Weakly Supervised Text-based Person Re-Identification

Shizhen Zhao[1], Changxin Gao[1*], Yuanjie Shao[1], Wei-Shi Zheng[2], Nong Sang[1]

[1]Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence
and Automation, Huazhong University of Science and Technology

[2]Sun Yat-sen University

Email: xbrainzsz@gmail.com, cgao@hust.edu.cn

Abstract

The conventional text-based person re-identification methods heavily rely on identity annotations. However, this labeling process is costly and time-consuming. In this paper, we consider a more practical setting called weakly supervised text-based person re-identification, where only the text-image pairs are available without the requirement of annotating identities during the training phase. To this end, we propose a Cross-Modal Mutual Training (CMMT) framework. Specifically, to alleviate the intra-class variations, a clustering method is utilized to generate pseudo labels for both visual and textual instances. To further refine the clustering results, CMMT provides a Mutual Pseudo Label Refinement module, which leverages the clustering results in one modality to refine that in the other modality constrained by the text-image pairwise relationship. Meanwhile, CMMT introduces a Text-IoU Guided Cross-Modal Projection Matching loss to resolve the cross-modal matching ambiguity problem. A Text-IoU Guided Hard Sample Mining method is also proposed for learning discriminative textual-visual joint embeddings. We conduct extensive experiments to demonstrate the effectiveness of the proposed CMMT, and the results show that CMMT performs favorably against existing text-based person re-identification methods. Our code will be available at https://github.com/X-BrainLab/WS_Text-ReID.

1. Introduction

Text-based Person Re-Identification (Re-ID) [20] is a challenging task that aims to retrieve the corresponding person images by textual descriptions. In recent years, numerous fully-supervised textual-visual embedding methods [30, 15, 25, 23, 18] have made great progress. These methods follow a similar learning scheme: 1) The identity loss is utilized to suppress the intra-class variations in each
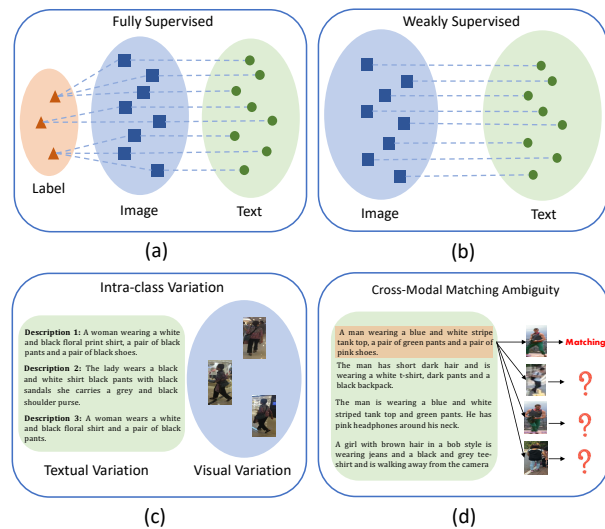
---

*Corresponding author

Figure 1. Illustration of (a) fully supervised text-based person re-identification, (b) our proposed weakly supervised text-based person re-identification, (c) intra-class variation in both textual and visual modalities and (d) cross-modal matching ambiguity.

modality; 2) The cross-modal matching is supervised by automatically generated positive or negative labels, based on whether it originates from the same identity or not. It can be observed that they heavily rely on the identity annotations, as shown in Figure 1(a). However, the identity labeling process across multiple non-overlapping camera views is costly and time-consuming. In this work, we consider a more practical setting called weakly supervised text-based person Re-ID, where *only text-image pairs are available without any identity annotations in the training phase,* as illustrated in Figure 1(b).

There are two main challenges due to the absence of identity annotations. 1) It is difficult to mitigate the effect caused by the intra-class variations in both textual and visual modalities. As shown in Figure 1(c), descriptions of a person may be syntactically different. Meanwhile, images captured by different cameras are always visually affected

by dramatic variations with respect to illumination, human pose, view angle, background, *etc*. Existing clustering-based methods [10, 9, 22] can resolve this problem to some extent by unsupervised representation learning with pseudo labels in each modality. However, different from unsupervised person Re-ID, the relationship between textual and visual modalities can be further leveraged to refine the clustering results. 2) As shown in Figure 1(d), it leads to a cross-modal matching ambiguity problem that, for one textual description, it is unable to assign positive or negative labels to all the images, except its paired one, when learning the cross-modal matching.

To address the aforementioned problems, we propose a Cross-Modal Mutual Training (CMMT) framework to facilitate visual-textual representation learning for weakly supervised text-based person Re-ID. First, in order to reduce the intra-class variations, a clustering method is leveraged to generate preliminary pseudo labels for both textual and visual instances. To further improve the clustering results, CMMT provides a Mutual Pseudo Label Refinement (MPLR) module, which utilizes the clustering results in one modality to refine that in the other modality constrained by the pairwise relationship between textual and visual modalities. Second, to mitigate the cross-modal matching ambiguity, CMMT employs a Text-IoU Guided Cross-Modal Projection Matching (Text-IoU CMPM) loss, which introduces a heuristic metric termed Intersection over Union of Text (Text-IoU) to assign similarity soft-labels. Furthermore, a Text-IoU guided Hard Sample Mining (Text-IoU HSM) method is presented to learn discriminative visual-textual joint embeddings by exploring the similarity consistency of embedding features and textual phrases.

