Look Before You Leap: Improving Text-based Person Retrieval by Learning A Consistent Cross-modal Common Manifold

Zijie Wang
Nanjing Tech University
Nanjing, China
zijiewang9928@gmail.com

Aichun Zhu*
Nanjing Tech University
Nanjing, China
aichun.zhu@njtech.edu.cn

Jingyi Xue
Nanjing Tech University
Nanjing, China
jyx981218@163.com

Xili Wan
Nanjing Tech University
Nanjing, China
xiliwan@njtech.edu.cn

Chao Liu
Jinling Institute of Technology
Nanjing, China
liuchao@jit.edu.cn

Tian Wang
Beihang University
Beijing, China
wangtian@buaa.edu.cn

Yifeng Li
Nanjing Tech University
Nanjing, China
lyffz4637@163.com

ABSTRACT

The core problem of text-based person retrieval is how to bridge the heterogeneous gap between multi-modal data. Many previous approaches contrive to learning a latent common manifold mapping paradigm following a **cross-modal distribution consensus prediction (CDCP)** manner. When mapping features from distribution of one certain modality into the common manifold, feature distribution of the opposite modality is completely invisible. That is to say, how to achieve a cross-modal distribution consensus so as to embed and align the multi-modal features in a constructed cross-modal common manifold all depends on the experience of the model itself, instead of the actual situation. With such methods, it is inevitable that the multi-modal data can not be well aligned in the common manifold, which finally leads to a sub-optimal retrieval performance. To overcome this **CDCP dilemma**, we propose a novel algorithm termed LBUL to learn a Consistent Cross-modal Common Manifold (C^{3}M) for text-based person retrieval. The core idea of our method, just as a Chinese saying goes, is to '*san si er hou xing*', namely, to **Look Before yoU Leap (LBUL)**. The common manifold mapping mechanism of LBUL contains a looking step and a leaping step. Compared to CDCP-based methods, LBUL considers distribution characteristics of both the visual and textual modalities before embedding data from one certain modality into C^{3}M to achieve a more solid cross-modal distribution consensus, and hence achieve a superior retrieval accuracy. We evaluate our proposed method on two text-based person retrieval datasets CUHK-PEDES and RSTPReid. Experimental results demonstrate that the proposed LBUL outperforms previous methods and achieves the state-of-the-art performance.

CCS CONCEPTS

• **Information systems → Image search**; • **Computing methodologies → Object identification**.

KEYWORDS

person retrieval, text-based person re-identification, cross-modal retrieval

ACM Reference Format:
Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022. Look Before You Leap: Improving Text-based Person Retrieval by Learning A Consistent Cross-modal Common Manifold. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548166

1 INTRODUCTION

Given a textual description query, text-based person retrieval aims to identify images of the corresponding pedestrian from a large-scale image database. Compared to the currently active research topic image-based person retrieval (aka. person re-identification)[8, 33, 34] which utilizes image-based queries, text-based queries are much easier to access in the realistic application scenarios. Due to its effectiveness and applicability, text-based person retrieval [10, 12, 13, 16, 18, 21, 29–31] has drawn more and more attention. However, the study of this task is still in its infancy and there is still plenty of room for further research.

The core problem of text-based person retrieval is how to bridge the heterogeneous gap between multi-modal data. As different modalities are diverse and inconsistent in data form and distribution, it is not so easy to directly measure the cross-modal affinity. Many of the previous approaches contrive to learning a common

*Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548166

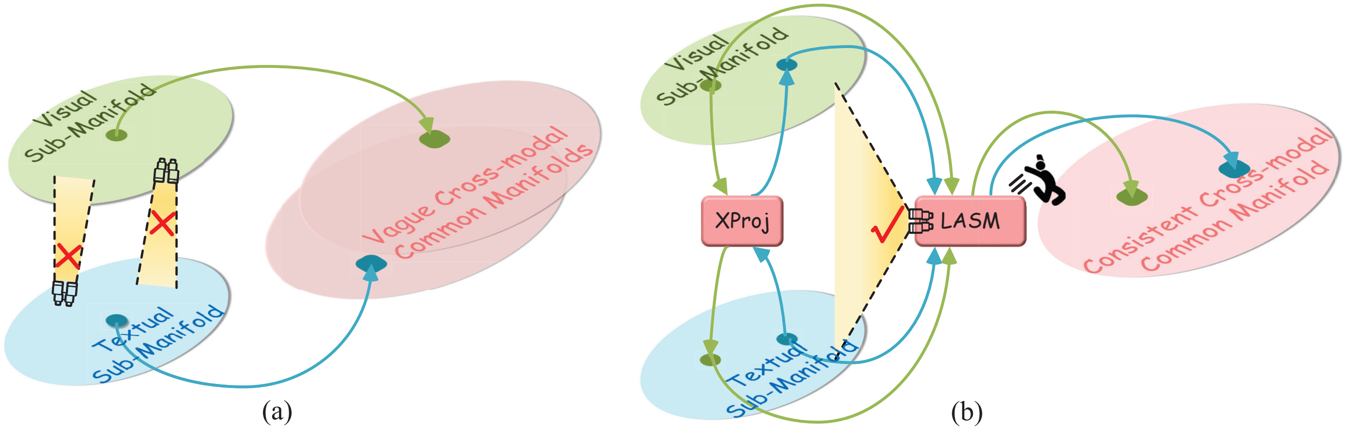


Figure 1: (a) For CDCP-based paradigms, when mapping features from distribution of one certain modality into the common manifold, feature distribution of the opposite modality is completely invisible. That is to say, how to achieve the cross-modal distribution consensus all depends on the experience of the model itself, instead of the actual situation. Consequently, there may exist multiple vague cross-modal common manifolds, which are adjacent to each other but not exactly identical. (b) For either a visual or textual sample, LBUL embeds it into C^{3}M after considering the distribution characteristics of both modalities to achieve a more solid cross-modal distribution consensus, instead of blindly predicting.

manifold mapping paradigm, either implicitly with separate submodels and extra constrains (e.g. attention mechanism), or explicitly with shared mapping blocks. These methods aim to transform heterogeneous multi-modal data into homogeneous feature representations, between which the cross-modal similarity can be calculated. However, most existing paradigms are proposed following a **cross-modal distribution consensus prediction (CDCP)** manner, which have their limitations. Specifically, as the multimodal data are from specific distribution of each modality, the process of the common manifold mapping can be deemed as trying to achieve a distribution consensus between the visual and textual modalities, so as to embed and align the multi-modal features in a constructed cross-modal common manifold. Nevertheless, when mapping features from distribution of one certain modality into the common manifold, feature distribution of the opposite modality is completely invisible. That is to say, how to achieve the cross-modal distribution consensus all depends on the experience of the model itself, instead of the actual situation. Consequently, as shown in Fig. 1 (a), there may exist multiple vague cross-modal common manifolds, which are adjacent to each other but not exactly identical. With such methods, it is inevitable that the multi-modal data will not be perfected aligned and matched with each other in a proper common manifold, which finally leads to a sub-optimal retrieval performance. This situation can be called a **CDCP dilemma**.

To overcome this CDCP dilemma, we consider to come up with a more effective common manifold mapping paradigm. In this paper, we propose a novel algorithm to learn a Consistent Cross-modal Common Manifold (C^{3}M) for text-based person retrieval. As illustrated in Fig. 1 (b), for either a visual or textual sample, our proposed method embeds it from one certain modality into C^{3}M after considering the distribution characteristics of both the visual and textual modalities to achieve a more solid cross-modal distribution consensus, instead of blindly predicting. The core idea of our method, just as a Chinese saying goes, is 'san si er hou xing', namely, to **Look Before yoU Leap**. So we name our proposed method **LBUL**, of which the common manifold mapping paradigm includes two steps, namely, a **looking step** and a **leaping step**. Compared with CDCP-based paradigms, LBUL is capable of achieving a more precise cross-modal distribution consensus. As a result, the multi-modal data can be embedded and aligned in a consistent common manifold with less information loss and higher accuracy, and hence achieve a superior retrieval performance. Specifically, with a proposed Uni-modal Sub-manifold Embedding Module (USEM), multi-granular features extracted from each modality are first distilled as a unified feature, which is embedded in a corresponding uni-modal sub-manifold (visual or textual). Then at the looking step of LBUL, in order to see distributions of both modalities before mapping data from one certain modality into C^{3}M, the uni-modal feature is projected into the opposite sub-manifold to give the distribution characteristics of the other modality by means of a Cross-modal Projection (XProj) module. In XProj, a Distribution Shifting ($\mathcal{DS}$) mechanism plays a key role in the statistical transformation of the feature according to the target modality, and thus enabling a proper feature projection. After the projection, for data from one certain modality, there exists two feature representations embedded in both the visual and textual modalities. Then at the leaping step, these two representations are processed together by a Leaping After Seeing Module (LASM), which conducts the common manifold mapping operation after seeing the distribution characteristics of both modalities. Through LASM, a consistent common representation can be obtained in C^{3}M for each sample, so that the cross-modal similarity can be properly measured. We evaluate our proposed method on two text-based person retrieval datasets including CUHK-PEDES [13] and RSTPReid [40]. Experimental results demonstrate that LBUL outperforms previous methods and achieves the state-of-the-art performance.

