Image-Centered Pseudo Label Generation for Weakly Supervised Text-Based Person Re-Identification

Weizhi Nie, Chengji Wang(✉), Hao Sun, and Wei Xie

Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, National Language Resources Monitoring and Research Center for Network Media, School of Computer Science, Central China Normal University, Wuhan, China
weizhi.nie@mails.ccnu.edu.cn, {wcj,haosun,xw}@ccnu.edu.cn

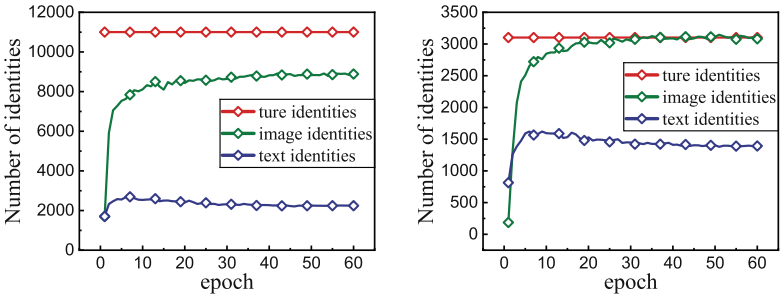
Abstract. Weakly supervised text-based person re-identification aims to identify a target person using textual descriptions, where the identity annotations are not available during the training phase. Previous methods attempted to cluster images and texts simultaneously for generating pseudo identity labels. However, we observed that the number of text clusters is significantly smaller than the number of true identities, while the number of image clusters is closer to the actual number of identities. This leads to uncertain pseudo identity labels. To address this issue, we propose a new approach called Image-Centered Pseudo Label Generation (ICPG) for weakly supervised text-based person re-identification. It directly generates pseudo labels for images and texts based on image clustering results. Firstly, we introduce a cross-modal distribution matching loss, which focuses on minimizing the KL divergence between the distributions of image-text similarity and normalized pseudo label matching distributions. Secondly, to enhance cross-modal associations, we propose a cross-modal hard sample mining method to explore challenging cross-modal examples. Experimental results demonstrate the effectiveness of our proposed methods. Compared to the state-of-the-art method, our approach achieves improvements of 3.6%, 2.4% and 3.0% in rank-1 accuracy on three datasets, respectively.

Keywords: Weakly Supervised Learning · Pseudo Label Generation · Text-based Person Re-Identification

1 Introduction

Weakly supervised text-based person re-identification [37] aims to retrieve target individuals based on textual descriptions without identity annotations available during training. It has received increasing attention in recent years. Compared to text-based person re-identification [15,19,30,32,34], weakly supervised text-based person re-identification lacks identity annotations. Cross-modal semantic differences and intra-class variations make it a challenging task. The variance

Z. Lin et al. (Eds.): PRCV 2024, LNCS 15042, pp. 1–15, 2024.
https://doi.org/10.1007/978-981-97-8858-3_33



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

Fig. 1. During training, we employ the original CLIP loss to compare the number of image and text pseudo labels in each dataset with the number of true identities.

of illumination condition, human posture, view angle, and other factors leads to the intra-class variations. The absence of identity labels makes it more difficult to mitigate the intra-class variations.

Previous methods believed that pairwise relationships between textual and visual modalities could enhance clustering performance. Therefore, they simultaneously conducted clustering on both text and images, generating pseudo labels for both textual and visual instances. In order to mitigate intra-class variations. CMMT [37] generates cross-modal soft labels according to the IoUs among texts. CPCL [38] proposed an Outlier Pseudo Label Mining module to mine implicit relationships between image-text pairs. However, as depicted in Fig. 1, we observe significant differences in the quantities of pseudo labels between the image and text modalities on the CUHK-PEDES [19] and ICFG-PEDES [6] datasets through experiments. Specifically, the number of text pseudo labels is significantly smaller than the number of true identities, while the number of image pseudo labels is closer to the actual number of identities. There is a obvious gap between the number of image and text clusters, which leads to uncertain pseudo identity labels and cross-modal matching ambiguity.

