Graph-Based Cross-Domain Knowledge Distillation for Cross-Dataset Text-to-Image Person Retrieval

Bingjun Luo, Jinpeng Wang, Zewen Wang, Junjie Zhu*, Xibin Zhao,

BNRist, KLISS, and School of Software, Tsinghua University
luobingjun@gmail.com, {wjp21,wang-zw19,zhujj18}@mails.tsinghua.edu.cn, zxb@tsinghua.edu.cn

Abstract

Video surveillance systems are crucial components for ensuring public safety and management in smart city. As a fundamental task in video surveillance, text-to-image person retrieval aims to retrieve the target person from an image gallery that best matches the given text description. Most existing text-to-image person retrieval methods are trained in a supervised manner that requires sufficient labeled data in the target domain. However, it is common in practice that only unlabeled data is available in the target domain due to the difficulty and cost of data annotation, which limits the generalization of existing methods in practical application scenarios. To address this issue, we propose a novel unsupervised domain adaptation method, termed Graph-Based Cross-Domain Knowledge Distillation (GCKD), to learn the cross-modal feature representation for text-to-image person retrieval in a cross-dataset scenario. The proposed GCKD method consists of two main components. Firstly, a graph-based multi-modal propagation module is designed to bridge the cross-domain correlation among the visual and textual samples. Secondly, a contrastive momentum knowledge distillation module is proposed to learn the cross-modal feature representation using the online knowledge distillation strategy. By jointly optimizing the two modules, the proposed method is able to achieve efficient performance for cross-dataset text-to-image person retrieval. Extensive experiments on three publicly available text-to-image person retrieval datasets demonstrate the effectiveness of the proposed GCKD method, which consistently outperforms the state-of-the-art baselines.

Introduction

Person retrieval is a fundamental problem in the field of video surveillance, which has attracted extensive attention from both scientific research and practical applications (Jing et al. 2020; Li et al. 2023). Text-to-image person retrieval is a subtask of person retrieval, which aims to retrieve the target person from the image gallery that matches the given text description (Li et al. 2017; Yan et al. 2023). This task is associated with both image-text retrieval (Li et al. 2021) and image-based person retrieval (He et al. 2021). Different from image-based person retrieval, text-to-image person retrieval is able to retrieve the target person only based on the given query text, which is more user-friendly and easily accessible than using query image (Jiang and Ye 2023; Bai et al. 2023). Thus, text-to-image person retrieval has attracted increasing attention in recent years (Zhu et al. 2021; Shu et al. 2022; Zhu et al. 2024). With the advancement of deep learning technology (Fan et al. 2024; Gu et al. 2020), the performance of text-to-image person retrieval has significantly improved. Recently, large language models (Radford et al. 2021; Li et al. 2021; Shao et al. 2024) have further enhanced this performance.

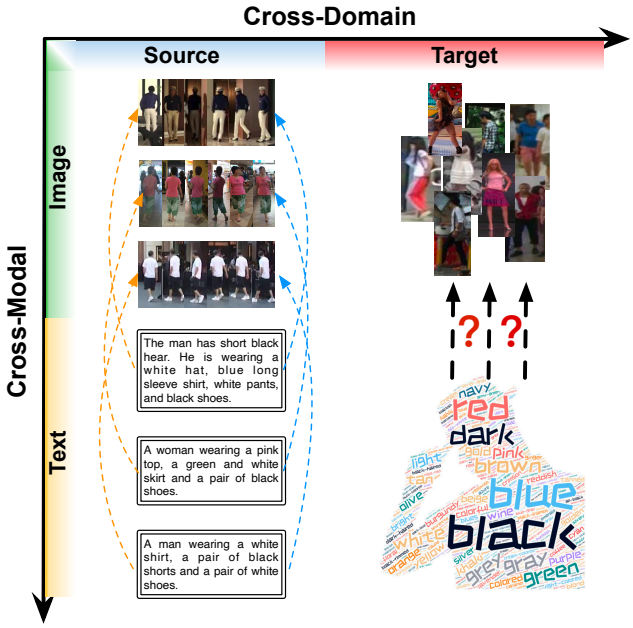

Figure 1: An illustration of cross-dataset text-to-image person retrieval task. The source dataset contains paired image-text annotations, while the target dataset only contains unpaired images and texts. This task faces both cross-domain shift and cross-modality gap challenges. Part of the figures comes from (Ding et al. 2021; Zhu et al. 2021, 2024).

However, most existing text-to-image person retrieval methods are trained in the supervised manner, which requires a large amount of high-quality annotated image-text alignment datasets in the target application scenario. These

*Corresponding author.

methods usually rely on large-scale alignment data collection and manual annotation, which are difficult to obtain and expensive. This limits the generalization of these methods in practical applications (Jiang and Ye 2023; Yan et al. 2023).

A feasible solution is to adopt the transfer learning paradigm, which transfers the knowledge learned from the existing large-scale supervised datasets to the target unlabeled dataset, i.e., cross-dataset text-to-image person retrieval as shown in Fig. 1. In the cross-dataset text-to-image person retrieval task, there are mainly two challenges: **1) Domain shift.** For the cross-domain transferring task, an important challenge is the domain shift, i.e., the distribution difference between the source and target datasets. **2) Modality gap.** Another important challenge for this task is the modality heterogeneity between visual and textual modalities (Jiang and Ye 2023). In the visual modality, data is usually high-dimensional and continuous, where differences mainly lie in the background, pose, lighting, etc. In contrast, in the textual modality, data is usually low-dimensional and discrete, where differences mainly lie in semantic ambiguity and order. In the cross-dataset scenario, due to the lack of text-image matching annotations on the target dataset, the modality heterogeneity will be further exacerbated. In the cross-dataset text-to-image person retrieval task, the superposition of these two challenges makes the performance of existing text-to-image person retrieval methods decrease in cross-domain scenarios, making this task more difficult.

