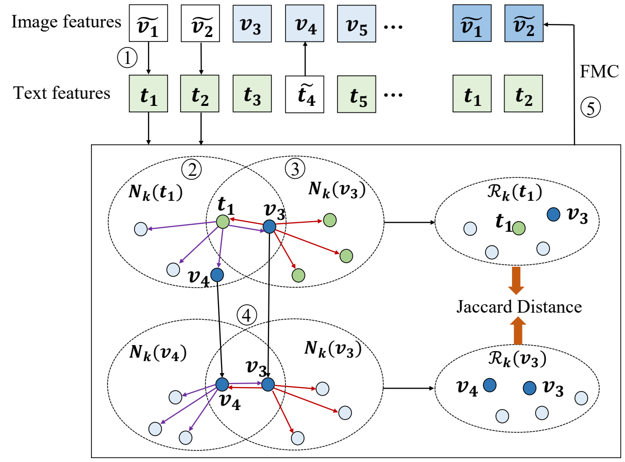

Figure 3: Illustration of high semantic similarity neighbor generation (HSNG) for missing image features.

3.2 Weighted Instance Granularity Alignment

To address the challenge of cross-modal matching ambiguity caused by the absence of true label supervision as shown in Figure 1 (d), we propose a weighted inter-instance granularity alignment module (WIGA). which adaptively applies different weights according to the matching probability between different instances, and adaptively adjusts the alignment of texts and images in the shared space. Our study is based on an empirical observation that noun phrases within two textual descriptions originating from the same pedestrian identity consistently exhibit either the same or synonymous attributes. For the provided textual description T_i, we employ NLTK [Loper and Bird, 2002] to extract relevant noun phrases from the text T_i, which are represented as,

$$P(T_i) = R_i = \{r_1, r_2, ..., r_l\}, \quad (12)$$

where P denotes the noun phrase extractor and l represents the count of noun phrases. Next, the Intersection over Union (IoU) based on the textual descriptions is defined as,

$$\text{IoU}_{i,j} = \frac{|R_i \cap R_j|}{|R_i \cup R_j|}, \quad (13)$$

here, $|R_i \cap R_j|$ represents the count of synonymous noun phrases shared between R_i and R_j. $|R_i \cup R_j|$ indicates the number of noun phrases in the union between R_i and R_j. The matching probability weights between different instances can be obtained by,

$$W_{i,j} = \frac{\text{IoU}_{i,j}}{\sum_{k=1}^{N} \text{IoU}_{i,k}}. \quad (14)$$

The WIGA dynamically adjusts the alignment of different instances by adding different similarity weights as,

$$L_{WIGA}^{i2t} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha I_{i,j} + (1-\alpha) W_{i,j}) L(v_i, t_j), \quad (15)$$

$$L_{WIGA}^{t2i} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha I_{i,j} + (1-\alpha) W_{i,j}) L(t_i, v_j), \quad (16)$$

$$L(v_i, t_j) = -log\frac{exp(\langle v_i, t_j\rangle/\tau)}{\sum_{k=1}^{N} exp(\langle v_i, t_k\rangle/\tau)}, \tag{17}$$

$$L(t_i, v_j) = -log\frac{exp(\langle t_i, v_j\rangle/\tau)}{\sum_{k=1}^{N} exp(\langle t_i, v_k\rangle/\tau)}, \tag{18}$$

where $\langle,\rangle$ represents the cosine similarity, and τ represents an temperature factor. N is the total number of image-text pairs, and $\alpha \in [0,1]$ is the prior probability that image v_i is matched with its paired text t_j. When $\alpha = 1$, we should use the one-hot labels I_{ij} for contrastive learning. However, to better align unpaired text feature t_j with image feature v_i, αI_{ij} provides supervision for paired image-text samples, while $(1-\alpha)W_{ij}$ supervises the unpaired samples. The overall objective of our WIGA loss is computed as,