To address the above problem, we propose a novel approach called Image-Centered Pseudo Label Generation (ICPG). It simultaneously assigning pseudo labels to both images and text based on the clustering results of images. Firstly, to explore a more effective cross-modal matching objective, we introduce a Cross-Modal Distribution Matching (CDM) loss. The CDM loss integrates the cosine similarity distributions of image-text pair embeddings into the KL divergence. It minimizes the KL divergence between the distributions of image-text similarity scores and normalized pseudo label matching distributions, aiming to discover additional potential image-text matching pairs. Secondly, in order to improve clustering performance and enhance the quality of pseudo labels, we propose a Cross-Modal Hard Sample Mining (CHM) method, it exploits the hard negative samples from the closest cluster, making different clusters more dispersed. This method learns discriminative textual-visual joint embeddings by exploring challenging hard negative samples. Our main contributions can be summarized as follows:

- We propose an Image-Centered Pseudo Label Generation method to address the issues of uncertain cross-modal pseudo identity labels.
- We introduce a Cross-Modal Distribution Matching loss. It utilizes the clustering results of images and the pairing relationships between the two modalities to explore potential image-text matching pairs.
- We propose a novel Cross-Modal Hard Sample Mining based on image pseudo labels, which exploits the hard negative samples from the closest cluster, making different clusters more dispersed.
- Extensive experiments on three public benchmark datasets show that our proposed ICPG consistently outperforms the state-of-the-arts by a margin.

2 Related Work

2.1 Text-Based Person Re-Identification

Text-based person re-identification [19,29,31] aims to retrieve images of target individuals based on provided textual descriptions. Early studies [1,18,19] utilized VGG [27] and LSTM [14] to extract representations from visual and textual modalities, followed by the utilization of matching loss functions to align them within a shared latent space. Subsequent research [2] enhanced feature extraction by employing Resnet [12] and BERT [5], and introduced novel cross-modal matching losses to align global image-text features within the joint embedding space. With increased attention on CLIP [24], more researchers [15,20] are endeavoring to exploit its powerful knowledge for constructing effective mapping relationships between images and text.

2.2 Unsupervised Person Re-Identification

The objective of unsupervised person re-identification [8,21,39] is to acquire discriminative feature representations in the absence of identity annotations. Unsupervised person re-identification methods primarily fall into two categories: completely unsupervised (USL) methods [21,36] and unsupervised domain adaptation (UDA) methods. Certain UDA methods rely on pseudo labels, initially training the network in the source domain and subsequently fine-tuning it in the target domain using these labels [3,9,22,28]. SpCL [10] is an unsupervised self-paced contrastive learning framework that utilizes instance-level hybrid memory. Cluster Contrast [4] employs cluster-level memory to mitigate the issue of inconsistent updating of memory cluster centroids. Unlike the unsupervised person re-identification, weakly supervised text-based person re-identification relies solely on image-text pairs during the training phase.

2.3 Weakly Supervised Person Re-Identification

Fully supervised Re-ID methods requires identity annotations. To overcome this constraint, CMMT [37] introduced weakly supervised text-based re-identification. CMMT uses pseudo labels for self-training in each modality,

and utilizes Text-IoU scores as soft labels of similarity to facilitate cross-modal matching learning. CPCL [38] introduces a prototype multimodal memory for cluster-level contrastive learning and instance-level cross-modal matching, facilitating cross feature learning between visual and textual modalities. Both CMMT and CPCL methods separately cluster images and texts, generating pseudo labels and allocating memory for cluster centroid features in each modality. In contrast, our method only requires clustering of images without requiring additional memory.

3 Method

3.1 Overview

The overview of ICPG is illustrated in Fig. 2. CLIP [24], pre-trained with abundant image-text pairs and possessing powerful underlying cross-modal alignment capabilities. We utilize CLIP's text encoder and image encoder to respectively extract text embedding f_i^t and visual embedding f_i^v. We propose an Image-Centered Pseudo Label Generation method to generate cross-modal pseudo labels. The Cross-Modal Distribution Matching (CDM) loss minimizes the KL divergence between the image-text similarity distributions and normalized pseudo label matching distributions, facilitating alignment of visual and textual embeddings. Additionally, Cross-Modal Hard Sample Mining (CHM) enhances the learning of discriminative textual-visual joint embeddings by mining challenging hard negative samples.

3.2 Image-Centered Pseudo Label Generation

Before each epoch, we first extract the features of all images and texts, where images features represent $F^v = \{f_1^v, f_2^v, \ldots, f_{n_v}^v\}$ and text features denote $F^t = \{f_1^t, f_2^t, \ldots, f_{n_t}^t\}$, n_v and n_t respectively represent the number of images and texts. Then, we use the DBSCAN [7] algorithm to cluster image features and obtain image pseudo labels:

$$Y^v = DBSCAN(F^v), \tag{1}$$

$Y^v = \{y_i^v\}_{i=1}^{n_v}$, where $y_i^v = -1$ indicates that the i-th image is an un-clustered instance; otherwise, y_i^v represents the index of the cluster. Leveraging the pairing relationships between images and texts, we assign pseudo labels $Y^t = \{y_i^t\}_{i=1}^{n_t}$ to text using the following rule:

$$y_i^t = \begin{cases} y_i^v, & y_i^v \neq -1 \\ -1, & \text{otherwise} \end{cases}. \tag{2}$$

Given a mini-batch with N image-text pairs $\{f_i^v, f_i^t\}_{i=1}^N$, the corresponding pseudo pairs labels are denoted as $\{y_i^v, y_i^t\}_{i=1}^N$, where $y_i^v = y_i^t$.