To address the above challenges, we propose a novel unsupervised domain adaptation method for cross-dataset text-to-image person retrieval, named **Graph-based Cross-domain Knowledge Distillation (GCKD)**. The proposed method is based on the vision language pre-training models, which have shown great power in the field of text-to-image person retrieval recently (Yan et al. 2023; Bai et al. 2023; Jiang and Ye 2023). The proposed GCKD method consists of two novel components. The first component is a graph-based multi-domain propagation (GMP) module, which aims to address the domain shift problem by propagating feature information between the source and target domains on the dynamic cross-domain graph. The second component is a contrastive momentum knowledge distillation (CMKD) module, which aims to address the modality gap by constructing high-confidence pseudo text-image similarity labels through momentum knowledge distillation, which can guide the model to learn modal-invariant feature representation. By integrating these two modules, our method can effectively address the domain shift and modality gap challenges, and achieve more accurate and robust performance in the cross-dataset text-to-image person retrieval task.

In summary, the contributions of this paper are three-fold:

- We propose a novel unsupervised domain adaptation method, i.e. GCKD, to improve the cross-domain performance of text-to-image person retrieval. To the best of our knowledge, this is the first work to adopt the vision language pre-training models for cross-dataset text-to-image person retrieval.
- We introduce a cross-domain graph propagation mechanism to address the domain shift challenge, and a contrastive momentum knowledge distillation strategy to address the modality gap problem. These components are integrated into a unified framework to learn the cross-modal representation.
- We conduct extensive experiments on three commonly used datasets, and the results demonstrate that our proposed method outperforms the state-of-the-art methods in the cross-dataset text-to-image person retrieval task.

Related Work

Text-to-Image Person Retrieval

With the advancement of smart city (Xi et al. 2024a,b), the integration of text-to-image person retrieval is gaining increasing attention. The primary challenge in text-to-image person retrieval is cross-modal alignment, which can be categorized into two main strategies: cross-modal interaction-based and cross-modal interaction-free methods. Cross-modal interaction-based methods (Li et al. 2017; Zheng et al. 2020; Zhu et al. 2021) utilize attention mechanisms to identify local correspondences (e.g., patch-word, patch-phrase) between images and texts, predicting matching scores for image-text pairs. Notably, DSSL (Zhu et al. 2021) separates visual data into person and surroundings information for effective surroundings-person distinction. In contrast, cross-modal interaction-free methods (Chen et al. 2021; Bai et al. 2023; Shu et al. 2022) achieve high performance without complex interactions. The success of Transformer architectures in Vision and Language Tasks has led to the development of various Transformer-based models (Li, Cao, and Zhang 2022; Shao et al. 2022), such as LGUR (Shao et al. 2022), which learns granularity-unified representations for text and image modalities in an end-to-end manner.

While existing methods typically require labeled data for downstream tasks, our work operates in an unsupervised setting, eliminating the need for extensive alignment data collection and manual annotation, thus demonstrating greater efficiency in cross-domain scenarios.

Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) seeks to transfer knowledge from a labeled source domain to an unlabeled target domain, but its application in text-to-image person retrieval is limited. MAN (Jing et al. 2020) introduces a moment alignment network for cross-modal text-image alignment, but its reliance on CNN and LSTM limits its adaptability to Transformer architectures. POUF (Tanwisuth et al. 2023) aligns prototypes and target data in latent space using transport-based distribution alignment and mutual information maximization. ReCLIP (Hu et al. 2024) presents a source-free domain adaptation method for vision-language models through cross-modality self-training with learned pseudo labels. However, both POUF and ReCLIP face modality heterogeneity issues specific to text-to-image person retrieval tasks.

In contrast to these approaches, our model leverages vision-language pre-training for unsupervised domain adaptation in text-to-image person retrieval, effectively addressing challenges related to domain shift and modality gaps.