Our **contributions** are listed as follows:

1. To the best of our knowledge, it is the first work that particularly addresses the problem of weakly supervised text-based person Re-ID.
2. The Mutual Pseudo Label Refinement module is proposed for pseudo label refinement to suppress intra-class variations by leveraging the pairwise relationship between textual and visual modalities.
3. We introduce Text-IoU, which measures the similarities of textual descriptions in phrase-level by treating the phrases as multi-labels. Text-IoU is further utilized to prompt cross-modal matching and hard sample mining.
4. Extensive experiments and component studies are conducted to demonstrate the superiority of the proposed approach for weakly supervised text-based person Re-ID. Experiment results show that, without any identity supervision, the proposed method even outperforms the state-of-the-art fully supervised text-based person Re-ID methods.

2. Related Work

2.1. Text-Based Person Re-Identification

Different from the typical person Re-ID problem [35, 4, 36], text-based person Re-ID [20, 19, 37, 3, 34, 30, 32, 8, 1, 16, 15, 2, 26, 7] aims to identify the target person by free-form natural language. For example, Li *et al.* [20] and Chen *et al.* [3] calculate image-word affinity to explore the local-level relation between visual and textual spaces. Zheng *et al.* [37] and Zhang *et al.* [34] focus on the cross-modal objective function for learning joint embeddings. Chen *et al.* [2] learn the local association through a noun phrase reconstruction method using image regions. Sarafianos *et al.* [25] introduce an adversarial cross-modal learning framework to obscure the modality information. Recent studies [30, 15] employ auxiliary algorithms (*i.e.* pose estimation, human parsing) to disentangle feature space of a person into multiple spaces, which corresponds to different human parts. All aforementioned are fully supervised methods, where identity labeling is costly and time-consuming.

2.2. Unsupervised Person Re-Identification

Unsupervised Person Re-ID [38, 29, 21, 6, 10, 9, 22] focuses on learning discriminative features without identity annotations in target datasets. There are three main categories of methods: style transfer, clustering-based pseudo label estimation, soft label learning. The most related works are the clustering-based pseudo label estimation methods. For example, Song *et al.* [29] iteratively assign pseudo labels for unlabeled data based on an encoder, which are further utilized for training the encoder. Ge *et al.* [10] present a self-paced contrastive learning framework with hybrid memory. Different from unsupervised person Re-ID, the training set consisting of text-image pairs is given in weakly supervised text-based person Re-ID. Therefore, considering the pairwise relationship between textual and visual modalities, the proposed MPLR utilizes the clustering results in one modality to refine that in the other modality.

2.3. Weakly Supervised Text-Image Retrieval

There are only a few studies [27, 11] about the weakly supervised text-image retrieval problem. Patel *et al.* [27] leverage entire text articles and image captions to supervise the textual-visual embeddings in both the local and global levels. Gomez *et al.* [11] extract feature embeddings from images and the paired captions, which then are utilized to learn textual-visual joint embeddings. Different from the general text-image retrieval problem, identity information is critical to learn the identity-specific feature embedding for text-based person Re-ID. Therefore, CMMT leverages the pseudo labels for self-training in each modality and utilizes the Text-IoU score as similarity soft-labels to facilitate the cross-modal matching learning and the hard sample mining.
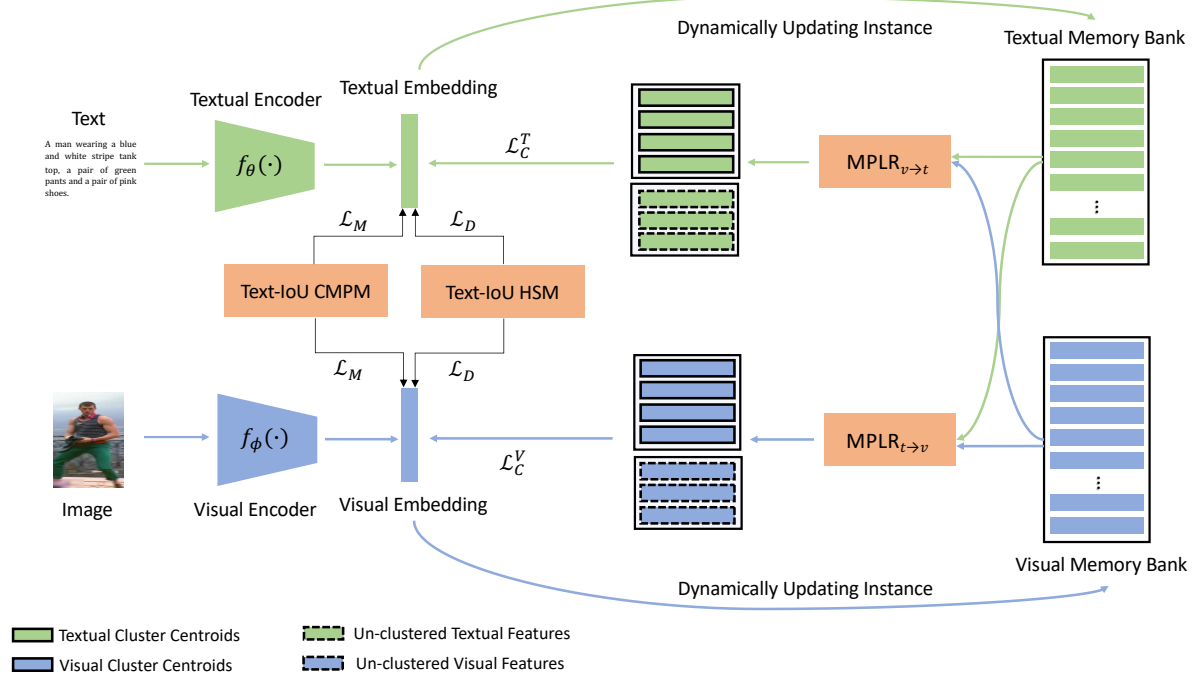
Figure 2. Illustration of the Cross-Modal Mutual Training (CMMT) framework. To mitigate the effect of intra-class variations, CMMT utilizes a clustering method to obtain preliminary pseudo labels. The Mutual Pseudo Label Refinement (MPLR) module leverages the clustering results in one modality to refine that in the other modality through the pairwise relationship between textual and visual modalities. The contrastive losses $\mathcal{L}_C^T$ and $\mathcal{L}_C^V$ are employed to supervise the identity representation learning for corresponding modality. Text-IoU CMPM is employed to relieve the cross-modal matching ambiguity on unpaired textual-visual instances. Text-IoU HSM is proposed to learn discriminative visual-textual joint embeddings. The Text-IoU CMPM loss and the discriminative embedding learning loss are denoted as $\mathcal{L}_M$ and $\mathcal{L}_D$, respectively.