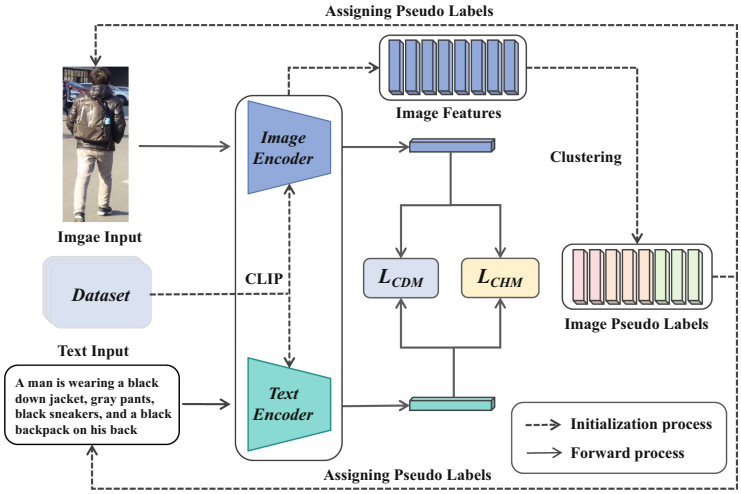


Fig. 2. An overview of our proposed ICPG framework.

3.3 Cross-Modal Distribution Matching

We introduce a Cross-Modal Distribution Matching (CDM) loss, which utilizes only image pseudo labels to supervise both image and text modalities. The purpose is to explore additional potential image-text matching pairs. The CDM loss integrates the cosine similarity distributions of the $N \times N$ image-text pairs embeddings into KL divergence, facilitating the alignment of representations across modalities.

Given a mini-batch of N image-text pairs $\{f_i^v, f_i^t\}_{i=1}^N$, we compute the cross-modal matching probability between i-th image and j-th text:

$$p_{i,j} = \frac{\exp(cos(f_i^v, f_j^t)/\tau)}{\sum_{k=1}^{N} \exp(cos(f_i^v, f_k^t)/\tau)}, \tag{3}$$

where $cos(,)$ denotes cosine similarity, and $\tau = 0.02$ serves as a temperature hyperparameter. The matching probability $p_{i,j}$ represents the ratio of the matching score between f_i^v and f_j^t to the total score between f_i^v and $\{f_j^t\}_{j=1}^N$ within the mini-batch. Then, we calculate the distance between the predicted distribution and the true distribution by the following formula:

$$\mathcal{L}_{cdm}^{i2t} = KL(\mathbf{p_i} \| \mathbf{q_i}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{i,j} \log(\frac{p_{i,j}}{q_{i,j} + \epsilon}), \tag{4}$$

where ϵ is a small number to avoid numerical problems, and $q_{i,j} = y_{i,j} / \sum_{k=1}^{N} y_{i,k}$ is the true cross-modal matching probability. When $y_i^v = y_j^t$, $y_{i,j} = 1$. Otherwise, $y_{i,j} = 0$.

The $\mathcal{L}_{cdm}^{i2t}$ is the image to text matching loss. Symmetrically, the text to image matching loss $\mathcal{L}_{cdm}^{t2i}$ can be formulated by exchanging f^v and f^t in Eqs. 3 and 4. The bi-directional CDM loss is then calculated by:

$$\mathcal{L}_{cdm} = \mathcal{L}_{cdm}^{i2t} + \mathcal{L}_{cdm}^{t2i}. \tag{5}$$

3.4 Cross-Modal Hard Sample Mining

In order to improve clustering performance and generate higher quality pseudo labels, we propose a Cross-Modal Hard Sample Mining (CHM) method, it exploits the hard negative samples from the closest cluster, making different clusters more dispersed. It learns discriminative textual-visual joint embeddings by exploring challenging hard negative samples.