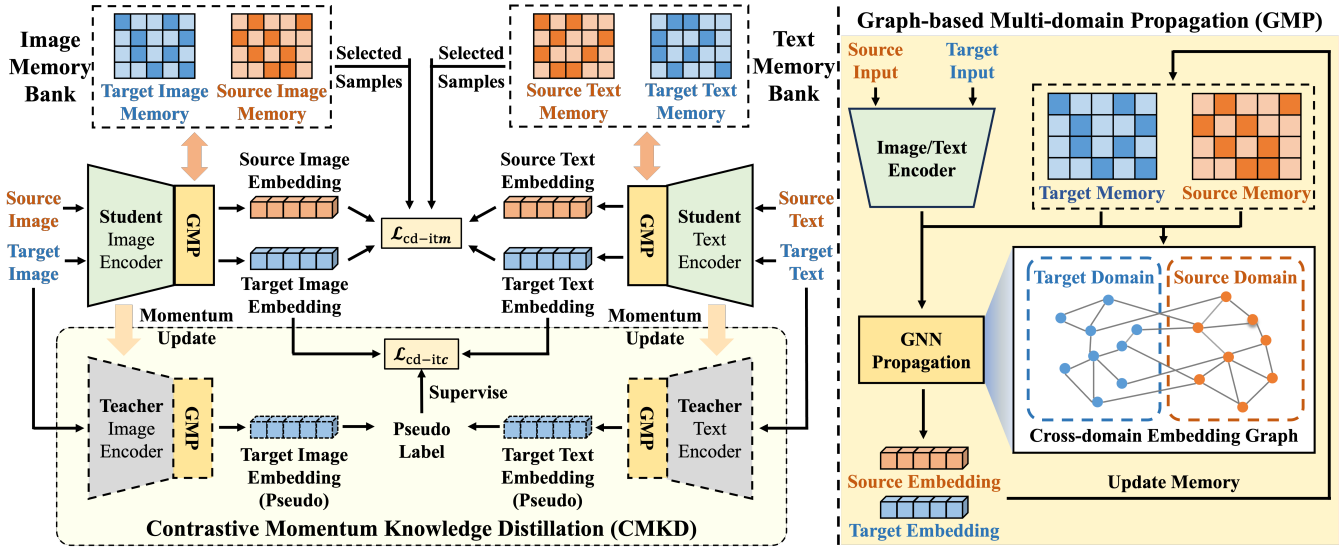

Figure 2: The main framework of the proposed Graph-based Cross-domain Knowledge Distillation (GCKD) method. The proposed method consists of two main components: Graph-based Multi-domain Propagation (GMP) module and Contrastive Momentum Knowledge Distillation (CMKD) module.

Problem Definition

Denote $\mathcal{D}_s = \{\mathcal{V}_s, \mathcal{T}_s\}$ as a well-annotated text-to-image person retrieval dataset from the source domain. The source dataset $\mathcal{D}_s$ consists of a labeled image set $\mathcal{V}_s = \{(\boldsymbol{v}_s^{(i)}, y_s^{(i)})\}_{i=1}^{M_s}$ of size M_s and a labeled text set $\mathcal{T}_s = \{(\boldsymbol{t}_s^{(j)}, y_s^{(j)})\}_{j=1}^{N_s}$ of size N_s respectively, where $\boldsymbol{v}_s^{(i)}$ and $y_s^{(i)}$ are the image and corresponding identity label of the i-th visual sample, and $\boldsymbol{t}_s^{(j)}$ and $y_s^{(j)}$ are the text and corresponding identity label of the j-th textual sample. Denote $\mathcal{D}_t = \{\mathcal{V}_t, \mathcal{T}_t\}$ as an unlabeled dataset from the target domain. The target dataset $\mathcal{D}_t$ only contains unpaired images $\mathcal{V}_t = \{\boldsymbol{v}_t^{(i)}\}_{i=1}^{M_t}$ and texts $\mathcal{T}_t = \{\boldsymbol{t}_t^{(j)}\}_{j=1}^{N_t}$, where $\boldsymbol{v}_t^{(i)}$ and $\boldsymbol{t}_t^{(j)}$ are the i-th image and j-th text sample in the target dataset, respectively. Given the labeled source dataset $\mathcal{D}_s$ and the unlabeled target dataset $\mathcal{D}_t$, the goal of cross-dataset text-to-image person retrieval is to learn a model that can effectively retrieve the target person image $\boldsymbol{v}_t$ for each given text query $\boldsymbol{t}_t$ in the target domain.

Our Model

In this section, we present the proposed Graph-based Cross-domain Knowledge Distillation (GCKD) method for cross-dataset text-to-image person retrieval.

Model Overview

In this paper, we propose a novel graph-based cross-dataset text-to-image person retrieval method, named **Graph-based Cross-domain Knowledge Distillation (GCKD)**. As shown in Fig. 2, the proposed GCKD method consists of two main components as the following. **Firstly**, a graph-based multi-domain propagation module is proposed to address the domain shift problem in the cross-modal retrieval task. By unifying visual and textual features from the source and target domains into a holistic cross-domain embedding graph, the module is designed to bridge the correlation and align the cross-modal features between the source and target domain. **Secondly**, a contrastive momentum knowledge distillation module is proposed to address the modality gap problem in the cross-domain scenario. Different from the single-domain visual language model, the module introduces momentum strategy into the domain adaptation task and proposes cross-domain fine-grained matching tasks to learn the shared representation among different modalities and domains. By jointly optimizing the two modules, the proposed model is able to achieve efficient performance for cross-dataset text-to-image person retrieval.

Graph-based Multi-domain Propagation

The domain shift problem, arising from variations in image conditions (e.g., resolutions, angles) and text styles (e.g., descriptive approaches, paragraph lengths) across datasets, significantly impacts the performance of text-to-image person retrieval models. To mitigate this, we propose a graph-based multi-domain propagation module that connects features from both source and target domains. By utilizing a unified graph, this module bridges correlations and reduces discrepancies across domains.

Embedding Memory The embedding memory stage is designed to store the embeddings from the source and target domains. We propose two types of memory banks to store the most recent C embedding of D-dimmension from the visual and textual modalities during training. The image memory bank is further composed of the source image memory $Q_{SI} \in \mathbb{R}^{C \times D}$ and $Q_{TI} \in \mathbb{R}^{C \times D}$, which stores the image embeddings of source and target domains respectively. The

setting is similar for the textual modal, which stores the most recent C text embeddings of source and target domains respectively, i.e., $Q_{ST} \in \mathbb{R}^{C\times D}$ and $Q_{TT} \in \mathbb{R}^{C\times D}$. At each training iteration, the image and text memories are updated iteratively from the recent embeddings of both source and target domains. By leveraging the memory banks and the corresponding updating mechanism, the module can capture the multi-modal embedding information across batches and even epochs, which effectively extends the scope of multi-domain propagation.

