Noisy-Correspondence Learning for Text-to-Image Person Re-identification

Yang Qin[1] Yingke Chen[2] Dezhong Peng[1,4,5] Xi Peng[1] Joey Tianyi Zhou[3] Peng Hu[1*]

[1]College of Computer Science, Sichuan University, Chengdu, 610095, China.
[2]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK.
[3]Centre for Frontier AI Research (CFAR) and Institute of High Performance Computing (IHPC), A*STAR, Singapore.
[4]Sichuan Newstrong UHD Video Technology Co., Ltd., Chengdu 610095, China.
[5]Chengdu Ruibei Yingte Information Technology Company Ltd., Chengdu 610065, China.

Abstract

Text-to-image person re-identification (TIReID) is a compelling topic in the cross-modal community, which aims to retrieve the target person based on a textual query. Although numerous TIReID methods have been proposed and achieved promising performance, they implicitly assume the training image-text pairs are correctly aligned, which is not always the case in real-world scenarios. In practice, the image-text pairs inevitably exist under-correlated or even false-correlated, a.k.a noisy correspondence (NC), due to the low quality of the images and annotation errors. To address this problem, we propose a novel Robust Dual Embedding method (RDE) that can learn robust visual-semantic associations even with NC. Specifically, RDE consists of two main components: 1) A Confident Consensus Division (CCD) module that leverages the dual-grained decisions of dual embedding modules to obtain a consensus set of clean training data, which enables the model to learn correct and reliable visual-semantic associations. 2) A Triplet Alignment Loss (TAL) relaxes the conventional Triplet Ranking loss with the hardest negative samples to a log-exponential upper bound over all negative ones, thus preventing the model collapse under NC and can also focus on hard-negative samples for promising performance. We conduct extensive experiments on three public benchmarks, namely CUHK-PEDES, ICFG-PEDES, and RSTPReID, to evaluate the performance and robustness of our RDE. Our method achieves state-of-the-art results both with and without synthetic noisy correspondences on all three datasets. Code is available at https://github.com/QinYang79/RDE.

(a) Clean correspondence (b) Noisy correspondence

Figure 1. The illustration of noisy correspondence. The figure shows an example of the NC problem, which occurs when the image-text pairs are wrongly aligned, *i.e*., false positive pairs (FPPs). Since the model does not know which pairs are noisy in practice, they will unavoidably degrade the performance by incorrect supervision information. As seen in the figure, (a) the clean image-text pair is semantically matched, while (b) the noisy pair is not, which would cause the cross-modal model to learn erroneous visual-textual associations. Note that both examples in (a) and (b) are from and actually exist in the RSTPReid dataset [50].

1. Introduction

Text-to-image person re-identification (TIReID) [21, 24, 38] aims to understand the natural language descriptions to retrieve the matched person image from a large gallery set. This task has received increasing attention from both academic and industrial communities recently, *e.g*., finding/tracking suspect/lost persons in a surveillance system [10, 40]. However, TIReID remains a challenging task due to the inherent heterogeneity gap across different modalities and appearance attribute redundancy.

To tackle these challenges, most of the existing methods explore global- and local-matching alignment to learn accurate similarity measurements for person re-identification. To be specific, some global-matching methods [38, 41, 48] leverage vision/language backbones to extract modality-specific features and employ contrastive learning to achieve global visual-semantic alignments. To capture fine-grained information, some local-matching methods [22, 29, 35, 39] explicitly align local body regions to textually described entities/objectives to improve the discriminability of pedestrian features. Recently, some works [16, 21, 42]

*Corresponding author: Peng Hu (penghu.ml@gmail.com).

propose to exploit visual/semantic knowledge learned by the pre-trained models, such as BERT [5], ViT [9], and CLIP [34], and achieve explicit global alignments or discover more fine-grained local correspondence, thus boosting the re-identification performance. Although these methods achieve remarkable progress, they implicitly assume that all training image-text pairs are aligned correctly.

In reality, this assumption is hard or even impossible to hold due to the person's pose, camera angle, illumination, and other inevitable factors in images, which may result in some inaccurate/mismatched textual descriptions of images (see Figure 1), *e.g.*, the RSTPReid dataset [50]. Moreover, we observe that excessive such imperfect/mismatched image-text pairs would cause an overfitting problem and degrade the performance of existing TIReID methods shown in Figure 5. Based on the observation, in this paper, we reveal and study a new problem in TIReID, *i.e.*, noisy correspondence (NC). Different from noisy labels, NC refers to the false correspondences of image-text pairs in TIReID, *i.e.*, False Positive Pairs (FPPs): some negative image-text pairs are used as positive ones for cross-modal learning. Inevitably, FPPs will misguide models to overfit noisy supervision and collapse to suboptimal solutions due to the memorization effect [1] of Deep Neural Networks (DNNs).

To address the NC problem, we propose a Robust Dual Embedding method (RDE) for TIReID in this paper, which benefits from an effective Confident Consensus Division mechanism (CCD) and a novel Triplet Alignment Loss (TAL). Specifically, CCD fuses the dual-grained decisions to consensually divide the training data into clean and noisy sets, thus providing more reliable correspondences for robust learning. To diversify the model grain, the basic global embeddings (BGE) and token selection embeddings (TSE) are presented for coarse-grained and fine-grained cross-modal interactions respectively, thus capturing visual-semantic associations comprehensively. Different from the widely-used Triplet Ranking loss with the hardest negatives, our TAL relaxes the similarity learning from the hardest negative samples to all negative ones by applying an upper bound, which brings a stable solution for the collapse of training under NC while also benefiting from the hardest negatives mining to achieve promising performance. As a result, our RDE can achieve robustness against NC thanks to the proposed reliable supervision and stable triplet loss. The contributions and innovations of this paper are summarized as follows:

- We reveal and study a new and ubiquitous problem in TIReID, termed noisy correspondence (NC). Different from class-level noisy labels, NC refers to erroneous correspondences in the person-description pairs that can mislead the model to learn incorrect visual-semantic associations. To the best of our knowledge, this paper could be the first work to explore this problem in TIReID.
- We propose a robust method, termed RDE, to mitigate the adverse impact of NC through the proposed Confident Consensus Division (CCD) and novel Triplet Alignment Loss (TAL). By using CCD and TAL, RDE can obtain convincing consensus pairs and reduce the misleading risks in training, thus embracing robustness against NC.
- Extensive experiments on three public image-text person benchmarks demonstrate the robustness and superiority of our method. Our method achieves the best performance both with and without synthetic noisy correspondence on all three datasets.