3. Cross-Modal Mutual Training

3.1. Notations and Definitions

In weakly supervised text-based person Re-ID, we are given a training dataset $X = \{I_i, T_i\}_{i=1}^N$, where I_i is the ith image, T_i denotes the ith textual description that pairs with I_i, N is the number of text-image pairs. In contrast to fully supervised text-based person Re-ID, identity labels are not given. Based on these, our goal is to learn discriminative visual-semantic embeddings with only text-images pairs, so that we can search a person image from $X_{\text{gallery}} = \{I_j^g\}_{j=1}^{N^g}$ with N^g images by a textual description T^q.

3.2. Overview

We propose CMMT to address the weakly supervised text-based person Re-ID problem. As shown in Figure 2, the textual embedding f_i^t and the visual embedding f_i^v are extracted by the textual encoder f_θ and the visual encoder f_ϕ, respectively. In order to reduce inra-class variations in both modalities, clustering is conducted to obtain preliminary pseudo labels for both textual and visual instances. The reliability criterion [10] is utilized to preserve

the most reliable clusters, which measures the independence and compactness of clusters. The clustered textual and visual embeddings are denoted as $t^c = \{t_1^c, \cdots, t_{n_c^t}^c\}$ and $v^c = \{v_1^c, \cdots, v_{n_c^v}^c\}$, where n_c^t and n_c^v are the number of clustered instances in textual and visual modalities, respectively. The un-clustered textual and visual embeddings are denoted as $t^o = \{t_1^o, \cdots, t_{n_o^t}^o\}$ and $v^o = \{v_1^o, \cdots, v_{n_o^v}^o\}$, where n_o^t and n_o^v are the number of un-clustered instances in textual and visual modalities, repectively. We leverage memory banks, which are dynamically updated through the training phase, to provide cluster centroids and un-clustered instance features for both textual and visual modalities. The details of memory bank updating can be referred to the paper [10]. After that, CMMT leverages MPLR to mine the valuable un-clustered instances. Then the unsupervised identity learning is conducted based on the refined pseudo labels in both modalities. Moreover, the Text-IoU CMPM loss is proposed to mitigate the cross-modal matching ambiguity on the unpaired instances. Furthermore, Text-IoU HSM is presented in order to learn more discriminative textual-visual joint embeddings. In the rest of this section, we describe more details for each component individually.
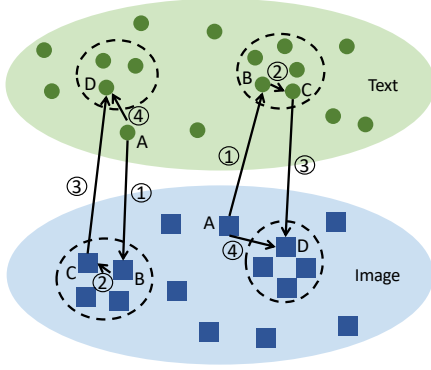
Figure 3. Illustration of MPLR. ① For each un-clustered instance A, we first search its paired instance B in the other modality. ② If B is clustered, we then find the nearest instance of B, marked as C. ③ After that, we search the paired instance D of C in the other modality. ④ We finally add A into the cluster that D belongs to, if D is clustered. More details can be found in Subsection 3.3.

3.3. Intra-Modal Self-Training by Pseudo Labels

In order to resolve the intra-class variation problem, one straightforward yet sub-optimal solution is to apply unsupervised representation learning in textual and visual modalities [31, 12, 10]. For example, after the clustering, a contrastive loss is utilized to supervised the identity representation learning for both modalities. The contrastive loss for the textual modality is give by,

$$\mathcal{L}_C^T = -\log \frac{\exp\left(\langle f_i^t, f^+ \rangle / \tau\right)}{\sum_{k=1}^{n^{ct}} \exp\left(\langle f_i^t, c_k^t \rangle / \tau\right) + \sum_{k=1}^{n_o^t} \exp\left(\langle f_i^t, t_k^o \rangle / \tau\right)},$$
(1)

where f^+ indicates the positive class prototype corresponding to f_i^t, the temperature τ is empirically set as 0.05. More specifically, if f_i^t belongs to the k-th cluster, $f^+ = c_k^t$ is the k-th textual cluster centroid. If f_i^t is a un-clustered outlier, we would have $f^+ = t_k^o$ as the outlier instance feature corresponding to f_i^t. Additionally, we conduct L2-normalization for all the features before calculating the loss. Meanwhile, the contrastive loss $\mathcal{L}_C^V$ for the visual modality can be defined similarly. Therefore, the overall contrastive loss is given by

$$\mathcal{L}_C = \mathcal{L}_C^T + \mathcal{L}_C^V.$$
(2)

Different from unsupervised person Re-ID, our training set consists of text-image pairs. Therefore, our motivation is that, in the ideal case, the clustering results in two modalities should be consistent due to the pairwise relationship. However, the intra-class variations lead to the inconsistency. The un-clustered instances, whose paired instances are clustered, may be crucial to learn discriminative features. Consequently, we exploit the pairwise relationship between two modalities to refine the clustering results.