Graph Construction For each training iteration, we are given a batch of image/text embeddings $\boldsymbol{F}_m = (\boldsymbol{f}_m^{(1)}, \cdots, \boldsymbol{f}_m^{(K)}) \in \mathbb{R}^{B\times K}$ from target domains, where $m = I$ denotes image modality embeddings and $m = T$ denotes text modality embeddings. Based on the input embeddings and the existing source and target memories, the dynamic cross-domain graph $\mathcal{G}$ is constructed using the K-nearest neighbor (KNN) algorithm.

Specifically, the graph $\mathcal{G}$ can be determined by the vertex set $\mathcal{V}$, the vertex embedding matrix $\boldsymbol{X}$, and the adjacency matrix $\boldsymbol{A}$. The vertex of the graph comes from three parts:

$$\mathcal{V} = \mathcal{V}_{\text{input}} \cup \mathcal{V}_{\text{src\_mem}} \cup \mathcal{V}_{\text{tgt\_mem}} \tag{1}$$

where $\mathcal{V}_{\text{input}} = \{v_i\}_{i=1}^{B}$ is the vertex set of the given image/text batch $\boldsymbol{F}_m$, $\mathcal{V}_{\text{src\_mem}} = \{v_i\}_{i=B+1}^{B+C}$ and $\mathcal{V}_{\text{tgt\_mem}} = \{v_i\}_{i=B+C+1}^{B+2C}$ are the vertex set of the source and target memories Q_{Sm}, Q_{Tm}, respectively. The vertex embedding matrix $\boldsymbol{X} \in \mathbb{R}^{(B+2C)\times K}$ is constructed by concatenating the input embeddings and the embeddings from the source and target memories:

$$\boldsymbol{X} = (\boldsymbol{F}_m^\top, Q_{Sm}^\top, Q_{Tm}^\top)^\top \tag{2}$$

The adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{(B+2C)\times(B+2C)}$ is dynamically computed by the KNN algorithm based on the cosine similarity in vertex embedding matrix $\boldsymbol{X}$:

$$\boldsymbol{A}_{ij} = \begin{cases} 1, & \text{if } v_i \text{ is one of the } K \text{ nearest neighbors of } v_j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

By constructing a dynamic cross-domain graph $\mathcal{G}$ from the input embeddings and source and target memories, the module maps multi-domain embeddings into a unified graph structure, bridging multi-domain correlations.

GNN Propagation After constructing the dynamic cross-domain graph $\mathcal{G}$, we introduce a Graph Neural Network (GNN) to propagate embeddings across domains, effectively learning the graph structure and capturing sample correlations. Specifically, we adopt a two-layer GNN model for this propagation. The GNN propagation is performed by iteratively updating the vertex embedding matrix $\boldsymbol{X}$ based on the adjacency matrix $\boldsymbol{A}$ and the vertex embedding matrix $\boldsymbol{X}$. For the l-th layer of the GNN, the update rule can be formulated as:

$$\boldsymbol{X}^{(l+1)} = \text{GNNConv}(\boldsymbol{X}^{(l)}, \boldsymbol{A}; \boldsymbol{\Theta}^{(l)}) \tag{4}$$

where $\text{GNNConv}(\cdot)$ is the GNN convolution operation, and $\boldsymbol{\Theta}^{(l)}$ is the learnable parameters of the l-th layer.

By iteratively updating the vertex embedding matrix $\boldsymbol{X}$ for $L = 2$ layers, the last layer is used to generate the final vertex embedding matrix $\boldsymbol{X}^{(L)}$, which is further used to compute the final domain-aware embeddings $\overline{\boldsymbol{F}}_m = \boldsymbol{X}_{1:B}^{(L)}$ for the image/text samples in the target domain. By propagating the embeddings across the dynamic cross-domain graph $\mathcal{G}$, the module is able to bridge the correlation among the samples from both source and target domains and reduce the domain discrepancy in the cross-dataset scenario.

Contrastive Momentum Knowledge Distillation

The modality difference between text and image data creates a significant challenge known as the modality gap. Existing contrastive learning methods, primarily designed for single-domain visual-language tasks, often degrade in cross-domain scenarios (Li et al. 2021). To address this, we propose a novel contrastive momentum knowledge distillation module that combines cross-modal contrastive learning with cross-domain knowledge distillation.

Cross-domain Momentum Distillation As shown in Fig. 2, the backbone model consists of an image encoder and a text encoder, referred to as the student model $\text{En}_{\text{Student}}(\cdot, \boldsymbol{\Theta}_{\text{Student}})$ in knowledge distillation, which extracts visual and textual features from input samples. We also create a teacher model with the same structure, $\text{En}_{\text{Teacher}}(\cdot, \boldsymbol{\Theta}_{\text{Teacher}})$. Both models are initialized with pre-trained weights from the source dataset, which provide rich knowledge from the source domain.

During the training process, the parameters of the teacher model are not optimized by the gradient descent method, but are updated by the exponential moving average (EMA) mechanism as follows:

$$\boldsymbol{\Theta}_{\text{Teacher}} \leftarrow m\boldsymbol{\Theta}_{\text{Teacher}} + (1-m)\boldsymbol{\Theta}_{\text{Student}} \tag{5}$$

where $m \in [0, 1]$ is the momentum coefficient. By updating the teacher model with EMA, its parameters lag behind those of the student model, retaining more source domain knowledge. This allows the teacher model to generate pseudo labels that encapsulate source domain knowledge, guiding the student model to learn better representations in the target domain.

Cross-modal Contrastive Learning After generating pseudo labels from the teacher model, the next challenge is effective cross-domain contrastive learning for knowledge transfer and improved representations in the target domain. We propose a cross-modal image-text contrast loss that leverages source domain knowledge from the teacher model to enhance target domain representations.