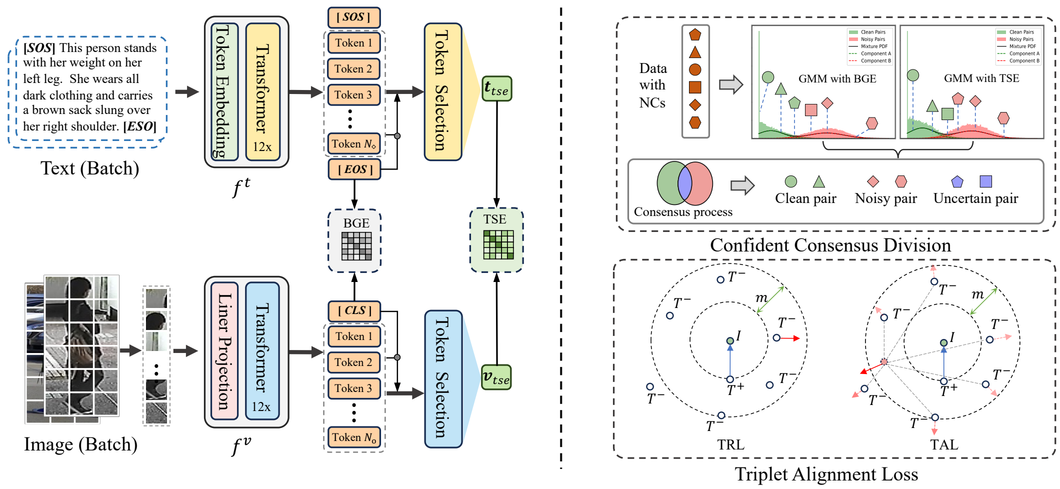
2. Related Work

2.1. Text-to-Image Person Re-identification

Text-to-image person re-identification (TIReID) is a novel and challenging task that aims to match a person image with a given natural language description [2–4, 24, 25, 28, 36, 37, 43, 48]. Existing TIReID methods could be roughly classified into two groups according to their alignment levels, *i.e.*, **global-matching** methods [38, 49, 50] and **local-matching** methods [13, 35, 39]. The former try to learn cross-modal embeddings in a common latent space by employing textual and visual backbones with a matching loss (*e.g.*, CMPM/C loss [48] and Triplet Ranking loss [11]) for TIReID. However, these methods mainly focus on global features while ignoring the fine-grained interactions between local features, which limits their performance improvement. To achieve fine-grained interactions, some of the latter methods explore explicit local alignments between body regions and textual entities for more refined alignments. However, these methods require more computational resources due to the complex local-level associations. Recently, inspired and benefited from **vision-language pre-training** models [34], some methods [16, 21, 42] expect to use the learned rich alignment knowledge of pre-trained models for local- or global-alignments. Although these methods achieve promising performance, almost all of them implicitly assume that all input training pairs are correctly aligned, which is hard to meet in practice due to the ubiquitous noise. In this paper, we address the inevitable and challenging noisy correspondence problem in TIReID.

2.2. Learning with Noisy Correspondence

As a special learning paradigm with noisy labels [12, 23, 27] in multi-modal/view community [18, 31, 31, 32, 45], the studies for noisy correspondence (NC) have recently attracted more and more attention in various tasks, *e.g.*, video-text retrieval [47], visible-infrared person re-identification [26, 44, 46], and image-text matching [20, 30], which means that the negative pairs are wrongly treated as positive ones, *i.e.*, false positive pairs (FPPs). To handle



(a) Cross-modal Embedding Model **(b) Robust Similarity Learning**

Figure 2. The overview of our RDE. (a) is the illustration of the cross-modal embedding model used in RDE, which consists of *basical global embedding* (BGE) and *token selection embedding* (TSE) modules with different granularity. By integrating them, RDE can capture coarse-grained cross-modal interactions while selecting informative local token features to encode more fine-grained representations for a more accurate similarity. (b) shows the core of RDE to achieve robust similarity learning, which consists of Confident Consensus Division (CCD) and Triplet Alignment Loss (TAL). CCD performs consensus division to obtain confident clean training data, thus avoiding misleading from noisy pairs. Unlike traditional Triplet Ranking Loss (TRL) [11], TAL exploits an upper bound to consider all negative pairs, thus embracing more stable learning.

this problem, numerous methods are proposed to learn with NC, which can be broadly categorized into **sample selection** [15, 20, 47] and **robust loss functions** [19, 30, 33, 44]. The former commonly leverage the memorization effect of DNNs [1] to gradually distinguish the noisy data, thus paying more attention to clean data while less attention to noisy data. Differently, the latter methods aim to develop noise-tolerance loss functions to improve the robustness of model training against NC. Although the aforementioned methods achieve promising performance in various tasks, they are not specifically designed for TIReID and may be inefficient or ineffective in person re-identification. In this paper, we propose a well-designed method to tackle the NC problem in TIReID, which not only performs superior in noisy scenarios but also achieves promising performance in ordinary scenarios.

3. Methodology

3.1. Problem Statement

The purpose of TIReID is to retrieve a pedestrian image from the gallery set that matches the given textual description. For clarity, we represent the gallery set as $\mathcal{V} = \{I_i, y_i^p, y_i^v\}_{i=1}^{N_v}$ and the corresponding text set as $\mathcal{T} = \{T_i, y_i^v\}_{i=1}^{N_t}$, where N_v is the number of images, N_t is the number of texts, $y_i^p \in \mathcal{Y}_p = \{1, \cdots, C\}$ is the class label (person identify), C is the number of identifies, and $y_i^v \in \mathcal{Y}_v = \{1, \cdots, N_v\}$ is the image label. The image-text pair set used in TIPeID can be defined as $\mathcal{P} = \{(I_i, T_i), y_i^v, y_i^p\}_{i=1}^{N}$, where the cross-modal samples of each pair have the same image label y_i^v and class label y_i^p. We define a binary correspondence label $l_{ij} \in \{0, 1\}$ to indicate the matched degree of any image-text pair. If $l_{ij} = 1$, the pair (I_i, T_j) is matched (positive pair), otherwise it is not (negative pair). In practice, due to ubiquitous annotation noise, some unmatched pairs $(l_{ij} = 0)$ are wrongly labeled as matched $(l_{ij} = 1)$, resulting in noisy correspondences (NCs) and performance degradation. To handle NC for robust TIReID, we present an RDE that leverages the Confident Consensus Division (CCD) and Triplet Alignment Loss (TAL) to mitigate the negative impact of label noise.

3.2. Cross-modal Embedding Model

In this section, we describe the cross-modal model used in our RDE. Following previous work [21], we utilize the visual encoder f^v and textual encoder f^t of the pre-trained model CLIP as modality-specific encoders to obtain token representations and implement cross-modal interactions through two embedding modules.