Mutual Pseudo Label Refinement. To further suppress the intra-class variations in both modalities, we propose MPLR to mine valuable un-clustered instances, instead of simply discarding them. As shown in Figure 3, MPLR leverages clustering results in one modality to refine that in the other modality through the pairwise relationship between textual and visual modalities. The MPLR processes in textual and visual modalities are denoted as $MPLR_{v \rightarrow t}$ and $MPLR_{t \rightarrow v}$, respectively. For example, in $MPLR_{v \rightarrow t}$, for an un-clustered textual feature t_i^o, the paired visual instance v_i can be found by

$$v_i = PIS(t_i^o),$$
(3)

where $PIS(\cdot)$ denotes the paired instance searching in the other modality. If the obtained paired visual instance is un-clustered $v_i \in \boldsymbol{v^o}$, we keep t_i^o un-clustered. In contrast, if $v_i \in \boldsymbol{v^c}$, the nearest instance can be obtained by

$$v_i^c = \underset{v_i^c \in C_k^v, v_i^c \notin U_k^v}{\arg\max} \langle v_i, v_i^c \rangle,$$
(4)

where k denotes the index of a cluster, C_k^v is the set of all the visual instances of the cluster that v_i belongs to, U_k^v is a set that initialized to $\{v_i\}$, $\langle \cdot, \cdot \rangle$ denotes the inner product between two feature vectors to measure their similarity. Then, for the obtained v_i^c, the paired textual instance t_i can be found by

$$t_i = PIS(v_i^c),$$
(5)

If $t_i \in \boldsymbol{t^c}$, we add the un-clustered textual feature t_i^o to the cluster C_k^t that t_i belongs to, which is given by

$$C_k^t \leftarrow [C_k^t, t_i^o],$$
(6)

where $[\cdot, \cdot]$ denotes the process of merging the latter to the former. If $t_i \in \boldsymbol{t^o}$, the process returns to Equation 4 and the original nearest instance v_i^c is added to U_k^v until $C_k^v = U_k^v$. Note that $C_k^v = U_k^v$ represents the paired textual instances of the visual instances in C_k^v are all un-clustered. If $C_k^v = U_k^v$, we create a new textual cluster for t_i^o by

$$C_{n^{ct}+1}^t \leftarrow [t_i, t_i^o],$$
(7)

where n^{ct} is the number of clusters in the textual modality and it is updated by $n^{ct} \leftarrow n^{ct} + 1$ if a new clustered is created. During MPLR, all the un-clustered instances in both modalities are traversed.

3.4. Text-IoU Guided Cross-Modal Projection Matching

As shown in Figure 1(d), it is unable to assign positive or negative labels when learning the cross-modal matching in weakly supervised text-based person Re-ID. This leads to the cross-modal ambiguity problem. A straightforward method is utilizing textual embeddings or visual embeddings to calculate similarity soft-labels when conducting the

A woman wearing a pair of brown shorts, a pair of white shoes, a red shirt with short sleeves.

A woman in a blue shirt, a black skirt and a pair of white shoes.

NPE

NPE

P_i

P_j

{woman, brown shorts, white shoes, red shirt, short sleeves}

{woman, blue shirt, black skirt, white shoes}

$P_i \cup P_j$

$P_i \cap P_j$

{brown shorts, white shoes, red shirt, short sleeves, blue shirt, black skirt}

{woman, white shoes}

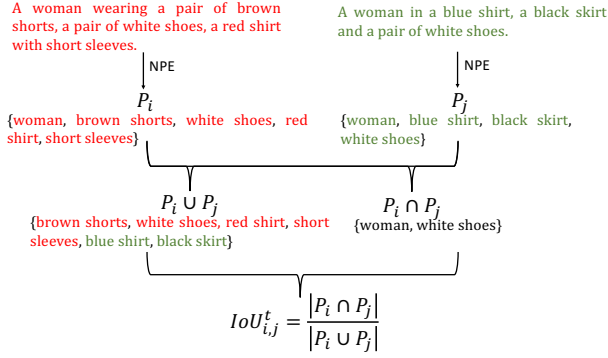$$IoU_{i,j}^t = \frac{|P_i \cap P_j|}{|P_i \cup P_j|}$$

Figure 4. Illustration of the calculation process of Text-IoU. First, the Noun Phrase Extraction (NPE) is conducted to obtain the set of noun phrases. Second, the sets of the intersection and the union of two phrases can be collected. Third, we calculate the Text-IoU score by dividing the element number of the intersection by the element number of the union.

cross-modal matching learning. However, there is still a number of outlier instances, which lead to inferior similarity soft-labels, especially in the beginning of the training phase. To resolve this problem, we propose the Text-IoU CMPM loss.