Specifically, given a batch of paired source domain samples $(\boldsymbol{v}_s, \boldsymbol{t}_s)$, and unpaired target domain visual and textual samples $\boldsymbol{v}_t$ and $\boldsymbol{t}_t$, the student model is used to generate the pseudo target domain image features $\hat{\boldsymbol{f}}_{TI}$ and pseudo target domain text features $\hat{\boldsymbol{f}}_{TT}$. At the same time, the student model is used to extract the source domain image features $\boldsymbol{f}_{SI}$ and source domain text features $\boldsymbol{f}_{ST}$, and the domain-aware target domain image features $\boldsymbol{f}_{TI}$ and target domain

text features $\boldsymbol{f}_{TT}$. The cross-domain image-text contrast loss is defined as:

$$\begin{aligned}\mathcal{L}_{cd-itc} = &- \sum_{\boldsymbol{f}_{TI},\boldsymbol{f}_{TT}} s_{\text{i2t}} \log \frac{\exp(d(\boldsymbol{f}_{TI},\boldsymbol{f}_{TT})/\tau)}{\sum_{q\in Q_{TT}}\exp(d(\boldsymbol{f}_{TI},q)/\tau)} \\ &- \sum_{\boldsymbol{f}_{TT},\boldsymbol{f}_{TI}} s_{\text{t2i}} \log \frac{\exp(d(\boldsymbol{f}_{TT},\boldsymbol{f}_{TI})/\tau)}{\sum_{q\in Q_{TI}}\exp(d(\boldsymbol{f}_{TT},q)/\tau)}\end{aligned} \tag{6}$$

where

$$s_{\text{i2t}} = \frac{\exp(d(\hat{\boldsymbol{f}}_{TI},\hat{\boldsymbol{f}}_{TT})/\tau)}{\sum_{q_{TT}\in Q_{TT}}\exp(d(\hat{\boldsymbol{f}}_{TI},q_{TT})/\tau)} \tag{7}$$

$$s_{\text{t2i}} = \frac{\exp(d(\hat{\boldsymbol{f}}_{TT},\hat{\boldsymbol{f}}_{TI})/\tau)}{\sum_{q_{TI}\in Q_{TI}}\exp(d(\hat{\boldsymbol{f}}_{TT},q_{TI})/\tau)} \tag{8}$$

are the pseudo similarity targets generated by the teacher model, $d(\cdot,\cdot)$ is the cosine similarity function, τ is the temperature parameter, $\boldsymbol{Q}_{TT}, \boldsymbol{Q}_{TI}$ are the target text and image memory respectively. By introducing the memory banks, the contrastive loss function is extended to almost all samples in the target domain, enabling the model to explore more positive and negative samples and facilitating feature extraction.

Cross-domain Fine-grained Matching As noted in (Li et al. 2021), the image-text matching task is a binary classification problem, with positive pairs from the same identity. To enhance fine-grained matching in the target domain, we propose a cross-domain image-text matching task using target-domain positive pairs and cross-domain hard negative pairs. Positive pairs are generated from high-confidence pseudo labels from the teacher model, while negative pairs consist of the most challenging cross-domain pairs for the student model to distinguish.

For each target domain image feature $\boldsymbol{f}_{TI}$, the positive text feature $\boldsymbol{f}_{TT}$ is selected from the target text memory Q_{TT} only if $d(\boldsymbol{f}_{TI},\boldsymbol{f}_{TT}) > \delta$ where δ is a predefined threshold. The negative text feature $\boldsymbol{f}_{ST}$ is selected from the source domain feature batch if $\boldsymbol{f}_{TI}$ has the highest cosine similarity with $\boldsymbol{f}_{ST}$, i.e., $\boldsymbol{f}_{TI}$ and $\boldsymbol{f}_{ST}$ comes from different identities (even different domains) but have the highest similarity score. Given the image and text sample pair, the multimodal encoder in the backbone network is utilized to produce the binary matching probability output $\hat{p}(\cdot,\cdot)$ between the given pairs. After a softmax function, the cross-domain image-text matching loss is defined as:

$$\begin{aligned}\mathcal{L}_{cd-itm} = &- \sum_{(\boldsymbol{f}_{TI},\boldsymbol{f}_{TT})} \log \text{Softmax}(\hat{p}(\boldsymbol{f}_{TI},\boldsymbol{f}_{TT})) \\ &- \sum_{(\boldsymbol{f}_{TI},\boldsymbol{f}_{ST})} \log(1-\text{Softmax}(\hat{p}(\boldsymbol{f}_{TI},\boldsymbol{f}_{ST})))\end{aligned} \tag{9}$$

The overall optimization objective is as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{cd-itc} + \lambda_2\mathcal{L}_{cd-itm} + \lambda_3\mathcal{L}_{mlm} \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the coefficient hyperparameters for each loss term, $\mathcal{L}_{cd-itc}$ is the cross-domain image-text contrast loss defined in Eq. (6), $\mathcal{L}_{cd-itm}$ is the cross-domain fine-grained matching loss defined in Eq. (9), $\mathcal{L}_{mlm}$ is the masked language model loss proposed by (Li et al. 2021).

Experiment Setup

In this section, we introduce the experiment setup of this paper, including datasets, baselines, task settings, evaluation metrics, and implementation details.