3.2.1 Token Representations

Give an input image $I_i \in \mathcal{V}$, we use the visual encoder f^v of CLIP to tokenize the image into a discrete token representation sequence with a length of $N_\circ + 1$, *i.e.*, $\boldsymbol{V}_i = f^v(I_i) = \{\boldsymbol{v}_g^i, \boldsymbol{v}_1^i, \boldsymbol{v}_2^i, \cdots, \boldsymbol{v}_{N_\circ}^i\}^\top \in \mathbb{R}^{(N_\circ+1)\times d}$, where d is the dimensionality of the shared latent space. These features include an encoded feature $\boldsymbol{v}_g^i$ of the [*CLS*] token and patch-level local features $\{\boldsymbol{v}_j^i\}_{j=1}^{N_\circ}$ of $N_\circ$ fixed-sized non-overlapping patches of I_i, wherein $\boldsymbol{v}_g^i$ can represent the global representation. For an input text $T_i \in \mathcal{T}$, we apply the textual encoder f^t of CLIP to obtain global and local representations. Specifically, following IRRA [21], we first tokenize the input text T_i using lower-cased byte pair encoding (BPE) with a 49,152 vocab size into a token sequence. The token sequence is bracketed with [*SOS*] and [*EOS*] tokens to represent the beginning and end of the sequence. Then, we feed the token sequence into f_t to obtain the features $\boldsymbol{T}_i = \{\boldsymbol{t}_s^i, \boldsymbol{t}_1^i, \cdots, \boldsymbol{t}_{N_\diamond}^i, \boldsymbol{t}_e^i\}^\top \in \mathbb{R}^{(N_\diamond+2)\times d}$, where $\boldsymbol{t}_s^i$ and $\boldsymbol{t}_e^i$ are the features of [*SOS*] and [*EOS*] tokens and $\{\boldsymbol{v}_j^i\}_{j=1}^{N_\diamond}$ are the word-level local features of $N_\diamond$ word tokens of text T_i. Generally, the $\boldsymbol{t}_e^i \in \mathbb{R}^{d\times 1}$ can be regarded as the sentence-level global feature of T_i.

3.2.2 Dual Embedding Modules

To measure the similarity between any image-text pair (I_i, T_j), we can directly use the global features of [*CLS*] and [*EOS*] tokens to compute the *Basic Global Embedding* (BGE) similarity by the cosine similarity, *i.e*., $S_{ij}^b = {\boldsymbol{v}_g^i}^\top \boldsymbol{t}_e^j / \|\boldsymbol{v}_g^i\| \|\boldsymbol{t}_e^j\|$, where the global features represent the global embedding representations of two modalities. However, optimizing the BGE similarities alone may not capture the fine-grained interactions between two modalities, which will limit performance improvement. To address this issue, we exploit the local features of informative tokens to learn more discriminative embedding representations, thus mining the fine-grained correspondences. In CLIP, the global features of the tokens ([*CLS*] and [*EOS*]) are obtained by a weighted aggregation of all local token features. These weights reflect the correlation between the global token and each local token. Following previous methods [42, 51], we could select the informative tokens based on these correlation weights to aggregate local features for a more representative global embedding.

In practice, these correlation weights can be obtained directly in the self-attention map of the last Transformer blocks of f^v and f^t, which reflects the relevance among the input $1 + N_\circ$ (or $2 + N_\diamond$) tokens. Given the output self-attention map $\boldsymbol{A}_i^v \in \mathbb{R}^{(1+N_\circ)\times(1+N_\circ)}$ of image I_i, the correlation weights between global token and local tokens are $\{a_{i,j}^v\}_{j=1}^{N_\circ} = \boldsymbol{a}_i^v = \boldsymbol{A}_i^v[0, 1 : N_\circ + 1] \in \mathbb{R}^{N_\circ}$. Similarly, for text T_i, the correlation weights are $\{a_{i,j}^t\}_{j=1}^{N_\diamond} = \boldsymbol{a}_i^t = \boldsymbol{A}_i^t[0, 1 : N_\diamond + 1] \in \mathbb{R}^{N_\diamond}$, where $\boldsymbol{A}_i^t \in \mathbb{R}^{(2+N_\diamond)\times(2+N_\diamond)}$ is the output self-attention map for text I_i. Then, we select a proportion ($TopK$) of the corresponding token features with higher scores for embedding. Specifically, for I_i, the selected token sequences and correlation weights are reorganized as $\boldsymbol{V}_i^s = \{\boldsymbol{v}_j^i\}_{j\in\boldsymbol{K}_i^v}$ and $\hat{\boldsymbol{a}}_i^v = \{a_{i,j}^v\}_{j\in\boldsymbol{K}_i^v}$, where $\boldsymbol{K}_i^v \in \mathbb{R}^{\lfloor \mathcal{R} N_\circ \rfloor}$ is the set of indices for the selected local tokens of I_i and $\mathcal{R}$ is the selection ratio. For text T_i, the selected token sequences and correlation weights are also reorganized as $\boldsymbol{T}_i^s = \{\boldsymbol{t}_j^i\}_{j\in\boldsymbol{K}_i^t}$ and $\hat{\boldsymbol{a}}_i^t = \{a_{i,j}^t\}_{j\in\boldsymbol{K}_i^t}$, where $\boldsymbol{K}_i^t \in \mathbb{R}^{\min(\lfloor \mathcal{R} N'_\diamond \rfloor, N_\diamond)}$ is the set of indices for the selected local tokens of T_i. $N'_\diamond$ is the maximum input sequence length of f^t. For I_i and T_i, we perform an embedding transformation on these selected token features to obtain subtle representations, instead of using complex fine-grained correspondence discovery used in CFine [42]. The transformation is performed by an embedding module like the residual block [17], as follows:

$$\begin{aligned} \boldsymbol{v}_{tse}^i &= MaxPool(MLP(\hat{\boldsymbol{V}}_i^s) + FC(\hat{\boldsymbol{V}}_i^s)), \\ \boldsymbol{t}_{tse}^i &= MaxPool(MLP(\hat{\boldsymbol{T}}_i^s) + FC(\hat{\boldsymbol{T}}_i^s)), \end{aligned} \tag{1}$$

where $MaxPool(\cdot)$ is the max-pooling function, $MLP(\cdot)$ is a multi-layer perceptron (MLP) layer, $FC(\cdot)$ is a linear layer, $\hat{\boldsymbol{V}}_i^s = L2Norm(\boldsymbol{V}_i^s)$, and $\hat{\boldsymbol{T}}_i^s = L2Norm(\boldsymbol{T}_i^s)$. $L2Norm(\cdot)$ is the ℓ_2-normalization function to normalize features. Finally, for any pair (I_i, T_j), we compute the cosine similarity S_{ij}^t between $\boldsymbol{v}_{tse}^i$ and $\boldsymbol{t}_{tse}^j$ as the *Token Selection Embedding* (TSE) similarity to measure the cross-modal matching degree for auxiliary training and inference.