Cross-Modal Projection Matching. The traditional cross-modal projection matching, which incorporates the cross-modal projection into the KL divergence measure to associate the representations across different modalities, is given by,

$$p_{i,j} = \frac{\exp\left(\langle t_i, \bar{v}_j \rangle\right)}{\sum_{k=1}^{B} \exp\left(\langle t_i, \bar{v}_k \rangle\right)} . \tag{8}$$

where the visual embedding is first normalized by $\bar{v}_j = \frac{v_j}{||v_j||}$, B denotes the batch size, and the probability $p_{i,j}$ represents the proportion of this inner product among all inner products between pairs in a batch. Thus, the more similar the textual embedding is to the visual embedding, the larger the inner product is.

Text-IoU. In order to supervise the cross-modal projection matching under the lack of identity supervision, we introduce Text-IoU to assign the similarity soft-labels. Text-IoU measures the similarities in the phrase level among textual instances. For example, there are two text-image pairs $\{I_i, T_i\}$ and $\{I_j, T_j\}$, which can be encoded as visual and textual embeddings $\{f_i^v, f_i^t\}$ and $\{f_j^v, f_j^t\}$ after the forward propagation. Our motivation is that, if two textual descriptions are originated from the same person, the noun phrases in these two descriptions are always the same or synonymous. Therefore, as shown in Figure 4, for the given textual description T_i, we utilize NLTK [24] to extract the noun phrase, which is given by

$$P_i = NPE(T_i) = \{p_1, \cdots, p_{m_i}\}, \tag{9}$$

where NPE represents the noun phrase extraction, and m_i denotes the number of noun phrases in the textual description T_i. Therefore, Text-IoU can be defined by

$$IoU_{i,j}^t = \frac{|P_i \cap P_j|}{|P_i \cup P_j|}, \tag{10}$$

where $|P_i \cap P_j|$ is the number of the same noun phrases and the synonymous noun phrases between P_i and P_j, and $|P_i \cup P_j|$ is the total number of phrases in the union of P_i and P_j.

Text-IoU Guided Cross-Modal Projection Matching Loss. Text-IoU can be utilized as the similarity soft-label, which is then normalized by

$$q_{i,j} = \frac{IoU_{i,j}^t}{\sum_{k=1}^{B}(IoU_{i,k}^t)}. \tag{11}$$

The Text-IoU CMPM loss that associates f_i^t with the correctly matched visual features is then defined as the KL divergence from the true matching distribution q_i to the probability of matching p_i. For each batch, the Text-IoU CMPM loss is defined by

$$L_M = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j} + \epsilon}\right), \tag{12}$$

where ϵ is a very small number for preventing division by zero.

3.5. Text-IoU Guided Hard Sample Mining

Text-IoU HSM is proposed to learn the discriminative textual-visual joint embeddings, which consists of Text-IoU Guided Cross-Modal Hard Sample Mining (Text-IoU CHSM) and Text-IoU Guided Intra-Modal Hard Sample Mining (Text-IoU IHSM).

Text-IoU Guided Cross-Modal Hard Sample Mining. If the textual-visual features $\{f_i^t, f_j^v\}$ has high cross-modal feature similarity $\langle f_i^t, f_j^v \rangle$, the pair is a cross-modal *similar pair*. Inspired by the Multilabel Reference Learning [33], we assume that if the corresponding textual descriptions of a similar pair has high Text-IoU scores, it is probably a positive pair. Otherwise, it is probably a hard negative pair. Therefore, hard negative pairs can be mined by considering both the cross-modal feature similarity and the Text-IoU score. Specifically, the similar pairs are defined as the pM pairs that have the highest feature similarities among all the $M = N \times (N-1)/2$ pairs within the training dataset, where p is a mining ratio. If a similar pair (f_i^t, f_j^v) is among the top pM pairs with the highest Text-IoU scores, we assign (i, j) to the positive set $\mathcal{P}$. Otherwise, we assign it to the hard negative set $\mathcal{N}$. Formally, we construct

$$\mathcal{P} = \{(i,j) | \langle f_i^t, f_j^v \rangle \geq Q, IoU_{i,j}^t \geq R\}$$
$$\mathcal{N} = \{(k,l) | \langle f_k^t, f_l^v \rangle \geq Q, IoU_{k,l}^t < R\} \tag{13}$$

where Q is the inner product of the pM-th pair after sorting M pairs in a descending order, and R is the threshold value for the Text-IoU score. Then the ==Text-IoU Guided Cross-Modal Discriminative Embedding Learning loss== can be formulated by

$$\mathcal{L}_D^C = -\log \frac{\overline{P}}{\overline{P} + \overline{N}}, \qquad (14)$$

where

$$\overline{P} = \frac{1}{|\mathcal{P}|} \Sigma_{(i,j) \in \mathcal{P}} \exp(-||f_i^t - f_j^v||_2^2),$$

$$\overline{N} = \frac{1}{|\mathcal{N}|} \Sigma_{(k,l) \in \mathcal{N}} \exp(-||f_k^t - f_l^v||_2^2).$$

By minimizing $\mathcal{L}_D^C$, we learn a discriminative textual-visual joint embedding. Note that $\mathcal{P}$ and $\mathcal{N}$ are constructed dynamically with up-to-date feature embeddings in every batch during model learning, and M is simply replaced by $M_{batch} = B \times (B - 1)/2$ in this case.