Datasets

In this paper, we conduct experiments on three publicly available text-to-image person retrieval datasets: ICFG-PEDES, RSTPReid, and CUHK-PEDES. **ICFG-PEDES** (Ding et al. 2021) is the largest public dataset for text-to-image person retrieval, which consists of 54,522 images from 4,102 identities in total. **RSTPReid** (Zhu et al. 2021) is a newly released text-to-image person retrieval dataset. The dataset comprises 20,505 images from 4,101 identities. **CUHK-PEDES** (Li et al. 2017) is also a commonly used dataset in the text-to-image person retrieval field, which is composed of 5 parts from different scenarios: SSM, VIPER, CUHK01, CUHK03, and Market-1501.

Baselines

To comprehensively evaluate the proposed method, we follow recent works (Hao et al. 2023; Zhu et al. 2024) to select state-of-the-art baselines and assess them in two training settings: Source Only (SO) and Source and Target (ST).

For the **Source Only (SO)** setting, we select several state-of-the-art methods from the single-domain text-to-image person retrieval task, trained solely on the labeled source dataset and tested directly in the target domain. For the main experiments, the baselines include RaSa (Bai et al. 2023), APTM (Yang et al. 2023), IRRA (Jiang and Ye 2023), CFine (Yan et al. 2023), IVT (Shu et al. 2022). For the intra-dataset experiments on CUHK-PEDES, the baselines include EAIBC (Zhu et al. 2024), RaSa (Bai et al. 2023), SSAN (Ding et al. 2021), MIA (Niu et al. 2020), SCAN (Lee et al. 2018), CMPM-CMPC (Zhang and Lu 2018).

For the **Source and Target (ST)** setting, due to the lack of unsupervised domain adaptation methods for text-to-image person retrieval with available code, we select several state-of-the-art UDA baselines from related text-image multimodal tasks. These include POUF (Tanwisuth et al. 2023) and ReCLIP (Hu et al. 2024) for main experiments, and MAN (Jing et al. 2020), ECN (Zhong et al. 2019), and ADDA (Tzeng et al. 2017) for intra-dataset experiments on CUHK-PEDES.

Evaluation Metrics and Settings

To quantitatively evaluate models in the cross-dataset text-to-image person retrieval task, we adopt two common metrics: Recall at Rank-K (Rank-K) and Mean Average Precision (mAP), following existing works (Bai et al. 2023; Yan et al. 2023). Rank-K measures the proportion of target person images in the top K ($K = 1, 5, 10$) results, while mAP reflects the mean average precision of all results. Higher values for both metrics indicate better performance.

In the main cross-dataset experiments, we evaluate various baselines in the cross-dataset scenario for the above 3 datasets including ICFG-PEDES, RSTPReid, and CUHK-PEDES. For each dataset as the source set, we train the baselines on the source set and test them on the other two datasets

Train Set	Method	Ref	ICFG-PEDES → RSTPReid				ICFG-PEDES → CUHK-PEDES			
			Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Supervised	IRRA	CVPR'23	60.20	81.30	88.20	47.17	73.38	89.93	93.71	66.13
	APTM	MM'23	67.50	85.70	91.45	52.56	76.53	90.04	94.15	66.91
	RaSa	IJCAI'23	66.90	86.50	91.35	52.31	76.51	90.29	94.25	69.38
Source Only	IVT	ECCVW'22	43.70	65.10	75.55	37.65	22.63	42.29	52.36	19.85
	CFine	TIP'23	47.40	70.60	79.35	42.30	32.67	54.03	63.68	25.47
	IRRA	CVPR'23	45.10	69.20	78.75	36.76	33.43	56.11	66.23	31.38
	APTM	MM'23	52.50	75.15	81.70	40.81	46.44	66.89	74.45	40.14
	RaSa	IJCAI'23	55.00	73.65	81.55	46.18	48.65	69.90	76.53	42.03
Source & Target	POUF	ICML'23	36.60	61.80	73.15	28.28	20.50	39.39	49.03	18.48
	ReCLIP	WACV'24	50.35	73.25	81.20	42.75	33.48	55.25	64.26	29.19
	Ours	/	**59.95**	**79.05**	**85.95**	**49.68**	**52.70**	**72.76**	**80.15**	**45.97**

Table 1: Comparison results with state-of-the-art methods on ICFG-PEDES as source dataset. The performance in the supervised setting is reported for reference. The best results are emphasized in **bold**. The second-best results are noted by underline.

Train Set	Method	Ref	RSTPReid → ICFG-PEDES				RSTPReid → CUHK-PEDES			
			Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Supervised	IRRA	CVPR'23	63.46	80.25	85.82	38.06	73.38	89.93	93.71	66.13
	APTM	MM'23	68.51	82.99	87.56	41.22	76.53	90.04	94.15	66.91
	RaSa	IJCAI'23	65.28	80.40	85.12	41.29	76.51	90.29	94.25	69.38
Source Only	IVT	ECCVW'22	19.43	35.22	43.83	13.52	16.73	33.65	43.01	14.42
	CFine	TIP'23	24.83	39.49	46.92	17.04	20.79	38.76	48.93	15.77
	IRRA	CVPR'23	32.35	49.68	57.74	20.57	32.65	55.20	65.38	30.17
	APTM	MM'23	44.01	60.09	66.25	25.60	44.92	64.90	73.75	38.47
	RaSa	IJCAI'23	41.30	56.18	62.36	22.39	42.85	61.97	69.79	35.64
Source & Target	POUF	ICML'23	21.27	37.87	46.66	11.04	17.84	35.62	46.72	16.30
	ReCLIP	WACV'24	27.18	42.51	50.28	17.83	21.04	40.29	50.28	18.92
	Ours	/	**46.40**	**61.17**	**66.97**	**26.06**	**51.01**	**70.29**	**77.29**	**43.71**

Table 2: Comparison results with state-of-the-art methods on RSTPReid as source dataset. The performance in the supervised setting is reported for reference. The best results are emphasized in **bold**. The second-best results are noted by underline.

respectively. In the intra-dataset cross-domain experiments, we follow the cross-domain settings proposed by (Jing et al. 2020) within the CUHK-PEDES dataset. Specifically, we select SSM (S) as the source domain and consider 4 transfer tasks on CUHK03, Market-1501, VIPER, and CUHK01.