Given a mini-batch, $\mathcal{L}_{chm}$ firstly selects a sample x_a as the anchor from one modality, then chooses a positive sample x_p from another modality paired with the anchor. Subsequently, it selects the hard negative sample x_n from the other modality, unpaired with the anchor but exhibiting the highest similarity to the anchor. This means that the anchor shares the same pseudo label with positive samples, i.e., $y_{x_a} = y_{x_p}$, but anchors have different pseudo labels from hard negative samples, i.e., $y_{x_a} \neq y_{x_n}$.When the image modality is used as an anchor, the loss is calculated as follows:

$$\mathcal{L}_{chm}^{i2t} = \sum_{i=1}^{B} \max[0, \beta_{\text{mar}} + sim(x_a^i, x_n) - sim(x_a^i, x_p)], \tag{6}$$

where B represents the batch size, $sim(,)$ denotes the cosine similarity, and β_{mar} is the hyperparameter margin in triplet loss [13]. Similarly, when employing the text modality as an anchor, $\mathcal{L}_{chm}^{t2i}$ can be obtained likewise to the aforementioned loss. The total loss function is then calculated by:

$$\mathcal{L}_{chm} = \mathcal{L}_{chm}^{i2t} + \mathcal{L}_{chm}^{t2i}. \tag{7}$$

The objective of this loss function is to minimize the distance between samples of the same identity while maximizing the distance between samples of different identities but with high similarity. This process helps to identify challenging examples, improves clustering, and enhances model performance.

3.5 Optimization

During the initial training stage, the representations are usually of low quality. Introducing hard samples during this phase might be counterproductive and may lead to the incorrect direction of model optimization from the beginning. To mitigate this concern, we introduce progressive learning, which forms the overall loss function:

$$\mathcal{L}_{icpg} = \begin{cases} \mathcal{L}_{itc} + \mathcal{L}_{cdm}, & \text{if epoch} \leq E_\lambda \\ \mathcal{L}_{itc} + \mathcal{L}_{cdm} + \mathcal{L}_{chm}, & \text{else} \end{cases} \tag{8}$$

where $\mathcal{L}_{itc}$ [23] is the baseline of our study and E_λ is a hyperparameter.

4 Experiment

4.1 Datasets and Evaluation Metrics

In this work, we evaluate our method on three challenging text-based person search datasets. Here is a brief introduction to these datasets:

CUHK-PEDES [19] comprises 40,206 images and 80,412 text descriptions belonging to 13,003 identities. According to official protocol, the dataset is split into 34,054 images and 68,108 descriptions for 11,003 identities in the training set, 3,078 images with 6,158 descriptions for 1,000 identities in the validation set, and 3,074 images with 6,156 descriptions of 1,000 persons in the testing set.

ICFG-PEDES [6] has more identities and textual descriptions, including a total of 54,522 images for 4,102 identities. The dataset is divided into a training set and a testing set, the former comprises 34,674 image-text pairs for 3,102 identities, and the latter contains 19,848 image-text pairs for the remaining 1,000 identities.

RSTPReid [40] contains 20,505 images of 4,101 identities captured by 15 cameras. Each identity is represented by 5 distinct images taken from various cameras, with each image annotated by 2 textual descriptions. Following the official data segmentation, the training, validation and test set contain 3,701, 200 and 200 identities, respectively.

Evaluation Metrics. We utilize the widely accepted Rank-k metrics (where $k = 1, 5, 10$) as our primary evaluation metrics. Additionally, we also adopt the mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP) [35] as another retrieval criterion. Higher values of Rank-k, mAP, and mINP indicate superior performance.

4.2 Implementation Details

We utilize the pre-trained CLIP-ViT-B/16 [24] as our backbone network, including both image and text encoders. During the training phase, all input images are resized to 384×128 with data augmentation, including random horizontally flipping, random cropping and random erasing. The maximum length of the textual token sequence L is set to 77. We set the batch size to 64 and train our model for 60 epochs using the Adam optimizer [16], with a learning rate initialized to 1×10^{-6} and cosine learning rate decay. The temperature parameter in CDM loss is set to 0.02, the margin hyperparameter β_{mar} is set to 0.3 in Eq. 6, and the hyperparameter E_λ set to 20 in Eq. 8. Following the approach of CMMT [37], we employ the DBSCAN [7] algorithm for clustering before each epoch. We use the CLIP-ViT-B/16 model fine-tuned with ITC [23] loss as the baseline for our study. Experiments are conducted using a single RTX4090 24GB GPU.

4.3 Comparisons with the State-of-the-Art Methods

In this section, we present comparison results with state-of-the-art methods on three public benchmark datasets.