3.3. Robust Similarity Learning

In this section, we detail how we use the image-text similarities computed by the dual embedding modules for robust TIReID, which involves Confident Consensus Division (CCD) and Triplet Alignment Loss (TAL).

3.3.1 Confident Consensus Division

To alleviate the negative impact of NC, the key is to filter the possible noisy pairs in the training data, which directly avoids false supervision information. Some previous work in learning with noisy labels [14, 20, 23] are inspired by the memorization effect [1] of DNNs to perform filtrations, *i.e*., the clean (easy) data tend to have a smaller loss value than that of noisy (hard) data in early training. Based on this, we can exploit the two-component Gaussian Mixture Model (GMM) to fit the per-sample loss distributions computed by the predictions of BGE and TSE to further identify the noisy pairs in the training data. Specifically, given a cross-modal model $\mathcal{M}$, we first define the per-sample loss as:

$$\ell(\mathcal{M}, \mathcal{P}) = \{\ell_i\}_{i=1}^N = \left\{\mathcal{L}(I_i, T_i)\right\}_{i=1}^N, \tag{2}$$

where $\mathcal{L}$ is the loss function for pair $(I_i, T_i) \in \mathcal{P}$ to bring them closer in the shared latent space. In our method, $\mathcal{L}$ is the proposed $\mathcal{L}_{tal}$ defined in Equation (5). Then, the per-sample loss is fed into the GMM to separate clean and noisy data, *i.e*., assigning the Gaussian component with a lower mean value as a clean set and the other as a noisy one, respectively. Following [20, 23], we use the Expectation-Maximization algorithm to optimize the GMM and compute the posterior probability $p(k|\ell_i) = p(k)p(\ell_i|k)/p(\ell_i)$ for the i-th pair as the probability of being clean/noisy pair, where $k \in \{0, 1\}$ is used to indicate whether it is a clean or a noisy component. Then, we set a threshold $\delta = 0.5$ to $\{p(k = 0|l_i)\}_{i=1}^{N}$ to divide the data into clean and noisy sets, *i.e.*,

$$\begin{aligned}\mathcal{P}^c =& \{(I_i, T_i)|p(k = 0|\ell_i) > \delta, \forall(I_i, T_i) \in \mathcal{P}\},\\ \mathcal{P}^n =& \{(I_i, T_i)|p(k = 0|\ell_i) \leq \delta, \forall(I_i, T_i) \in \mathcal{P}\},\end{aligned} \tag{3}$$

where $\mathcal{P}^c$ and $\mathcal{P}^n$ are the divided clean and noisy sets, respectively. For BGE and TSE, the divisions conducted with Equation (3) are $\mathcal{P} = \mathcal{P}^c_{bge} \cup \mathcal{P}^n_{bge}$ and $\mathcal{P} = \mathcal{P}^c_{tse} \cup \mathcal{P}^n_{tse}$, separately.

To obtain the final reliable divisions, we propose to exploit the consistency between the two divisions to find the consensus part as the final confident clean set, *i.e.*, $\hat{\mathcal{P}}^c = \mathcal{P}^c_{bge} \cap \mathcal{P}^c_{tse}$. The rest of the data can be divided into noisy and uncertain subsets, *i.e.*, $\hat{\mathcal{P}}^n = \mathcal{P}^n_{bge} \cap \mathcal{P}^n_{tse}$ and $\hat{\mathcal{P}}^u = \mathcal{P} - (\hat{\mathcal{P}}^c \cup \hat{\mathcal{P}}^n)$. Finally, we exploit the divisions to further recalibrate the correspondence labels, *e.g.*, for i-th pair, the process can be expressed as:

$$\hat{l}_{ii} = \begin{cases} 1, & \text{if } (I_i, T_i) \in \hat{\mathcal{P}}^c, \\ 0, & \text{if } (I_i, T_i) \in \hat{\mathcal{P}}^n, \\ Rand(\{0, 1\}), & \text{if } (I_i, T_i) \in \hat{\mathcal{P}}^u, \end{cases} \tag{4}$$

where $Rand(X)$ is the function to randomly select an element from the collection X.

3.3.2 Triplet Alignment Loss

The Triplet Ranking Loss (TRL) is a common matching loss that is widely used in cross-modal learning, and achieves promising performance by employing the hardest negatives, *e.g.*, image-text matching [6], video-text retrieval [8], *etc*. However, we find that this strategy may lead to bad local minima or even model collapse for TIReID under NC in the early stages of training. In contrast, the summation version of TRL that considers all negative samples, namely TRL-S, can maintain better stability and avoid model collapse, but suffers from insufficient performance due to the lack of attention to hard negatives (see Section 3.3.3 for more discussion). Therefore, we propose a novel Triplet Alignment Loss (TAL) to guide TIReID, which differs from TRL in that it relaxes the optimization of the hardest negatives to all negatives with an upper bound (see Lemma 1). Thanks to the relaxation, TAL reduces the risk of the optimization being dominated by the hardest negatives, thereby making the training more stable and comprehensive by considering all pairs. For an input pair (I_i, T_i) in a mini-batch $\mathbf{x}$, TAL is defined as

$$\begin{aligned}\mathcal{L}_{tal}(I_i, T_i) =& \big[m - S^+_{i2t}(I_i) + \tau \log(\sum_{j=1}^{K} q_{ij} \exp(S(I_i, T_j)/\tau))\big]_+\\ &+ \big[m - S^+_{t2i}(T_i) + \tau \log(\sum_{j=1}^{K} q_{ji} \exp(S(I_j, T_i)/\tau))\big]_+,\end{aligned} \tag{5}$$

where m is a positive margin coefficient, τ is a temperature coefficient to control hardness, $S(I_i, T_j) \in \{S^b_{ij}, S^t_{ij}\}$, $[x]_+ \equiv \max(x, 0)$, $\exp(x) \equiv e^x$, $q_{ij} = 1 - l_{ij}$, and K is the size of $\mathbf{x}$. From Lemma 1, as $\tau \to 0$, TAL approaches TRL and focuses more on hard negatives. Since multiple positive pairs from the same identity may appear in the mini-batch, $S^+_{i2t}(I_i) = \sum_{j=1}^{K} \alpha_{ij} S(I_i, T_j)$ is the weighted average similarity of positive pairs for image I_i, where $\alpha_{ij} = \frac{l_{ij} \exp(S(I_i, T_j)/\tau)}{\sum_{k=1}^{N} l_{ik} \exp(S(I_i, T_k)/\tau)}$. Similarly, $S^+_{i2t}(T_i)$ is the weighted average similarity of positive pairs for text T_i.