$$L_{WIGA} = L_{WIGA}^{i2t} + L_{WIGA}^{t2i}. \tag{19}$$

3.3 Enhanced Prototype-wise Alignment

Besides, we propose the enhanced prototype-wise granularity alignment (EPGA) module that can achieve a more effective alignment of global visual and textual embeddings by utilizing cross-modal unified prototypes for both image and text modality as intermediaries. To begin, we establish trainable unified prototypes for both image and text modalities, denoted as $C = \{c_1, c_2, ..., c_K\}$, where K signifies the count of trainable prototype vectors. More precisely, for each image-text embedding pair (v_i, t_i), we assign v_i and t_i to K unified prototypes in C, and obtain two soft prototype assignment codes $q_{v,i} \in \mathbb{R}^K$ and $q_{t,i} \in \mathbb{R}^K$ by using the Sinkhorn Knopp algorithm [Cuturi, 2013]. Following this, we compute the visual and textual softmax probabilities, $p_{v,i} \in \mathbb{R}^K$ and $p_{t,i} \in \mathbb{R}^K$, respectively. Here $p_{v,i}$ and $p_{t,i}$ can be acquired by applying the softmax function to the cosine similarities between v_i and all cross-modal unified prototypes in C, as well as between t_i and all cross-modal unified prototypes in C as,

$$p_{v,i}^k = \frac{exp(v_i^\top c_k/\tau)}{\sum_{k'} exp(v_i^\top c_{k'}/\tau)}, \tag{20}$$

$$p_{t,i}^k = \frac{exp(t_i^\top c_k/\tau)}{\sum_{k'} exp(t_i^\top c_{k'}/\tau)}, \tag{21}$$

where τ denotes a cluster-level temperature factor, and k represents the k-th vector in unified prototypes C. EPGA can be achieved by optimizing the cross-entropy loss as,

$$L(v_i, q_{t,i}) = -\sum_k q_{t,i}^k \mathrm{log} p_{v,i}^k, \tag{22}$$

$$L(t_i, q_{v,i}) = -\sum_k q_{v,i}^k \mathrm{log} p_{t,i}^k, \tag{23}$$

where the EPGA is executed by utilizing the soft text prototype assignment $q_{t,i}$ as the "pseudo-label" for training the visual embedding v_i, while the soft image assignment $q_{v,i}$ is employed as the "pseudo-label" for training the textual embedding t_i. The overall L_{EPGA} objective is calculated by,

$$L_{EPGA} = \frac{1}{2N}\sum_{i=1}^{N}(L(v_i, q_{t,i}) + L(t_i, q_{v,i})). \tag{24}$$

By employing the losses L_{WIGA}, L_{EPGA} and L_{FMC}, our model is trained by minimizing the loss as,

$$L = L_{WIGA} + L_{EPGA} + L_{FMC}. \tag{25}$$

4 Experiments

4.1 Experimental Setup

Datasets. CUHK-PEDES [Li *et al.*, 2017b] comprises 40,206 pedestrian images along with 80,412 text descriptions corresponding to 13,003 distinct pedestrian identities. Each individual image is accompanied by a minimum of two corresponding text descriptions. The training set includes 34,054 images, 68,108 textual descriptions, and 11,003 person identities. The test set contains 3,074 images and 6,156 textual descriptions, with 1,000 distinct person identities. **ICFG-PEDES** [Ding *et al.*, 2021] comprises 54,522 images with 4,102 distinct identities. Each person's image includes a corresponding textual description. The training set encompasses 34674 image-text pairs for 3102 different person identities. The test set consists of 19,848 image-text pairs.

Challenging Data Partitions. We define three distinct settings to represent varying levels of difficulty. For the easy setting, we use 50% of the training set as the complete image-text pair data, 25% as missing image data, and 25% as missing text data, denoted as (50%, 25%, 25%). Similarly, we establish the medium setting, defined as (30%, 35%, 35%), and the hard setting as (10%, 45%, 45%) to elevate the training complexity. We employ Rank-k (where k = 1, 5, 10), a commonly used metric in text-image person retrieval.

Implementation Details. In our experiments, we adopt the image encoder and text encoder components of the Clip [Radford *et al.*, 2021] model to serve as the feature extractors. During training, image data augmentation is applied through the incorporation of random horizontal flipping, random cropping, and random erasing techniques. All images are resized to 384 × 128 pixels. For the text modality, the maximum length of text tokens is set to 80. The model is optimized via the Adam optimizer [Kingma and Ba, 2014] with a 0.0001 learning ratio. The batch size is set to 64, and the training process spans across a total of 60 epochs. The temperature parameter τ (Equations 19 and 24) is set to 0.02.