Table 1. Performance comparisons with SOTA methods on CUHK-PEDES. The "G" and "L" in the "Type" column represent global/local matching methods.

Method	Type	Rank-1	Rank-5	Rank-10	mAP	mINP
Fully Supervised Text-based Person Re-Identification						
Han et al. [11]	G	64.08	81.73	88.19	60.08	–
SAF [17]	L	64.13	82.62	88.40	–	–
IVT [26]	G	65.59	83.11	89.21	–	–
UniPT [25]	G	68.50	84.67	90.38	–	–
CFine [34]	L	69.57	85.93	91.15	–	–
Wei et al. [20]	G	71.59	87.95	92.45	65.03	49.97
IRRA [15]	G	**73.38**	**89.93**	**93.71**	**66.13**	**50.24**
Weakly Supervised Text-based Person Re-Identification						
CMMT [37]	G	57.10	78.14	85.23	–	–
CAIBC [33]	L	58.64	79.02	85.93	–	–
CPCL [38]	G	70.03	87.28	91.78	63.19	47.54
Baseline (CLIP-ViT-B/16)	G	68.29	86.31	91.73	61.19	44.96
ICPG (Ours)	G	**73.68**	**88.84**	**93.12**	**65.87**	**49.78**

Table 2. Performance comparisons with SOTA methods on ICFG-PEDES.

Method	Type	Rank-1	Rank-5	Rank-10	mAP	mINP
Fully Supervised Text-based Person Re-Identification						
IVT [26]	G	56.04	73.60	80.22	–	–
UniPT [25]	G	60.09	76.19	82.46	–	–
CFine [34]	L	60.83	76.55	82.42	–	–
Wei et al. [20]	G	60.93	77.96	84.11	36.44	7.79
IRRA [15]	G	**63.46**	**80.25**	**85.82**	**38.06**	**7.93**
Weakly Supervised Text-based Person Re-Identification						
CPCL [38]	G	62.60	79.07	84.46	36.16	6.31
Baseline (CLIP-ViT-B/16)	G	56.86	75.70	82.21	31.89	5.27
ICPG (Ours)	G	**65.03**	**80.60**	**85.95**	**37.79**	**6.98**

Performance Comparisons on CUHK-PEDES. We first evaluate the proposed method on CUHK-PEDES dataset. As shown in Table 1. ICPG achieves 73.68%, 88.84% and 93.12% on Rank-1, Rank-5 and Rank-10, respectively. Compared to weakly supervised methods, our approach surpasses the recent state-of-the-art method CPCL [38] by margins of +3.65%, +1.56% and +1.34%, respectively. This is attributed to our method only using image pseudo labels to jointly supervise both modalities, effectively alleviating cross-modal ambiguity issues. Additionally, compared to fully supervised methods that also

Table 3. Performance comparisons with SOTA methods on RSTPReid.

Method	Type	Rank-1	Rank-5	Rank-10	mAP	mINP
Fully Supervised Text-based Person Re-Identification						
IVT [26]	G	46.70	70.00	78.80	–	–
CFine [34]	L	50.55	72.50	81.60	–	–
UniPT [25]	G	51.85	74.85	82.85	–	–
Wei et al. [20]	G	56.65	77.40	84.70	45.27	26.02
IRRA [15]	G	**60.20**	**81.30**	**88.20**	**47.17**	**25.28**
Weakly Supervised Text-based Person Re-Identification						
CPCL [38]	G	58.35	81.05	87.65	45.81	23.87
Baseline (CLIP-ViT-B/16)	G	54.90	78.85	86.55	43.36	22.56
ICPG (Ours)	G	**61.40**	**81.30**	**88.55**	**47.15**	**24.37**

use CLIP as the backbone network, our method surpasses the IRRA [15] method in Rank-1 accuracy and closely approaches its performance in all other metrics. This shows that our method achieves strong competitiveness without using any identity labels, relying solely on image pseudo labels generated by clustering.

Performance Comparisons on ICFG-PEDES. The experimental results on the ICFG-PEDES dataset are reported in Table 2. Compared with the Baseline, ICPG achieves a significant improvement of 8.17%, 4.90% and 3.74% on Rank-1, Rank-5 and Rank-10, respectively. In comparison with weakly supervised methods, our method ICPG achieves state-of-the-art performance in all metrics. Additionally, both the CMMT [37] and CPCL [38] methods separately cluster images and texts, generating pseudo labels and allocating memory for cluster centroid features in each modality. In contrast, our method only requires clustering of images without requiring additional memory. Compared to the fully supervised method IRRA [15], our method ICPG outperforms it on Rank-1, Rank-5, and rank-10. In addition, IRRA employed a masked language model that introduced additional parameters, while our method did not introduce any training parameters.