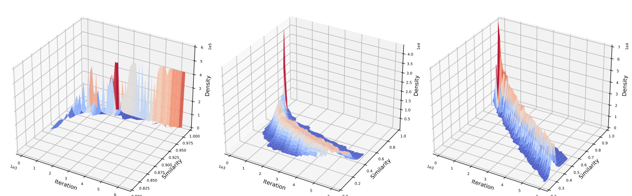
Lemma 1 *TAL is the upper bound of TRL, i.e.,*

$$\begin{aligned}\mathcal{L}_{trl}(I_i, T_i) =& \big[m - S^+_{i2t}(I_i) + S(I_i, \hat{T}_i)\big]_+\\ &+ \big[m - S^+_{t2i}(T_i) + S(\hat{I}_i, T_i)\big]_+ \leq \mathcal{L}_{tal}(I_i, T_i),\end{aligned} \tag{6}$$

where $\hat{T}_i \in \{T_j | l_{ij} = 0, \forall j \in \{1, \cdots, K\}\}$ *is the hardest negative text for* I_i *and* $\hat{I}_i \in \{I_j | l_{ji} = 0, \forall j \in \{1, \cdots, K\}\}$ *is the hardest negative image for* I_i*, respectively.*

3.3.3 Revisit Triplet Raking Loss

To explore the behaviors of the triplet losses in the noisy case, we record the similarity distributions versus iterations of TRL, TRL-S, and the proposed TAL under 50% noise. From Figure 3a, one can see that the similarities of all pairs are gradually gathered to 1 during training with TRL, *i.e.*, all samples *collapses* to a narrow neighborhood space on a hypersphere, resulting in a trivial solution and a bad performance (3.64%). To delve deeper into the underlying reason, we performed a gradient analysis. For ease of representation and analysis, we only consider one direction since



(a) TRL (3.64%) (b) TRL-S (44.93%) (c) TAL (63.35%)

Figure 3. The difference between TRL, TRL-S, and proposed TAL on the similarity distribution versus iterations. The y-z plane represents the similarity density. The corresponding Rank-1 scores of testing are placed in brackets for convenience.

image-to-text retrieval and text-to-image retrieval are symmetrical. And, we suppose that there is only one paired text for each image in the mini-batch. Due to the truncation operation $[x]_+$, we only discuss the case of $\mathcal{L} > 0$ that could generate gradients. Taking the image-to-text direction as an example, the gradients generated by TRL, TRL-S, and TAL are

$$\frac{\partial \mathcal{L}_{trl}}{\partial \boldsymbol{v}_i} = \hat{\boldsymbol{t}}_i - \boldsymbol{t}_i, \quad \frac{\partial \mathcal{L}_{trl}}{\partial \boldsymbol{t}_i} = -\boldsymbol{v}_i, \quad \frac{\partial \mathcal{L}_{trl}}{\partial \hat{\boldsymbol{t}}_i} = \boldsymbol{v}_i, \tag{7}$$

$$\frac{\partial \mathcal{L}_{trls}}{\partial \boldsymbol{v}_i} = \sum_{j\in\mathcal{Z}}(\boldsymbol{t}_j - \boldsymbol{t}_i), \frac{\partial \mathcal{L}_{trls}}{\partial \boldsymbol{t}_i} = -|\mathcal{Z}|\boldsymbol{v}_i, \frac{\partial \mathcal{L}_{trls}}{\partial \boldsymbol{t}_j} = \boldsymbol{v}_i, \tag{8}$$

$$\frac{\partial \mathcal{L}_{tal}}{\partial \boldsymbol{v}_i} = \sum_{j\neq i}^{K} \beta_j(\boldsymbol{t}_j - \boldsymbol{t}_i), \frac{\partial \mathcal{L}_{tal}}{\partial \boldsymbol{t}_i} = -\boldsymbol{v}_i, \frac{\partial \mathcal{L}_{tal}}{\partial \boldsymbol{t}_j} = \beta_j\boldsymbol{v}_i, \tag{9}$$

where $\mathcal{Z} = \{z \mid \left[m - S(I_i,T_i) + S(I_i,T_z)\right]_+ > 0, z \neq i, z \in \{0,\cdots,K\}\}$, $\beta_j = \frac{\exp(\boldsymbol{v}_i^\top \boldsymbol{t}_j/\tau)}{\sum_{k\neq i}^{K}\exp(\boldsymbol{v}_i^\top \boldsymbol{t}_k/\tau)}$, $\hat{\boldsymbol{t}}_i$, $\boldsymbol{t}_j$ and $\boldsymbol{t}_i$ are the hardest negative sample, negative sample, and positive sample of the anchor sample $\boldsymbol{v}_i$, respectively. Since the hardest sample is most similar to the positive one, $\frac{\partial_{trl}}{\partial \boldsymbol{v}_i}$ would easily approach $\mathbf{0}$ and the gradients for other negative samples except for the hardest negative one are all $\mathbf{0}$, which may lead to bad local minima early on in training and even cause the worst-case scenario, *i.e.*, model collapse (see Figure 3a). Unlike TRL, TRL-S aims to push all negative samples away from the anchor by a constant margin and produces stronger gradients for the anchor, *i.e.*, $\|\frac{\partial_{trls}}{\partial \boldsymbol{v}_i}\|_2 \geq \|\frac{\partial_{trl}}{\partial \boldsymbol{v}_i}\|_2$, thus avoiding model collapse (see Figure 3b). However, the drawback is that TRL-S treats every negative sample equally while ignoring challenging ones, which limits performance improvement. Different from TRL and TRL-S, from Equation (9), our TAL can comprehensively consider all negative samples and exploits the anchor-negative semantic relationships to adaptively adjust the gradients for each negative, thus paying more attention to hard negatives. As a result, TAL would avoid model collapse under NC while achieving superior performance (63.35% *vs.* 44.93% *vs.* 3.64%). More details for the derivations of gradients are provided in the supplementary material.