4.2 Comparison with State-of-the-Art Methods

Comparisons on Incomplete Modal Data. We initially evaluate the proposed ECCA method on the widely-used CUHK-PEDES and ICFG-PEDES using unsupervised incomplete modal data. As shown in Tables 1 and 2, our ECCA outperforms SOTA text-image person retrieval approaches in three distinct settings, including unsupervised IRRA and supervised AXM-Net, LGUR, SSAN, ViTAA, SCAN, MIA and CMPM/C methods. More specifically, our ECCA improves the unsupervised IRRA method (same feature extractor as ours) by 1.43%, 4.61% and 5.96% Rank-1 accuracy on CUHK-PEDES, by 2.53%, 3.68% and 4.47% Rank-1 accuracy on ICFG-PEDES under three different settings, respectively. It can be observed that these methods suffer significant performance degradation when encountering incomplete data. Therefore, the performance improvements on the easy setting are not as prominent as on the hard setting. Under the hard setting, our method achieves 56.38% and 42.08% Rank-1 accuracy on CUHK-PEDES and ICFG-PEDES, which fully demonstrates that our method can effectively deal with incomplete data and improve the robustness of the model.

Methods	ID	Easy Setting			Medium Setting			Hard Setting		
		Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
CMPM/C[Zhang and Lu, 2018]	✓	40.79	63.01	74.58	39.82	62.54	74.84	26.96	52.48	63.88
MIA [Niu *et al.*, 2020]	✓	46.23	68.56	77.64	43.64	66.37	74.11	29.78	54.94	65.71
SCAN [Lee *et al.*, 2018]	✓	49.84	71.96	79.38	46.87	69.43	77.64	31.83	53.77	64.93
ViTAA [Wang *et al.*, 2020]	✓	49.32	70.67	79.51	47.24	69.56	78.79	31.48	54.96	65.02
SSAN [Ding *et al.*, 2021]	✓	53.41	74.34	82.31	49.05	71.76	79.73	34.04	57.74	68.35
AXM-Net [Farooq *et al.*, 2022]	✓	57.28	77.18	84.11	53.23	74.24	81.97	36.64	59.98	69.74
LGUR [Shao *et al.*, 2022]	✓	58.77	78.36	85.41	53.95	74.79	81.77	35.61	59.36	69.18
IRRA [Jiang and Ye, 2023]	✗	63.80	83.28	89.27	59.09	79.50	86.59	50.42	73.76	81.61
ECCA	✗	**65.23**	**85.14**	**91.29**	**63.70**	**83.11**	**89.84**	**56.38**	**77.24**	**85.07**

Table 1: Performance comparisons under three different settings on the CUHK-PEDES benchmark dataset.

Methods	ID	Easy Setting			Medium Setting			Hard Setting		
		Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
CMPM/C [Zhang and Lu, 2018]	✓	34.69	55.86	65.71	29.73	51.17	60.84	15.62	30.26	41.59
MIA [Niu *et al.*, 2020]	✓	39.13	60.34	69.16	36.17	58.52	67.83	19.54	35.44	45.95
SCAN [Lee *et al.*, 2018]	✓	42.53	65.41	71.82	41.37	63.95	71.89	22.24	40.27	51.86
ViTAA [Wang *et al.*, 2020]	✓	43.79	65.86	73.42	42.78	64.55	72.41	22.61	40.62	51.37
SSAN [Ding *et al.*, 2021]	✓	46.27	67.06	75.52	44.86	66.72	74.58	24.83	42.64	53.19
AXM-Net [Farooq *et al.*, 2022]	✓	50.31	70.62	77.93	48.26	67.14	76.69	28.26	49.53	59.49
LGUR [Shao *et al.*, 2022]	✓	52.73	70.55	78.41	48.32	68.73	76.91	29.73	51.26	60.19
IRRA [Jiang and Ye, 2023]	✗	51.65	71.66	78.98	47.49	68.56	76.48	37.61	59.22	68.09
ECCA	✗	**54.18**	**74.34**	**81.10**	**51.17**	**71.04**	**78.32**	**42.08**	**62.95**	**73.16**

Table 2: Performance comparisons under three different settings on the ICFG-PEDES benchmark dataset.