Performance Comparisons on RSTPReid. We also present our experimental results on the RSTPReid dataset in Table 3. Compared with existing weakly supervised methods, our method ICPG achieves state-of-the-art performance across all evaluation metrics. This demonstrates the effectiveness of our image-centered pseudo label generation approach in alleviating cross-modal ambiguity. Furthermore, our method outperforms most existing fully supervised methods, even surpassing the state-of-the-art fully supervised method IRRA [15] in Rank-1 and Rank-10. This also highlights our strong competitiveness.

In summary, our ICPG consistently achieves the best performance for all metrics on all three benchmark datasets. This demonstrate the generalization and robustness of our proposed method.

Table 4. Ablation studies on our proposed components of ICPG.

No.	Methods	Components			CUHK-PEDES			ICFG-PEDES			RSTPReid		
		TAug	$\mathcal{L}_{cdm}$	$\mathcal{L}_{chm}$	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
0	Baseline				68.29	86.31	91.73	56.86	75.70	82.21	54.90	78.85	86.55
1	+TAug	✓			71.16	88.73	93.10	59.86	78.08	84.04	57.70	80.75	88.40
2	$+\mathcal{L}_{cdm}$		✓		70.37	87.41	92.24	60.72	78.08	84.12	58.75	80.65	88.65
3	$+\mathcal{L}_{chm}$			✓	70.45	87.10	91.94	62.51	78.49	84.12	57.40	79.45	87.25
4	$+\mathcal{L}_{cdm}+\mathcal{L}_{chm}$		✓	✓	71.17	87.41	91.98	63.23	79.19	84.86	59.95	80.45	87.50
5	ICPG	✓	✓	✓	**73.68**	**88.84**	**93.12**	**65.03**	**80.60**	**85.95**	**61.40**	**81.30**	**88.55**

4.4 Ablation Study

As demonstrated in Table 4, we conducted an analysis of the effectiveness of each component in the ICPG method. $\mathcal{L}_{chm}$ represents the cross-modal hard sample mining module, $\mathcal{L}_{cdm}$ denotes the cross-modal distribution matching loss. Additionally, TAug stands for text data augmentation. Specifically, when provided with an input text description, we randomly mask out text tokens with a probability of 15% and replace them with the special token [MASK].

Cross-Modal Distribution Matching. To demonstrate the effectiveness of our proposed Cross-Modal Distribution Matching (CDM) loss, simply adding the CDM to Baseline improves the Rank-1 accuracy by 2.08%, 3.86% and 3.85% on the three datasets, respectively. Furthermore, the efficacy of CDM is further evidenced by the experimental results of entries No. 3 and No. 4 in Table 4. These results demonstrate that the proposed CDM loss effectively aligns the features representations between the two modalities.

Cross-Modal Hard Sample Mining. The effectiveness is demonstrated by the experimental results of No. 0 *vs.* No. 3 and No. 2 *vs.* No. 4 in Table 4. Compared to the baseline, merely adding this module improves the Rank-1 accuracy by 2.16%, 5.65% and 2.5% on three datasets, respectively. These results clearly indicate that the CHM module is beneficial for learning discriminative textual-visual joint embeddings and improving model performance.

Text Data Augmentation. The efficacy of TAug is revealed via the experimental results of No. 0 *vs.* No. 1 and No. 4 *vs.* No. 5 in Table 4. Merely adding the TAug to Baseline improves the Rank-1 accuracy by 2.8%, 3.0% and 2.8% on the three datasets, respectively. Randomly dropping some keywords allows the model pay more attention to other words, enabling the model to deeply explore fine-grained image and text features, effectively enhancing cross-modal fine-grained semantic associations.

Influence of Parameters. We evaluated two key parameters E_λ and β_{mar} in our modeling. As depicted in Fig. 3, the performance peaks at $E_\lambda = 20$ and $\beta_{mar} = 0.3$. When E_λ is set between 20 and 50, the performance remains relatively stable. It is worth noting that setting $E_\lambda = 0$ yields the worst performance. This is because the pseudo labels generated in the early stage of training are not accurate, and mining hard samples at this time introduces too much noise, resulting in decreased model performance.