3.3.4 Training and Inference

To train the model robustly, we use the corrected label $\hat{l}_{ii}$ instead of the original correspondence label l_{ii} to compute the final matching loss, *i.e.*,

$$\mathcal{L}_m = \sum_{i=1}^{K} \hat{l}_{ii}(\mathcal{L}^b(I_i,T_i) + \mathcal{L}^t(I_i,T_i)), \tag{10}$$

where $\mathcal{L}^b(I_i,T_i)$ and $\mathcal{L}^t(I_i,T_i)$ are the TAL losses computed by Equation (5) with BGE and TSE similarities, respectively. The training process of RDE is shown in Algorithm 1. For the joint inference, we compute the final similarity of the image-text pair as the average of the similarities computed by both embedding modules, *i.e.*, $S = (S^b + S^t)/2$.

Algorithm 1 The training process of our RDE

Input: The training data $\mathcal{P}$ with N image-text pairs, maximal epoch N_e, the cross-modal model $\mathcal{M}(\Theta)$, and the hyper-parameters $\mathcal{R}, m, \tau$;

```
Initialize the backbones with the weights of the pre-trained CLIP except for the TSE module, which is randomly initialized;
for e = 1, 2, ..., N_e do
    Calculate the per-sample loss ℓ(M, P);
    Divide the training data with the predictions of BGE and TSE using Equation (3), respectively;
    Obtain the consensus divisions to recalibrate the correspondence labels {l̂_ii}_{i=1}^N with Equation (4);
    for x in mini-batches {x_m}_{m=1}^M do
        Extract the BGE and TSE features of x;
        Compute the similarities between K image-text pairs in x with above features;
        Calculate the final matching loss L_m with Equation (10);
        Θ = Optimizer (Θ, L_m);
    end for
end for
```

Output: The optimized parameters $\hat{\Theta}$.

4. Experiments

In this section, we conduct extensive experiments to verify the effectiveness and superiority of the proposed RDE on three widely-used benchmark datasets.

4.1. Datasets and Settings

4.1.1 Datasets

In the experiments, we use the CHUK-PEDES [24], ICFG-PEDES [7], and RSTPReid [50] datasets to evaluate our RDE. We follow the data partitions used in IRRA [21] to split the datasets into training, validation, and test sets, wherein the ICFG-PEDES dataset only has training and validation sets. More details are provided in the supplementary material.

4.1.2 Evaluation Protocols

For all experiments, we mainly employ the popular Rank-K metrics (K=1,5,10) to measure the retrieval performance. In addition to Rank-K, we also adopt the mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP) as auxiliary retrieval metrics to further evaluate performance following [21].

Noise	Methods		CUHK-PEDES R-1	CUHK-PEDES R-5	CUHK-PEDES R-10	CUHK-PEDES mAP	CUHK-PEDES mINP	ICFG-PEDES R-1	ICFG-PEDES R-5	ICFG-PEDES R-10	ICFG-PEDES mAP	ICFG-PEDES mINP	RSTPReid R-1	RSTPReid R-5	RSTPReid R-10	RSTPReid mAP	RSTPReid mINP
0%	SSAN	Best	61.37	80.15	86.73	-	-	54.23	72.63	79.53	-	-	43.50	67.80	77.15	-	-
	IVT	Best	65.59	83.11	89.21	-	-	56.04	73.60	80.22	-	-	46.70	70.00	78.80	-	-
	CFine	Best	69.57	85.93	91.15	-	-	60.83	76.55	82.42	-	-	50.55	72.50	81.60	-	-
	IRRA	Best	<u>73.38</u>	<u>89.93</u>	<u>93.71</u>	<u>66.13</u>	<u>50.24</u>	<u>63.46</u>	<u>80.25</u>	<u>85.82</u>	<u>38.06</u>	**7.93**	<u>60.20</u>	<u>81.30</u>	<u>88.20</u>	<u>47.17</u>	<u>25.28</u>
	RDE	Best	**75.94**	**90.14**	**94.12**	**67.56**	**51.44**	**67.68**	**82.47**	**87.36**	**40.06**	<u>7.87</u>	**65.35**	**83.95**	**89.90**	**50.88**	**28.08**
20%	SSAN	Best	46.52	68.36	77.42	42.49	28.13	40.57	62.58	71.53	20.93	2.22	35.10	60.00	71.45	28.90	12.08
		Last	45.76	67.98	76.28	40.05	24.12	40.28	62.68	71.53	20.98	2.25	33.45	58.15	69.60	26.46	10.08
	IVT	Best	58.59	78.51	85.61	57.19	45.78	50.21	69.14	76.18	34.72	<u>8.77</u>	43.65	66.50	75.70	37.22	20.47
		Last	57.67	78.04	85.02	56.17	44.42	48.70	67.42	75.06	34.44	**9.25**	37.95	63.35	73.75	34.24	19.67
	IRRA	Best	69.74	87.09	92.20	62.28	45.84	60.76	78.26	84.01	35.87	6.80	58.75	81.90	88.25	46.38	24.78
		Last	69.44	87.09	92.04	62.16	45.70	60.58	78.14	84.20	35.92	6.91	54.00	77.15	85.55	43.20	22.53
	CLIP-C	Best	66.41	85.15	90.89	59.36	43.02	55.25	74.76	81.32	31.09	4.94	54.45	77.80	86.70	42.58	21.38
		Last	66.10	86.01	91.02	59.77	43.57	55.17	74.58	81.46	31.12	4.97	53.20	76.25	85.40	41.95	21.95
	DECL	Best	70.29	87.04	91.93	62.84	46.54	61.95	<u>78.36</u>	83.88	36.08	6.25	61.75	80.70	86.90	47.70	26.07
		Last	70.08	87.20	92.14	<u>62.86</u>	46.63	61.95	<u>78.36</u>	83.88	36.08	6.25	60.85	80.45	86.65	47.34	25.86
	RDE	Best	<u>74.46</u>	**89.42**	**93.63**	**66.13**	**49.66**	**66.54**	**81.70**	<u>86.70</u>	<u>39.08</u>	7.55	**64.45**	<u>83.50</u>	**90.00**	<u>49.78</u>	<u>27.43</u>
		Last	**74.53**	<u>89.23</u>	<u>93.55</u>	**66.13**	<u>49.63</u>	<u>66.51</u>	**81.70**	**86.71**	**39.09**	7.56	<u>63.85</u>	**83.85**	<u>89.45</u>	**50.27**	**27.75**
50%	SSAN	Best	13.43	31.74	41.89	14.12	6.91	18.83	37.70	47.43	9.83	1.01	19.40	39.25	50.95	15.95	6.13
		Last	11.31	28.07	37.90	10.57	3.46	17.06	37.18	47.85	6.58	0.39	14.10	33.95	46.55	11.88	4.04
	IVT	Best	50.49	71.82	79.81	48.85	36.60	43.03	61.48	69.56	28.86	6.11	39.70	63.80	73.95	34.35	18.56
		Last	42.02	65.04	73.72	40.49	27.89	36.57	54.83	62.91	24.30	5.08	28.55	52.05	62.70	26.82	13.97
	IRRA	Best	62.41	82.23	88.40	55.52	38.48	52.53	71.99	79.41	29.05	4.43	56.65	78.40	86.55	42.41	21.05
		Last	42.79	64.31	72.58	36.76	21.11	39.22	60.52	69.26	19.44	1.98	31.15	55.40	65.45	23.96	9.67
	CLIP-C	Best	64.02	83.66	89.38	57.33	40.90	51.60	71.89	79.31	28.76	4.33	53.45	76.80	85.50	41.43	21.17
		Last	63.97	83.74	89.54	57.35	40.88	51.49	71.99	79.32	28.77	4.37	52.35	76.35	85.25	40.64	20.45
	DECL	Best	65.22	83.72	89.28	57.94	41.39	<u>57.50</u>	75.09	<u>81.24</u>	<u>32.64</u>	<u>5.27</u>	<u>56.75</u>	<u>80.55</u>	<u>87.65</u>	<u>44.53</u>	23.61
		Last	65.09	83.58	89.26	57.89	41.35	57.49	<u>75.10</u>	81.23	32.63	5.26	55.00	80.50	86.50	43.81	23.31
	RDE	Best	**71.33**	**87.41**	**91.81**	<u>63.50</u>	<u>47.36</u>	**63.76**	**79.53**	**84.91**	**37.38**	**6.80**	**62.85**	**83.20**	**89.15**	**47.67**	**23.97**
		Last	<u>71.25</u>	<u>87.39</u>	<u>91.76</u>	**63.59**	**47.50**	**63.76**	**79.53**	**84.91**	**37.38**	**6.80**	**62.85**	**83.20**	**89.15**	**47.67**	<u>23.96</u>