Text-IoU Guided Intra-Modal Hard Sample Mining. Similarly, we can mine the intra-modal hard samples (i.e., , (f_i^t, f_j^t) and (f_i^v, f_j^v)) in each individual modalities through the same way, where Text-IoU is still utilized as the criterion of positive/negative pairs. Therefore, the losses of the ==Text-IoU Guided Intra-Modal Discriminative Embedding Learning== for textual and visual modalities can be defined as $\mathcal{L}_D^T$ and $\mathcal{L}_D^V$. The overall discriminative embedding learning loss is given by

$$\mathcal{L}_D = \mathcal{L}_D^C + \mathcal{L}_D^T + \mathcal{L}_D^V. \qquad (15)$$

3.6. Overall Loss

By combining the losses defined above, the final objective of CMMT is formulated as:

$$\mathcal{L}_{overall} = \mathcal{L}_C + \alpha \mathcal{L}_M + \beta \mathcal{L}_D, \qquad (16)$$

where α and β aim to control the relative importance of $\mathcal{L}_M$ and $\mathcal{L}_D$, respectively.

4. Experiment

In this section, we first describe the dataset and settings for evaluation. Then we compare our approach with the state-of-the-art methods and conduct the ablation studies. Finally, we show the qualitative results of our method.

4.1. Experimental Setting

Dataset. The CUHK-PEDES dataset is currently the only dataset for text-based person Re-ID. We follow the same data split as [20]. The training set has 11003 identities, 34054 images and 68126 textual descriptions. The test set

Table 1. Method comparison (%) on the CUHK-PEDES dataset. The methods in the 1st group are the fully-supervised text-based Re-ID methods. The methods in the 2nd group are the weakly supervised text-based Re-ID methods. The 3rd group is our method. Ticks in the ID column represent the methods that have identity supervision, and crosses denote no identity supervision. The best results are in bold.

Method	ID	Feature	R@1	R@5	R@10
GNA-RNN [20]	✓	global	19.05	-	53.64
CMCE [19]	✓	global	25.94	-	60.48
PWM-ATH [3]	✓	global	27.14	49.45	61.02
Dual Path [37]	✓	global	44.40	66.26	75.07
CMPM+CMPC [34]	✓	global	49.37	-	79.27
MIA [26]	✓	global+region	53.10	75.00	82.90
PMA [15]	✓	global+keypoint	53.81	73.54	81.23
GALM [14]	✓	global+keypoint	54.12	75.45	82.97
ViTAA [30]	✓	global+attribute	55.97	75.84	83.52
MM-TIM [11]	×	global	45.35	63.78	70.63
CMPM [34] + SpCL [10]	×	global	51.13	71.54	80.03
CMPM [34] + MMT [9]	×	global	50.51	70.23	78.98
CMMT	×	global	**57.10**	**78.14**	**85.23**

has 1000 identities, 3074 images and 6156 textual descriptions. Note that, although the identity labels are available in the training set, we do not use them in our method.

Evaluation protocols. We adopt Rank@K (K=1, 5, 10) to evaluate the performance. Given a query description, all the test images are ranked by their similarities with the query. A successful search is achieved if top-k images contain the corresponding identity.

Implementation Details. In our experiments, we set the textual encoder f_θ as Bi-LSTM [28] with one hidden layer. And both the dimensions of the hidden layer and the feature space are 1024. Meanwhile, the visual encoder f_ϕ is set as ResNet-50 [13] pre-trained on the ImageNet classification task. We use DBSCAN [5] for preliminary clustering before each epoch. Parameters for the final loss function are $\alpha = 0.8$ and $\beta = 0.9$. The mining ratio p is set to 3‰. The model is optimized with the Adam [17] optimizer, with a learning rate of 0.00032. It is decayed by 0.1 after 20 epochs. The batch size is set to 128, and the training stops at 50 epochs.

4.2. Comparisons with the State-of-the-Arts

We first compare the proposed approach with the existing methods on the CUHK-PEDES dataset. Since existing methods are not specifically designed for weakly supervised text-based person Re-ID, we select two groups of the compared methods, including fully-supervised approaches proposed for text-based person Re-ID and weakly supervised models for text-image retrieval, as listed in Table 1. Additionally, for the second group, we try our best to implement MM-TIM [11], and CMPM [34]) with two state-of-the-art unsupervised person Re-ID methods (i.e., SpCL [10] and MMT [9]).

Comparison with the fully-supervised text-based Re-ID methods. As shown in Table 1, we compare the proposed

Table 2. Component analysis (%) of the proposed method on the CUHK-PEDES dataset.

Method	R@1	R@5	R@10
Baseline	51.13	71.54	80.03
+ MPLR	53.28	74.67	82.21
+ Text-IoU CMPM	55.59	76.74	83.69
+ Text-IoU HSM	57.10	78.14	85.23

Table 3. Analysis of the Mutual Pseudo Label Refinement (MPLR) module. MPLR consists of $MPLR_{t \to v}$ and $MPLR_{v \to t}$.