Implementation Details

The proposed model is implemented based on *PyTorch 1.10* framework on Python 3.8 and Ubuntu 20.04. For each dataset, the image-text data is split according to the existing protocol (Bai et al. 2023; Yang et al. 2023). ALBEF (Li et al. 2021) is adopted as the backbone of the vision language pre-training model and initialized with the pre-training weights on the source dataset. During training, the batch size is set to 4, and the optimizer is AdamW with an initial learning rate of $1e-5$ and cosine scheduler strategy. The hyperparameters of the proposed method are set as follows: the number of graph layers $L = 2$, the number of neighbors $K = 10$, the temperature $\tau = 0.07$, the momentum coefficient $\alpha = 0.999$, the loss coefficients $\lambda_1 = \lambda_2 = 0.5, \lambda_3 = 1$. All the experiments are conducted on NVIDIA GeForce RTX 4090.

Result and Analysis

In this section, we present the experimental results of our method and state-of-the-art baselines on the cross-dataset text-to-image person retrieval task. We also conduct ablation studies to analyze the effectiveness of each component.

Comparison with State-of-the-arts

To have a comprehensive evaluation of the proposed method, we compare it with state-of-the-art baselines on the cross-dataset text-to-image person retrieval task. As mentioned in Baselines, the compared baselines include the text-to-image person retrieval methods in the Source Only (SO) setting and the unsupervised domain adaptation methods in the Source and Target (ST) setting. The performance of the supervised training setting is also reported for reference. The results on ICFG-PEDES as the source dataset are shown in Table 1. The results on RSTPReid as the source dataset are shown in Table 2. The results on the CUHK-PEDES in intra-dataset cross-domain settings are shown in Table 3. From the results, we can make the following observations:

(1) The proposed method consistently outperforms the compared baselines on different transfer tasks. Compared with the second-best baseline, the proposed method achieves

Train Set	Method	S → C03		S → M		S → V		S → C01	
		Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
Source Only	CMPM-CMPC	42.30	69.20	63.40	85.10	57.80	84.70	44.80	70.90
	MIA	49.00	76.70	66.20	86.20	55.10	84.70	50.20	75.90
	SCAN	50.20	75.90	64.20	86.20	55.10	81.10	48.20	76.80
	SSAN	54.50	78.50	71.10	88.60	66.30	89.30	60.50	81.30
	EAIBC	55.10	79.60	72.50	89.40	67.40	91.30	62.40	81.90
	RaSa	<u>70.63</u>	<u>89.17</u>	<u>82.57</u>	<u>95.40</u>	<u>85.20</u>	<u>97.45</u>	<u>77.34</u>	<u>91.56</u>
Source & Target	SPGAN	44.70	72.50	63.30	85.30	60.70	85.70	45.30	71.20
	ADDA	45.10	72.80	63.90	85.70	61.40	86.00	45.70	71.60
	ECN	45.80	73.20	64.30	86.10	62.50	86.40	46.60	72.10
	MAN	48.50	74.80	65.10	87.40	64.20	87.20	48.20	73.20
	Ours	**71.25**	**89.90**	**83.54**	**95.52**	**86.73**	**98.47**	**78.91**	**92.03**

Table 3: Comparison results with state-of-the-art methods on intra-dataset cross-domain settings within CUHK-PEDES dataset. The best results are emphasized in **bold**. The second-best results are noted by <u>underline</u>.

Method	ICFG-PEDES → RSTPReid				ICFG-PEDES → CUHK-PEDES			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Baseline	55.00	73.65	81.55	46.18	48.65	69.90	76.53	42.03
CMKD	58.20	78.85	85.25	48.50	52.04	72.42	79.46	45.33
CMKD + GMP (Proposed)	**59.95**	**79.05**	**85.95**	**49.68**	**52.70**	**72.76**	**80.15**	**45.97**

Table 4: Results of ablation study on ICFG-PEDES as the source dataset. The best results are emphasized in **bold**.

4.37% improvement of Rank-1 Recall and 3.29% improvement of mAP on average. The results demonstrate that the proposed method can effectively address the domain shift and modality heterogeneity challenges, and successfully transfer the knowledge learned from the source dataset to the target dataset for better text-to-image person retrieval performance.

(2) Existing single-domain text-to-image person retrieval methods generally suffer from significant performance degradation in cross-dataset scenarios. For example, the SOTA methods, APTM and RaSa, suffer an average performance drop of 25.30% and 24.35% in Rank-1 Recall respectively. This is mainly due to the challenge of data distribution differences between the source and target domains. The domain shift between different datasets makes it difficult for the single-domain model to transfer the knowledge and generalize well to the target domain.

(3) Existing unsupervised domain adaptation methods generally underperform. In the *Source & Target* setting, these baselines consistently lag behind single-domain retrieval baselines. The large image gallery scale exacerbates modality heterogeneity challenges in text-to-image person retrieval, limiting UDA method performance.