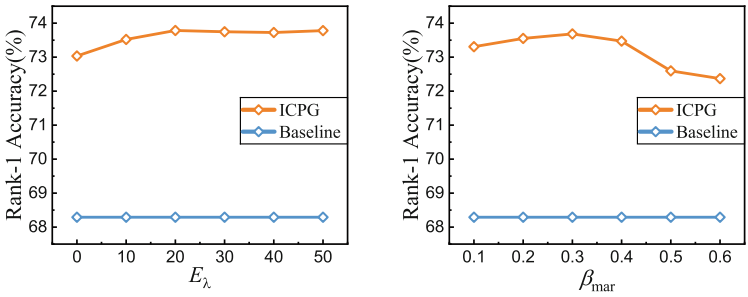


Fig. 3. Evaluation (%) of the hyperparameter E_λ and β_{mar} using Rank-1 accuracy on the CUHK-PEDES dataset.

Table 5. Comparison of using different pseudo labels in the CHM and CDM loss.

No.	Methods	Pseudo Label		CUHK-PEDES			ICFG-PEDES			RSTPReid		
		Image	Text	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
0	$\mathcal{L}_{chm}$	✓	✓	70.19	86.31	91.36	62.20	78.33	84.10	56.55	78.05	85.70
1			✓	70.06	86.50	91.42	62.32	78.48	84.10	57.00	78.85	86.60
2		✓		70.45	87.10	91.94	62.51	78.49	84.12	57.40	79.45	87.25
3	$\mathcal{L}_{cdm}$	✓	✓	70.26	86.91	91.80	59.88	78.00	84.04	56.60	**80.85**	88.10
4			✓	70.31	87.25	**92.66**	59.73	77.95	83.82	58.20	80.30	87.85
5		✓		70.37	87.41	92.24	60.72	78.08	84.12	58.75	80.65	**88.65**
6	$\mathcal{L}_{chm} + \mathcal{L}_{cdm}$	✓	✓	70.61	87.25	91.86	63.19	**79.30**	84.86	57.85	78.85	85.15
7			✓	71.08	87.36	92.12	63.12	79.27	84.81	57.00	78.85	86.60
8		✓		**71.16**	**87.41**	91.98	**63.23**	79.19	**84.86**	**59.95**	80.45	87.50

4.5 Pseudo Label Generation

In our study, we conducted extensive experiments to investigate the performance of various combinations of pseudo labels in the two proposed components. For the CHM module, comparing entries No. 0, 1 and 2 in Table 5, it is evident that using only image pseudo labels yielded the best performance across all metrics on the three datasets, outperforming the use of pseudo labels from both image and text modalities simultaneously, as well as using only text pseudo labels. Similarly, in the CDM module, comparing entries No. 3, 4 and 5 in Table 5 reveals superior performance with only image pseudo labels in most metrics across all three datasets. When both CDM and CHM modules were used simultaneously, comparing entries No. 6, 7 and 8 in Table 5, it was observed that using only image pseudo labels outperformed the other configurations in almost all metrics. These results indicate that, in the two modules we propose, using only image pseudo labels consistently yields optimal performance.

4.6 Qualitative Analysis

Figure 4 displays the top-10 retrieval images obtained using textual descriptions from the Baseline and our proposed ICPG on CUHK-PEDES dataset. Compared with the baseline method, ICPG achieves much more accurate retrieval

Fig. 4. Comparison of the top-10 retrieved results by the Baseline (first row) and our ICPG (second row) on the CUHK-PEDES dataset for each text query. The correctly retrieved images are marked with green rectangular boxes.

results and obtains accurate retrieval results when Baseline fails to retrieve them. This is mainly due to the Cross-Modal Hard Sample Mining (CHM) module we designed, which enhances the discriminative capacity of pedestrian features by exploring challenging hard negative samples, thereby effectively distinguishing different pedestrians with high similarity.

5 Conclusion

In this study, we only use image pseudo labels for jointly supervising both image and text modalities, with the aim of addressing the challenge of inconsistent pseudo label generation across modalities. We introduce a cross-modal distribution matching loss to amplify the correlation disparities between matched and unmatched pairs. To achieve further cross-modal alignment, we propose a cross-modal hard sample mining module to facilitate discriminative learning of textual-visual joint embeddings. Experimental results across three benchmark datasets have demonstrated the effectiveness of our proposed ICPG method.

Acknowledgement. This work was partially supported by the China Postdoctoral Science Foundation (2023M741305), the Fundamental Research Funds for the Central Universities (CCNU23XJ001).