Method	R@1	R@5	R@10
CMMT	57.10	78.14	85.23
CMMT (w/o $MPLR_{v \to t}$)	56.03	76.75	84.03
CMMT (w/o $MPLR_{t \to v}$)	55.96	76.48	84.16
CMMT (w/o MPLR)	54.92	75.08	83.02

Table 4. Analysis of the Text-IoU CMPM loss. The first method and the second method calculate similarity soft-labels by the inner product of textual embeddings (TE) and visual embeddings (VE), respectively.
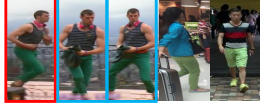
Method	R@1	R@5	R@10
TE CMPM	50.15	70.37	79.40
VE CMPM	48.59	67.25	76.78
Text-IoU CMPM	**57.10**	**78.14**	**85.23**

CMMT with the fully-supervised methods, which are supervised by identity annotations. We just copy and paste the results from these papers. The results show that ViTAA is superior among these methods. For example, ViTAA achieves 55.97% Rank-1 accuracy, 75.84% Rank-5 accuracy, and 83.52% Rank-10 accuracy, which outperforms the previous methods. This is because ViTAA leverages global and local features on both visual and textual modalities to improve the cross-modal feature alignment, and an auxiliary segmentation layer is employed for the knowledge distillation on the local level. Even compared with the fully-supervised methods, our method shows the best performance. For example, CMMT outperforms ViTAA by 1.13% Rank-1 accuracy, 2.30% Rank-5 accuracy, and 1.71% Rank-10 accuracy. It indicates that CMMT can learn the discriminative textual-visual joint embeddings without identity annotations.

Comparison with the weakly supervised text-image retrieval methods. MM-TIM simply performs the cross-modal matching without learning the discriminative identity features. CMPM + MMT and CMPM + SpCL perform self-training in textual and visual modalities, and all the textual-visual instances (except the paired ones) are treated as negatives when performing cross-modal matching. As shown in Table 1, without the supervision of identity labels, they are not comparable with the fully-supervised methods. For example, CMPM + SpCL reaches 51.13% Rank-1 accuracy, 71.54% Rank-5 accuracy, and 80.03% Rank-1 accuracy. Our method surpasses CMPM + SpCL by a large margin with 57.10% Rank-1 accuracy, 78.14% Rank-5 accuracy, and 85.23% Rank-10 accuracy. This is because our proposed CMMT explicitly utilizes the textual-visual pairwise relationship and Text-IoU to facilitate cross-modal identity-specific representation learning.

4.3. Ablation Studies

Contributions of Individual Components. In Table 2, we evaluate the contributions of three components to the full model. For the baseline method, all the textual-visual instances (except their paired ones) are set to be negative during training when training the cross-modal matching. Additionally, we try to build a baseline with high performance to demonstrate the effectiveness of our proposed

modules because of the high baseline performance of ViTAA(global branch) [30]. The results show that all of our proposed components are effective on their own. For example, the baseline with MPLR outperforms the baseline method by 2.15% Rank-1 accuracy, 3.13% Rank-5 accuracy, and 2.18% Rank-10 accuracy. Text-IoU CMPM further contributes the improvement of 2.31% Rank-1 accuracy, 2.07% Rank-5 accuracy, and 1.48% Rank-10 accuracy. Moreover, when combined all the components, the best performance is achieved. This validates our design consideration in that they are complementary and should be combined.

Analysis of MPLR. As shown in Table 3, we conduct the ablation study of MPLR, which consists of $MPLR_{t \to v}$ and $MPLR_{v \to t}$. We observe that both $MPLR_{t \to v}$ and $MPLR_{v \to t}$ contribute to the performance of MPLR. This validates the effectiveness of $MPLR_{t \to v}$ and $MPLR_{v \to t}$. For example, if $MPLR_{t \to v}$ is removed, the Rank-1, Rank-5, and Rank-10 accuracies are reduced by 1.14%, 1.66%, and 1.07%, respectively.

Analysis of Text-IoU CMPM. To resolve the cross-modal matching ambiguity problem, a straightforward method is setting the inner product of textual embeddings or visual embeddings from our self-training models as the similarity soft-labels. As shown in Table 4, the results show that Text-IoU CMPM achieves better results than the other two methods by large margins. For example, Text-IoU CMPM surpasses the method, which utilizes text embeddings to calculate similarity soft-labels, by 6.95% Rank-1 accuracy, 7.77% Rank-5 accuracy, and 5.83% Rank-10 accuracy. These results are even worse than that of the baseline. We believe the reason is that the embeddings in the beginning of the training phase are not discriminative enough,
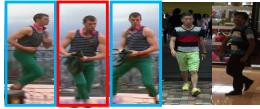
A man wearing a blue and white stripe tank top, a pair of green pants and a pair of pink shoes.

The man is wearing a white shirt and gray shorts. He is also wearing black sandals.

The man is wearing a blue and white striped tank top and green pants. He has pink headphones around his neck.

A man in a gray shirt, a pair of brown shorts and a pair of black shoes.

The man is wearing a black and white striped tank with green leggings and pink headphones around his neck.

A man wearing a pair of brown shorts, a pair of black shoes, a red shirt with short sleeves.

Figure 5. Qualitative Results of the proposed CMMT on the CUHK-PEDES dataset. In each group, we show the top-5 rank galley images for each query text. Both red rectangles and blue rectangles represent the correct retrieval results. The former indicates the paired image of the query text. The latter represents the unpaired image of the query text.

Table 5. Analysis of the Text-IoU HSM method. Text-IoU HSM consists of Text-IoU CHSM and Text-IoU IHSM.

Method	R@1	R@5	R@10
CMMT	57.10	78.14	85.23
CMMT (w/o Text-IoU CHSM)	56.02	77.05	84.12
CMMT (w/o Text-IoU IHSM)	56.42	77.49	84.56
CMMT (w/o Text-IoU HSM)	55.59	76.74	83.69

which leads to inferior similarity soft-labels. Then the network may fall to the local optimality. These results validate the effectiveness of our proposed Text-IoU CMPM.