Ablation Study

To analyze the effectiveness of each component in the proposed method, we conduct ablation studies on ICFG-PEDES and CUHK-PEDES as the source dataset respectively. Specifically, we add each component to the baseline step by step, train the model according to the same settings as the main experiments, and evaluate the model on the target dataset. Since the GMP module does not have any loss function and must rely on the training of the CMKD module, we evaluate the following three possible combinations:

- **Baseline:** The backbone only.
- **CMKD:** The model with the CMKD module.
- **CMKD + GMP (Proposed):** The model with both the CMKD and GMP modules, i.e. the proposed GCKD.

The ablation study results are shown in Table 4. As observed from the results, the baseline method achieves the lowest performance, which is due to the lack of domain adaptation. Adding the CMKD module to the baseline model improves the performance by a large margin, which demonstrates the effectiveness of the CMKD module in addressing the domain shift challenge. The proposed GMP module further improves the performance, which demonstrates the effectiveness of the proposed method in addressing the modality heterogeneity challenges. The results demonstrate the effectiveness of the proposed components in the cross-dataset text-to-image person retrieval task.

Conclusion

This paper presents a novel unsupervised domain adaptation method named Graph-Based Cross-Domain Knowledge Distillation (GCKD) for cross-dataset text-to-image person retrieval. In this method, a graph-based multi-modal propagation module and a contrastive knowledge distillation module are proposed to bridge the cross-domain correlation among the visual and textual samples and learn the cross-modal feature representation using the momentum knowledge distillation strategy. In the future, we plan to explore various advanced techniques to further improve cross-dataset retrieval performance, such as metric learning and adversarial learning.

Acknowledgments

This research is sponsored in part by the NSFC Program (No. U20A6003), Industrial Technology Infrastructure Public Service Platform Project "Public Service Platform for Urban Rail Transit Equipment Signal System Testing and Safety Evaluation" (No. 2022-233-225), Science and technology innovation project of Hunan Province (No.2023RC4014).

References

Bai, Y.; Cao, M.; Gao, D.; Cao, Z.; Chen, C.; Fan, Z.; Nie, L.; and Zhang, M. 2023. RaSa: relation and sensitivity aware representation learning for text-based person search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 555–563.

Chen, Y.; Huang, R.; Chang, H.; Tan, C.; Xue, T.; and Ma, B. 2021. Cross-modal knowledge adaptation for language-based person search. *IEEE Transactions on Image Processing*, 30: 4057–4069.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*.

Fan, Z.; Cong, W.; Wen, K.; Wang, K.; Zhang, J.; Ding, X.; Xu, D.; Ivanovic, B.; Pavone, M.; Pavlakos, G.; Wang, Z.; and Wang, Y. 2024. InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds. arXiv:2403.20309.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Hao, X.; Zhang, W.; Wu, D.; Zhu, F.; and Li, B. 2023. Dual alignment unsupervised domain adaptation for video-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18962–18972.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15013–15022.

Hu, X.; Zhang, K.; Xia, L.; Chen, A.; Luo, J.; Sun, Y.; Wang, K.; Qiao, N.; Zeng, X.; Sun, M.; et al. 2024. ReCLIP: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2994–3003.

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Jing, Y.; Wang, W.; Wang, L.; and Tan, T. 2020. Cross-modal cross-domain moment alignment network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10678–10686.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 201–216.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, S.; Cao, M.; and Zhang, M. 2022. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2724–2728. IEEE.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1970–1979.

Li, S.; Xu, X.; Shen, F.; and Yang, Y. 2023. Multi-granularity Separation Network for Text-Based Person Retrieval with Bidirectional Refinement Regularization. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 307–315.

Niu, K.; Huang, Y.; Ouyang, W.; and Wang, L. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29: 5542–5556.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shao, Z.; Xi, H.; Lu, H.; Wang, Z.; Bell, M. G.; and Gao, J. 2024. STLLM-DF: A Spatial-Temporal Large Language Model with Diffusion for Enhanced Multi-Mode Traffic System Forecasting. *arXiv preprint arXiv:2409.05921*.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th acm international conference on multimedia*, 5566–5574.

Shu, X.; Wen, W.; Wu, H.; Chen, K.; Song, Y.; Qiao, R.; Ren, B.; and Wang, X. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, 624–641. Springer.

Tanwisuth, K.; Zhang, S.; Zheng, H.; He, P.; and Zhou, M. 2023. POUF: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *International Conference on Machine Learning*, 33816–33832. PMLR.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

Xi, H.; Nelson, J. D.; Hensher, D. A.; Hu, S.; Shao, X.; and Xie, C. 2024a. Evaluating travel behavior resilience

across urban and rural areas during the COVID-19 pandemic: contributions of vaccination and epidemiological indicators. *Transportation research part A: policy and practice*, 180: 103980.

Xi, H.; Wang, Y.; Shao, Z.; Zhang, X.; and Waller, T. 2024b. Optimizing mobility resource allocation in multiple MaaS subscription frameworks: a group method of data handling-driven self-adaptive harmony search algorithm. *Annals of Operations Research*, 1–29.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2023. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*.

Yang, S.; Zhou, Y.; Zheng, Z.; Wang, Y.; Zhu, L.; and Wu, Y. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4492–4501.

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

Zheng, K.; Liu, W.; Liu, J.; Zha, Z.-J.; and Mei, T. 2020. Hierarchical gumbel attention network for text-based person search. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3441–3449.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 598–607.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 209–217.

Zhu, A.; Wang, Z.; Xue, J.; Wan, X.; Jin, J.; Wang, T.; and Snoussi, H. 2024. Improving Text-Based Person Retrieval by Excavating All-Round Information Beyond Color. *IEEE Transactions on Neural Networks and Learning Systems*.