References

1. Chen, T., Xu, C., Luo, J.: Improving text-based person search by spatial matching and adaptive threshold. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1879–1887. IEEE (2018)
2. Chen, Y., Huang, R., Chang, H., Tan, C., Xue, T., Ma, B.: Cross-modal knowledge adaptation for language-based person search. IEEE Trans. Image Process. **30**, 4057–4069 (2021)

3. Dai, Y., Liu, J., Bai, Y., Tong, Z., Duan, L.Y.: Dual-refinement: joint label and feature refinement for unsupervised domain adaptive person re-identification. IEEE Trans. Image Process. **30**, 7815–7829 (2021)
4. Dai, Z., Wang, G., Yuan, W., Zhu, S., Tan, P.: Cluster contrast for unsupervised person re-identification. In: Proceedings of the Asian Conference on Computer Vision (ACCV), pp. 1142–1160 (2022)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). arXiv:1810.04805
6. Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification (2021). arXiv:2107.12666
7. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, vol. 96, pp. 226–231 (1996)
8. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification. In: proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6112–6121 (2019)
9. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification (2020). arXiv:2001.01526
10. Ge, Y., Zhu, F., Chen, D., Zhao, R., et al.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. Adv. Neural. Inf. Process. Syst. **33**, 11309–11321 (2020)
11. Han, X., He, S., Zhang, L., Xiang, T.: Text-based person search with limited data (2021). arXiv:2110.10807
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification (2017). arXiv:1703.07737
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
15. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2787–2797 (2023)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv:1412.6980
17. Li, S., Cao, M., Zhang, M.: Learning semantic-aligned feature representation for text-based person search. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2724–2728. IEEE (2022)
18. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1890–1899 (2017)
19. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5187–5196 (2017)
20. Li, W., Tan, L., Dai, P., Zhang, Y.: Prompt decoupling for text-to-image person re-identification (2024). arXiv:2401.02173
21. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8738–8745 (2019)

22. Lin, Y., Xie, L., Wu, Y., Yan, C., Tian, Q.: Unsupervised person re-identification via softened similarity learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3387–3396 (2020)
23. Oord, A.V.D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2018). arXiv:1807.03748
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 8748–8763. PMLR (2021)
25. Shao, Z., Zhang, X., Ding, C., Wang, J., Wang, J.: Unified pre-training with pseudo texts for text-to-image person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11174–11184 (2023)
26. Shu, X., Wen, W., Wu, H., Chen, K., Song, Y., Qiao, R., Ren, B., Wang, X.: See finer, see more: Implicit modality alignment for text-based person retrieval. In: European Conference on Computer Vision, pp. 624–641. Springer (2022)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556
28. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: theory and practice. Pattern Recogn. **102**, 107173 (2020)
29. Wang, C., Luo, Z., Lin, Y., Li, S.: Text-based person search via multi-granularity embedding learning. In: IJCAI, pp. 1068–1074 (2021)
30. Wang, C., Luo, Z., Lin, Y., Li, S.: Improving embedding learning by virtual attribute decoupling for text-based person search. Neural Comput. Appl. 1–23 (2022)
31. Wang, C., Luo, Z., Zhong, Z., Li, S.: Divide-and-merge the embedding space for cross-modality person search. Neurocomputing **463**, 388–399 (2021)
32. Wang, Z., Fang, Z., Wang, J., Yang, Y.: Vitaa: visual-textual attributes alignment in person search by natural language. In: Computer Vision–ECCV 2020, pp. 402–420. Springer (2020)
33. Wang, Z., Zhu, A., Xue, J., Wan, X., Liu, C., Wang, T., Li, Y.: Caibc: capturing all-round information beyond color for text-based person retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 5314–5322 (2022)
34. Yan, S., Dong, N., Zhang, L., Tang, J.: Clip-driven fine-grained text-image person re-identification. IEEE Trans. Image Process. **32**, 6032–6046 (2023)
35. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2022)
36. Zeng, K., Ning, M., Wang, Y., Guo, Y.: Hierarchical clustering with hard-batch triplet loss for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13654–13662 (2020)
37. Zhao, S., Gao, C., Shao, Y., Zheng, W.S., Sang, N.: Weakly supervised text-based person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11395–11404 (2021)
38. Zheng, Y., Zhao, X., Lan, C., Zhang, X., Huang, B., Yang, J., Yu, D.: Cpcl: cross-modal prototypical contrastive learning for weakly supervised text-based person re-identification (2024). arXiv:2401.10011

39. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: Proceedings of the European conference on computer vision (ECCV), pp. 172–188 (2018)
40. Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., Hua, G.: Dssl: deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 209–217 (2021)