Comparisons on Complete Modal Data. To further verify the more robust advantage of our ECCA model for fine-grained cross-modal semantic alignment, we compare our model with several SOTA text-image person retrieval methods. In Table 3, for CUHK-PEDES with full modality data, our model surpasses unsupervised CMMT, MM-TIM and fully supervised AXM-Net, LGUR, CAIBC, IVT, TextReID and SSAN methods, and achieves 68.13% on Rank-1, 87.26% on Rank-5 and 91.88% on Rank-10. These experimental results fully demonstrate that our method can be applied to more realistic scenarios, such as the lack of identity labels.

Methods	ID	Rank-1	Rank-5	Rank-10
SSAN [Ding *et al.*, 2021]	✓	61.37	80.15	86.73
TextReID [Han *et al.*, 2021]	✓	64.08	81.73	88.19
IVT [Shu *et al.*, 2022]	✓	64.00	82.72	88.95
CAIBC [Wang *et al.*, 2022]	✓	64.43	82.87	88.37
LGUR [Shao *et al.*, 2022]	✓	64.21	81.94	87.93
AXM-Net [Farooq *et al.*, 2022]	✓	64.44	80.52	86.77
IRRA [Jiang and Ye, 2023]	✓	73.38	89.93	93.71
MM-TIM [Gomez *et al.*, 2019]	✗	45.35	63.78	70.63
CMMT [Zhao *et al.*, 2021]	✗	57.10	78.14	85.23
ECCA (our)	✗	**68.13**	**87.26**	**91.88**

Table 3: Performance comparisons on CUHK-PEDES benchmark under the full multimodal data.

4.3 Ablation Studies

Analysis of Feature-level Missing Modality Completion (FMC). As illustrated in Table 4, to verify the effectiveness of our feature-level missing modality completion (FMC), our method is trained under the medium setting on four distinct training sets, including 1) only complete modal data, 2) missing visual data, 3) missing textual data and 4) complete all missing data. It can be observed from these experimental results that the accuracy of Rank-1, 5, 10 using only the complete data is the worst. As we reconstruct incomplete visual data or incomplete text data, experimental performance is gradually improved. The experimental performance reaches a maximum until all incomplete data are completed. The ECCA with complete all missing data improves ECCA with only complete modal data by 5.23% Rank-1 accuracy and 4.25% Rank-5 accuracy on CUHK-PEDES dataset under the medium setting. These results demonstrate that our feature-level missing modality completion can reduce the impact of performance degradation caused by incomplete data.

Ablations on High Semantic Similarity Neighbor Generation. In Table 5, to verify the effectiveness of our feature-level missing modality completion (FMC) with high semantic similarity neighbor generation (HSNG) in our ECCA, we conduct ablation experiments on CUHK-PEDES dataset under the easy setting. WIGA + EPGA represents weighted inter-instance granularity alignment and enhanced prototype-wise granularity alignment. WIGA + EPGA adding HSNG

Training Set Setting	Rank-1	Rank-5	Rank-10
only complete-modal data	58.47	78.86	86.51
dataset w/o image-modal data	60.65	80.21	87.36
dataset w/o text-modal data	61.03	80.84	87.75
the full dataset	**63.70**	**83.11**	**89.84**

Table 4: Performance comparisons with different training sets on CUHK-PEDES dataset under the Medium setting.

WIGA+EPGA	HSNG	FMC	Rank-1	Rank-5	Rank-10
✓	✓		61.68	81.91	87.21
✓		✓	62.14	82.56	87.63
✓	✓	✓	**65.23**	**85.14**	**91.29**

Table 5: Ablation Study: Performance comparisons for HSNG on CUHK-PEDES dataset under the Easy setting.

Loss functions	Rank-1	Rank-5	Rank-10
CMPM	59.17	79.38	86.51
Ranking loss	61.76	80.94	87.02
InfoNCE	62.39	81.26	87.71
WIGA (Our)	66.45	85.73	90.01
WIGA+EPGA (Our)	**68.13**	**87.26**	**91.88**

Table 6: Performance comparisons of different losses on CUHK-PEDES dataset under the full multimodal data.

achieve 61.68 % on Rank-1 and 81.91 % on Rank-5, and WIGA + EPGA adding FMC with general nearest neighbor completion achieve 62.14% on Rank-1 and 82.56% on Rank-5. However, our WIGA + EPGA adding FMC with HSNG consistently surpasses both cases. This verifies that the FMC with HSNG can effectively recover the missing features, and fully mine the side information of the missing data.