Table 1. Performance comparison under different noise rates on three benchmarks. "Best" means choosing the best checkpoint on the validation set to test, and "Last" stands for choosing the checkpoint after the last training epoch to conduct inference. R-1,5,10 is an abbreviation for Rank-1,5,10 (%) accuracy. The best and second-best results are in **bold** and <u>underline</u>, respectively.

4.1.3 Implementation Details

As mentioned earlier, we adopt the pre-trained model CLIP [34] as our modality-specific encoders. In fairness, we use the same version of CLIP-ViTB/16 as IRRA [21] to conduct experiments. During training, we introduce data augmentations to increase the diversity of the training data. Specifically, we utilize random horizontal flipping, random crop with padding, and random erasing to augment the training images. For training texts, we employ random masking, replacement, and removal for the word tokens as the data augmentation. Moreover, the input size of images is 384×128 and the maximum length of input word tokens is set to 77. We employ the Adam optimizer to train our model for 60 epochs with a cosine learning rate decay strategy. The initial learning rate is $1e-5$ for the original model parameters of CLIP and the initial one for the network parameters of TSE is initialized to $1e-3$. The batch size is 64. Following IRRA [21], we adopt an early training process with a gradually increasing learning rate. For hyperparameter settings, the margin value m of TAL is set to 0.1, the temperature parameter τ is set to 0.015, and the selection ratio $\mathcal{R}$ is 0.3.

4.2. Comparison with State-of-the-Art Methods

In this section, we evaluate the performance of our RDE on three benchmarks under different scenarios. For a comprehensive comparison, we compare our method with several state-of-the-art methods, including both ordinary methods and robust methods. Moreover, we use two synthetic noise levels (*i.e.*, noise rates), 20%, and 50%, to simulate the real-world scenario where the image-text pairs are not well-aligned. We randomly shuffle the text descriptions to inject NCs into the training data. We compare our RDE with five state-of-the-art baselines: SSAN [7], IVT [38], IRRA [21], DECL [30], and CLIP-C. SSAN, IVT, and IRRA are recent ordinary methods that are not designed for NC. DECL is a general framework that can enhance the robustness of image-text matching methods against NC. We use the model of IRRA as the base model of DECL for TIReID. CLIP-C is a strong baseline that fine-tunes the CLIP(ViT-B/16) model with only clean image-text pairs. We report the results of both the best checkpoint on the validation set and the last checkpoint to show the overfitting degree. Furthermore, we also evaluate our RDE on the original datasets without synthetic NC to demonstrate its superiority in Table 1. We compare our RDE with two local-matching methods: SSAN [7] and CFine [42]); and two global-matching methods: IVT [38] and IRRA [21]. More comparisons with other methods are provided in the supplementary material.

From Table 1, one can see that our RDE achieves state-of-the-art performance on three datasets and we can draw three observations: (1) On the datasets with synthetic NC, the ordinary methods suffer from remarkable performance degradation or poor performance as the noise rate increases. In contrast, our RDE achieves the best results on all metrics. Moreover, by comparing the 'Best' performance with the 'Last' ones in Table 1, we can see that our RDE can effectively prevent the performance deterioration caused by overfitting against NC. (2) Compared with the robust framework DECL and the strong baseline CLIP-C, our RDE also shows obvious advantages, which indicates that our solution against NC is effective and superior in TIReID. For instance, on CUHK-PEDES under 50% noise, our RDE achieves 71.33%, 87.41%, and 91.81% in terms of Rank-1,5,10 on the 'Best' rows, respectively, which surpasses the best baseline DECL by a large margin, *i.e.*, +6.11%, +3.69%, and +2.53%, respectively. (3) On the datasets without synthetic NC, our RDE outperforms all baselines by a large margin. Specifically, RDE achieves performance gains of +2.56%, +4.22%, and +5.15% in terms of Rank-1 compared with the best baseline IRRA on three datasets, respectively, demonstrating the effectiveness and advantages of our method.

4.3. Ablation Study

In this section, we conduct ablation studies on the CUHK-PEDES dataset with 50% noise to investigate the effects and contributions of each proposed component in RDE. We compare different combinations of our components in Table 2. From the experimental results, we could draw the following observation: (1) RDE achieves the best performance by using both BGE and TSE for joint inference, which demonstrates that these two modules are complementary and effective. (2) RDE benefits from CCD, which can enhance the robustness and alleviate the overfitting effect caused by NC. (3) Our TAL outperforms the widely-used Triplet Ranking Loss (TRL) and SDM loss [21], which demonstrates the superior stability and robustness of our TAL against NC.