Analysis of Text-IoU HSM. As shown in Table 5, the ablation study of Text-IoU HSM is conducted, which consists of Text-IoU CHSM and Text-IoU IHSM. The results show that both Text-IoU CHSM and Text-IoU IHSM are beneficial to the performance of Text-IoU HSM, especially Text-IoU CHSM. For example, if Text-IoU CHSM is removed, the Rank-1, Rank-5, and Rank-10 accuracies are less by 1.08%, 1.09%, and 1.11%, respectively.

Influence of Parameters. We evaluate two key parameters in our modeling, the loss weights α and β in Equation 16. As shown in Figure 6, the performance peaks at $\alpha = 0.8$ and $\beta = 0.9$. When α and β are set between 0.6 and 1.0, and 0.5 and 1.0, respectively, the performance does not change dramatically, which indicates that CMMT is insensitive to the α and β in the value ranges.

4.4. Qualitative analysis

We conduct a qualitative evaluation for our proposed CMMT. Figure 5 shows six person Re-ID results with natural language descriptions by CMMT. We can conclude that: 1) As shown in the first column, different textual descriptions of the same identity retrieve similar person images. Meanwhile, as shown in the second column, one textual description can retrieve the images regardless of variations
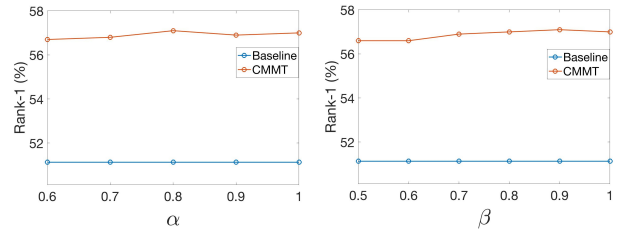

Figure 6. Evaluation (%) of the loss weights α and β in Equation 16 using Rank-1 accuracy on the CUHK-PEDES dataset.

such as camera-view, low illumination, *etc*. It indicates the proposed CMMT suppresses the intra-class variations in both textual and visual modalities. 2) The CMMT can retrieve both the paired image and the unpaired image. It is because CMMT leverages Text-IoU CMPM to solve the cross-modal matching ambiguity.

5. Conclusions

We have considered a new text-based person Re-ID challenge: weakly supervised text-based person Re-ID. To address its particular challenges, we have proposed CMMT, which mainly consists of MPLR, Text-IoU CMPM, and Text-IoU HSM. MPLR and Text-IoU CMPM are specifically designed to address the problems caused by the intra-class variations and the cross-modal matching ambiguity. Moreover, Text-IoU HSM is further presented to learn more discriminative textual-visual joint embeddings. We have conducted extensive experiments on the CUHK-PEDES dataset and demonstrated the effectiveness of the proposed model, which outperforms existing text-based person re-identification methods.

Acknowledgment

References

[1] Farooq Ammarah, Awais Muhammad, Yan Fei, Kittler Josef, Akbari Ali, and Khalid Syed Safwan. A convolutional baseline for person re-identification using vision and language descriptions. *arXiv preprint arXiv:2003.00808*, 2020. 2

[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018. 2

[3] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *WACV*, 2018. 2, 6

[4] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 2

[5] Martin Ester, Hans-Peter Kriegel, Joerg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 6

[6] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, 2019. 2

[7] 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*, 2020. 2

[8] Jing Ge, Guangyu Gao, and Zhen Liu. Visual-textual association with hardest and semi-hard negative pairs mining for person search. *arXiv preprint arXiv:1912.03083*, 2019. 2

[9] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 2, 6

[10] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NIPS*, 2020. 2, 3, 4, 6

[11] Raul Gomez, Gomez Lluis, Gibert Jaume, and Karatzas Dimosthenis. Self-supervised learning from web data for multimodal retrieval. *Multimodal Scene Understanding*, 2019. 2, 6

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 6

[14] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. *arXiv preprint arXiv:1809.08440*, 2018. 6

[15] 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 *AAAI*, 2020. 1, 2, 6

[16] Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Cross-modal cross-domain moment alignment network for person search. In *CVPR*, 2020. 2

[17] Diederik P. Kingma and J. Adam Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[18] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019. 1

[19] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, 2017. 2, 6

[20] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017. 1, 2, 6

[21] Yutian Lin, Xuanyi Dong, Yan Yan Zheng, Liang, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019. 2

[22] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *CVPR*, 2020. 2

[23] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, 2020. 1

[24] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint arXiv:0205028*, 2002. 5

[25] Sarafianos Nikolaos, Xu Xiang, and A. Kakadiaris Ioannis. Adversarial representation learning for text-to-image matching. In *ICCV*, 2019. 1, 2

[26] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *arXiv preprint arXiv:1906.09610*, 2019. 2, 6

[27] Yash Patel, Gomez Lluis, Rusiñol Marçal, Karatzas Dimosthenis, and C. V. Jawahar. Self-supervised visual representations for cross-modal retrieval. In *ICMR*, 2019. 2

[28] Mike Schuster and Paliwal Kuldip K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997. 6

[29] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *arXiv preprint arXiv:1807.11334*, 2018. 2

[30] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *ECCV*, 2020. 1, 2, 6, 7

[31] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 4

[32] Fei Yan, Mikolajczyk Krystian, and Kittler Josef. Person re-identification with vision and language. In *ICPR*, 2017. 2

[33] 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 *CVPR*, 2019. 5

[34] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018. 2, 6

[35] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 2

[36] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2

[37] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2020. 2, 6

[38] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, 2018. 2