Performance comparisons of different losses. To demonstrate the effectiveness of our proposed weighted inter-instance granularity alignment (WIGA) and enhanced prototype-wise granularity alignment (EPGA), as shown in Table 6, we compare experimental results of the commonly used CMPM loss [Zhang and Lu, 2018], Ranking loss [Faghri *et al.*, 2017], InfoNC loss [Oord *et al.*, 2018] with our proposed WIGA and EPGA loss on CUHK-PEDES dataset under the full multimodal data. Specifically, our WIGA + EPGA achieves 68.13% on Rank-1, 87.26 % on Rank-5 and 91.88% on Rank-10, and consistently outperforms CMPM loss, Ranking loss, and InfoNC loss by 8.96%, 6.37% and 5.74 % Rank-1 accuracy, respectively. This is because our WIGA and EPGA effectively handle the cross-modal matching ambiguity caused by the absence of true label supervision, which encourages the model to map semantically similar image-text pairs more compactly. This fully validates that our WIGA and EPGA can achieve tighter fine-grained cross-modal semantic alignment in unsupervised scenarios.

4.4 Parameter Analysis

In this section, we conduct hyperparameter analysis experiments on the CUHK-PEDES dataset for the temperature τ, the number of the shared prototypes s in D of Equations (1) and (2) under the full data, the mutual neighbor values k and $k^{'}$ in Equations (5) and (7) under the easy settings. We transform the values of τ, s, k and $k^{'}$ in a certain range, report the corresponding Rank-1 and mAP values, respectively, and then perform experimental analysis. As illustrated in Figure 4, a) experimental results show that as the value of τ rises, the Rank-1 and mAP accuracy initially increase and then decrease more. Our method achieves a more stable performance when τ is set to 0.02. b) When s is approximately equal to 400, the model reaches the optimal value, which indicates that the number of prototypes is sufficient for learning shared features between images and texts. c, d) The peak performance is achieved when $k = 6$ and $k^{'} = 4$ on the CUHK-PEDES dataset. This trend underscores that excessively large values of k and $k^{'}$ will increase the probability of false neighbors belonging to distinct person identities, leading to a reduction on Rank-1 and mAP.

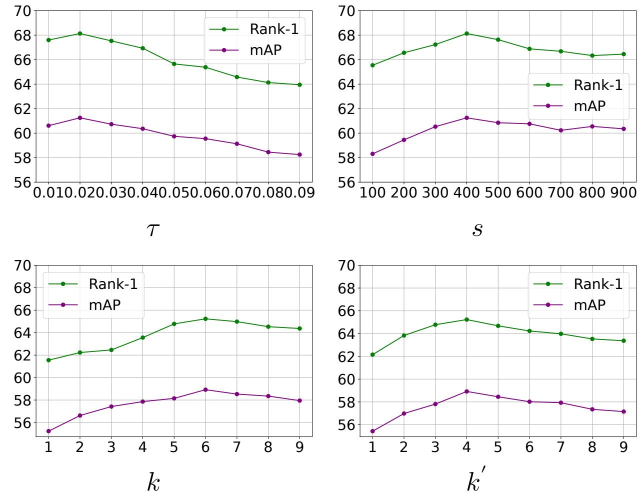

Figure 4: Parameter analysis with different values of temperature τ, number of prototypes s under full data and nearest neighbors k and $k^{'}$ under the Easy setting on CUHK-PEDES.

5 Conclusions

In this paper, we propose a novel enhancing cross-modal completion and alignment (ECCA) framework for unsupervised incomplete text-image person retrieval task. Specifically, we introduce a feature-level cross-modal completion technique tailored for incomplete data. In addition, we achieve a tighter semantic fine-grained alignment between images and texts by integrating weighted inter-instance granularity alignment and enhanced prototype-wise granularity alignment. Extensive experimental results on public datasets fully demonstrate the effectiveness of our method in the face of substantial missing data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62172212, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20230031.

References

[Bai *et al.*, 2022] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7339–7348, 2022.