No.	S^b	S^e	CCD	Loss	R-1	R-5	R-10	mAP	mINP
#1	✓	✓	✓	TAL	**71.33**	**87.41**	**91.81**	**63.50**	**47.36**
#2	✓	✓	✓	TRL	6.40	16.08	22.14	6.53	2.51
#3	✓	✓	✓	TRL-S	67.38	85.35	90.64	60.04	43.60
#4	✓	✓	✓	SDM	69.33	86.99	91.68	61.99	45.34
#5		✓	✓	TAL	70.70	86.60	91.16	62.67	46.19
#6	✓		✓	TAL	69.07	86.09	91.13	61.69	45.40
#7	✓	✓		TAL	63.11	81.04	87.22	55.42	38.68

Table 2. Ablation studies on the CHUK-PEDES dataset.

4.4. Parametric Analysis

To study the impact of different hyperparameter settings on performance, we perform sensitivity analyses for two key hyperparameters (*i.e.*, m and τ) on the CHUK-PEDES dataset with 50% noise. From Figure 4, we can see that: (1) Too large or too small m will lead to suboptimal performance. We choose $m = 0.1$ in all our experiments. (2) Too small τ will cause training failure, while the increasing τ will gradually decrease the separability (hardness) of positive and negative pairs for suboptimal performance. We choose $\tau = 0.015$ in all our experiments.

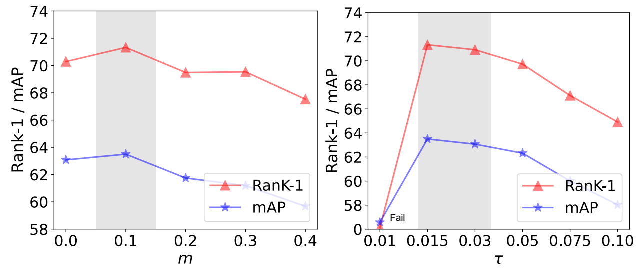
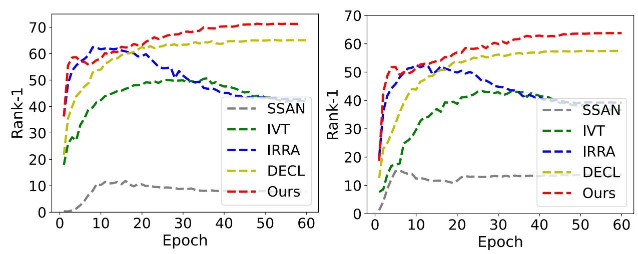


Figure 4. Variation of performance with different m and τ.

4.5. Robustness Study

In this section, we provide some visualization results during cross-modal training to verify the robustness and effectiveness of our method. As shown in Figure 5, one can clearly see that our RDE not only achieves excellent performance under noise but also effectively alleviates noise overfitting.



(a) CUHK-PEDES (b) ICFG-PEDES

Figure 5. Test performance (Rank-1) versus epochs on the CHUK-PEDES and ICFG-PEDES datasets with 50% noise.

5. Conclusion

In this paper, we reveal and study a novel challenging problem of noisy correspondence (NC) in TIReID, which violates the common assumption of existing methods that image-text data is perfectly aligned. To this end, we propose a robust method, *i.e.*, RDE, to effectively handle the revealed NC problem and achieve superior performance. Extensive experiments are conducted on three datasets to comprehensively demonstrate the superiority and robustness of RDE both with and without synthetic NCs.

Acknowledgments

This work was supported in part by NSFC under Grant U21B2040, 62176171, 62372315, and 62102274, in part by Sichuan Science and Technology Program under Grant 2022YFH0021 and 2023ZYD0143; in part by Chengdu Science and Technology Project under Grant 2023-XT00-00004-GX; in part by the SCU-LuZhou Sciences and Technology Coorperation Program under Grant 2023CDLZ-16; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303 and YJ202140.

References

[1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 233–242. PMLR, 2017. 2, 3, 4

[2] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: relation and sensitivity aware representation learning for text-based person search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 555–563, 2023. 2

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

[4] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. *arXiv preprint arXiv:2308.10045*, 2023. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[6] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1218–1226, 2021. 5

[7] 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. 6, 7

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021. 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. *Advances in neural information processing systems*, 32, 2019. 1

[11] 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. 2, 3

[12] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Rono: Robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11610–11619, 2023. 2

[13] 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. 2

[14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 4

[15] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7517–7526, 2023. 3

[16] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*, 2021. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[18] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5403–5413, 2021. 2

[19] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2023. 3

[20] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 2, 3, 4, 5

[21] 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. 1, 2, 3, 4, 6, 7, 8

[22] 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*, pages 11189–11196, 2020. 1

[23] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 2, 4, 5

[24] 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. 1, 2, 6

[25] Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. Dcel: Deep cross-modal evidential learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6292–6300, 2023. 2

[26] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity correspondence learning from

long-term noisy videos. *arXiv preprint arXiv:2401.16702*, 2024. 2

[27] Yangdi Lu, Yang Bo, and Wenbo He. An ensemble model for combating label noise. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 608–617, 2022. 2

[28] Yiwei Ma, Xiaoshuai Sun, Jiayi Ji, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. Beat: Bi-directional one-to-many embedding alignment for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4157–4168, 2023. 2

[29] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020. 1

[30] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 2, 3, 7

[31] Yalan Qin, Nan Pu, and Hanzhou Wu. Edmc: Efficient multi-view clustering via cluster and instance space learning. *IEEE Transactions on Multimedia*, 2023. 2

[32] Yalan Qin, Nan Pu, and Hanzhou Wu. Elastic multi-view subspace clustering with pairwise and high-order correlations. *IEEE Transactions on Knowledge and Data Engineering*, 2023. 2

[33] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[34] 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. 2, 7

[35] 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*, 2022. 1, 2

[36] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184, 2023. 2

[37] Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8922–8931, 2023. 2

[38] 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. 1, 2, 7

[39] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. Text-based person search via multi-granularity embedding learning. In *IJCAI*, pages 1068–1074, 2021. 1, 2

[40] Zheng Wang, Ruimin Hu, Yi Yu, Chao Liang, and Wenxin Huang. Multi-level fusion for person re-identification with incomplete marks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1267–1270, 2015. 1

[41] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633, 2021. 1

[42] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *arXiv preprint arXiv:2210.10276*, 2022. 1, 2, 4, 7

[43] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6202–6211, 2023. 2

[44] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022. 2, 3

[45] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2022. 2

[46] Mouxing Yang, Zhenyu Huang, and Xi Peng. Robust object re-identification with coupled noisy labels. *International Journal of Computer Vision*, pages 1–19, 2024. 2

[47] Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. Robust video-text retrieval via noisy pair calibration. *IEEE Transactions on Multimedia*, 2023. 2, 3

[48] 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. 1, 2

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

[50] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217, 2021. 1, 2, 6

[51] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4702, 2022. 4