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

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

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

[Dou *et al.*, 2022] Shuguang Dou, Xinyang Jiang, Qingsong Zhao, Dongsheng Li, and Cairong Zhao. Towards privacy-preserving person re-identification via person identify shift. *arXiv preprint arXiv:2207.07311*, 2022.

[Faghri *et al.*, 2017] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[Farooq *et al.*, 2022] Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4477–4485, 2022.

[Gao *et al.*, 2021] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*, 2021.

[Gomez *et al.*, 2019] Raul Gomez, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas. Self-supervised learning from web data for multimodal retrieval. In *Multimodal Scene Understanding*, pages 279–306. Elsevier, 2019.

[Gong *et al.*, 2023a] Tiantian Gong, Kaixiang Chen, Liyan Zhang, and Junsheng Wang. Debiased contrastive curriculum learning for progressive generalizable person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[Gong *et al.*, 2023b] Tiantian Gong, Guodong Du, Junsheng Wang, Yongkang Ding, and Liyan Zhang. Prototype-guided cross-modal completion and alignment for incomplete text-based person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5253–5261, 2023.

[Guo and Zhu, 2019] Jun Guo and Wenwu Zhu. Collective affinity learning for partial cross-modal hashing. *IEEE Transactions on Image Processing*, 29:1344–1355, 2019.

[Han *et al.*, 2021] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*, 2021.

[Hou *et al.*, 2019] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7183–7192, 2019.

[Hou *et al.*, 2021] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2014–2023, 2021.

[Jiang and Ye, 2023] Ding Jiang and Mang Ye. 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*, pages 2787–2797, 2023.

[Jing *et al.*, 2020a] Mengmeng Jing, Jingjing Li, Lei Zhu, Ke Lu, Yang Yang, and Zi Huang. Incomplete cross-modal retrieval with dual-aligned variational autoencoders. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3283–3291, 2020.

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

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.

[Li *et al.*, 2017a] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1890–1899, 2017.

[Li *et al.*, 2017b] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979, 2017.

[Li *et al.*, 2018] Zechao Li, Jinhui Tang, and Tao Mei. Deep collaborative embedding for social image understanding. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2070–2083, 2018.

[Li *et al.*, 2022] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2724–2728. IEEE, 2022.

[Liu *et al.*, 2019] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. Deep adversarial graph attention convolution network for text-based person search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 665–673, 2019.

[Liu *et al.*, 2022] Yaxin Liu, Jianlong Wu, Leigang Qu, Tian Gan, Jianhua Yin, and Liqiang Nie. Self-supervised correlation learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2022.

[Loper and Bird, 2002] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[Niu *et al.*, 2020] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person reidentification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020.

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

[Patel *et al.*, 2019] Yash Patel, Lluis Gomez, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised visual representations for cross-modal retrieval. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 182–186, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[Shu *et al.*, 2022] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, pages 624–641. Springer, 2022.

[Sun *et al.*, 2018] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[Tang *et al.*, 2016] Jinhui Tang, Xiangbo Shu, Guo-Jun Qi, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1662–1674, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2020] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. 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*, pages 402–420. Springer, 2020.

[Wang *et al.*, 2022] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5314–5322, 2022.

[Xian *et al.*, 2023] Yuqiao Xian, Jinrui Yang, Fufu Yu, Jun Zhang, and Xing Sun. Graph-based self-learning for robust person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4789–4798, 2023.

[Xiang *et al.*, 2023] Suncheng Xiang, Jingsheng Gao, Mengyuan Guan, Jiacheng Ruan, Chengfeng Zhou, Ting Liu, Dahong Qian, and Yuzhuo Fu. Learning robust visual-semantic embedding for generalizable person re-identification. *arXiv preprint arXiv:2304.09498*, 2023.

[Xuan and Zhang, 2021] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person reidentification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11926–11935, 2021.

[Yu *et al.*, 2019] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2148–2157, 2019.

[Zeng *et al.*, 2021] Zhixiong Zeng, Shuai Wang, Nan Xu, and Wenji Mao. Pan: Prototype-based adaptive network for robust cross-modal retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1125–1134, 2021.

[Zhang and Lu, 2018] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–701, 2018.

[Zhang *et al.*, 2022] Junwu Zhang, Mang Ye, and Yao Yang. Learnable privacy-preserving anonymization for pedestrian images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7300–7308, 2022.

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