The main contributions of this paper can be summarized as threefold:

- A novel LBUL method is proposed to learn a Consistent Cross-modal Common Manifold ($\mathrm{C}^3\mathrm{M}$) for text-based person retrieval, which embeds data from one certain modality into $\mathrm{C}^3\mathrm{M}$ after considering the distribution characteristics of both the visual and textual modalities to achieve a more solid cross-modal distribution consensus, instead of blindly predicting.
- A two-step common manifold mapping mechanism which includes a looking step and a leaping step is proposed. By conducting the common manifold mapping operation after seeing the distribution characteristics of both modalities, LBUL is capable of learning consistent common representations with less information loss and higher retrieval accuracy.
- Extensive experimental analysis is carried out on CUHK-PEDES [13] and RSTPReid [40] to evaluate the proposed LBUL method for text-based person retrieval. Experimental results demonstrate that LBUL significantly outperforms existing methods and achieves the state-of-the-art performance.

2 RELATED WORKS

2.1 Person Re-identification

Person re-identification has drawn increasing attention in both academical and industrial fields. This technology addresses the problem of matching pedestrian images across disjoint cameras. The key challenges lie in the large intra-class and small inter-class variation caused by different views, poses, illuminations, and occlusions. Existing methods can be grouped into handed-crafted descriptors, metric learning methods and deep learning methods. With the development of deep learning [19, 22, 27, 37], deep learning methods are in general playing a major role in current state-of-the-art works. Yi et al. [34] firstly proposed deep learning methods to match people with the same identification. To boost the ReID model training efficiency in multi-label classification, Wang et al. [26] further proposed the memory-based multi-label classification loss (MMCL). MMCL works with memory-based non-parametric classifier and integrates multi-label classification and single-label classification in an unified framework. Jin et al. [9] introduce a global distance-distributions separation (GDS) constraint over two distributions to encourage the clear separation of positive and negative samples from a global view. Yuan et al. [35] propose a Gabor convolution module for deep neural networks based on Gabor function, which has a good texture representation ability and is effective when it is embedded in the low layers of a network. Taking advantage of the hinge function, they also design a new regularizer loss function to make the proposed Gabor Convolution module meaningful. A model that has joint weak saliency and attention aware is presented by Ning et al. [17], which can obtain more complete global features by weakening saliency features. In recent years, methods for unsupervised person re-identification have gradually emerged. Unsupervised person re-identification means that the target data set is unlabeled but the auxiliary source data set is not necessarily unlabeled [5]. Existing unsupervised person ReID works can be concluded into three categories. The first category utilizes hand-craft features [14]. But the features made by hand can not be robust and discriminative. To solve this problem, second category [6] adopts clustering to estimate pseudo labels to train the CNN. However, these methods require good trained model. Recently, the third category is proposed, which improves unsupervised person ReID by using transfer learning. Some works [15, 28] utilize transfer learning and minimize the attribute-level discrepancy by using extra attribute annotations.

2.2 Text-based Person Retrieval

Text-based person retrieval aims to search for the corresponding pedestrian image according to a given text query. This task is first introduced by Li et al. [13] and a GNA-RNN model is employed to handle the multi-modal data. Later, an efficient patch-word matching model [3] is proposed to capture the local similarity between image and text. Jing et al. [10] utilize pose information as soft attention to localize the discriminative regions. Niu et al. [18] adopt a Multi-granularity Image-text Alignments (MIA) model exploit the combination of multiple granularities. Nikolaos et al. [21] design a Text-Image Modality Adversarial Matching approach (TIMAM) to learn modality-invariant feature representation by means of adversarial and cross-modal matching objectives. Liu et al. [16] generate fine-grained structured representations from images and texts of pedestrians with an A-GANet model to exploit semantic scene graphs. CMAAM is introduced by Aggarwal et al. [1] which learns an attribute-driven space along with a class-information driven space by introducing extra attribute annotation and prediction. Zheng et al. [38] propose a Gumbel attention module to alleviate the matching redundancy problem and a hierarchical adaptive matching model is employed to learn subtle feature representations from three different granularities. Zhu et al. [40] proposed a Deep Surroundings-person Separation Learning (DSSL) model to effectively extract and match person information. Besides, they construct a Real Scenarios Text-based Person Re-identification (RSTPReid) dataset based on MSMT17 [32] to benefit future research on text-based person retrieval. Most of the above mentioned approaches are proposed following a cross-modal distribution consensus prediction (CDCP) manner. By means of the LBUL mechanism, a more solid cross-modal distribution consensus can be achieved and hence a more consistent cross-modal common manifold can be constructed.

3 METHODOLOGY

3.1 Problem Formulation

The goal of the proposed framework (shown in Fig. 2) is to measure the similarity between multi-modal data, namely, a given textual description query and a gallery person image. Formally, let $D = \{p_i, q_i\}_{i=1}^{N}$ denotes a dataset consists of N image-text pairs. Each pair contains a pedestrian image p_i captured by one certain surveillance camera and its corresponding textual description query q_i. The IDs of the Q pedestrians in the dataset are denoted as $Y = \{y_i\}_{i=1}^{Q}$. Given a textual description, the aim is to identify images of the most relevant pedestrian from a large scale person image gallery.

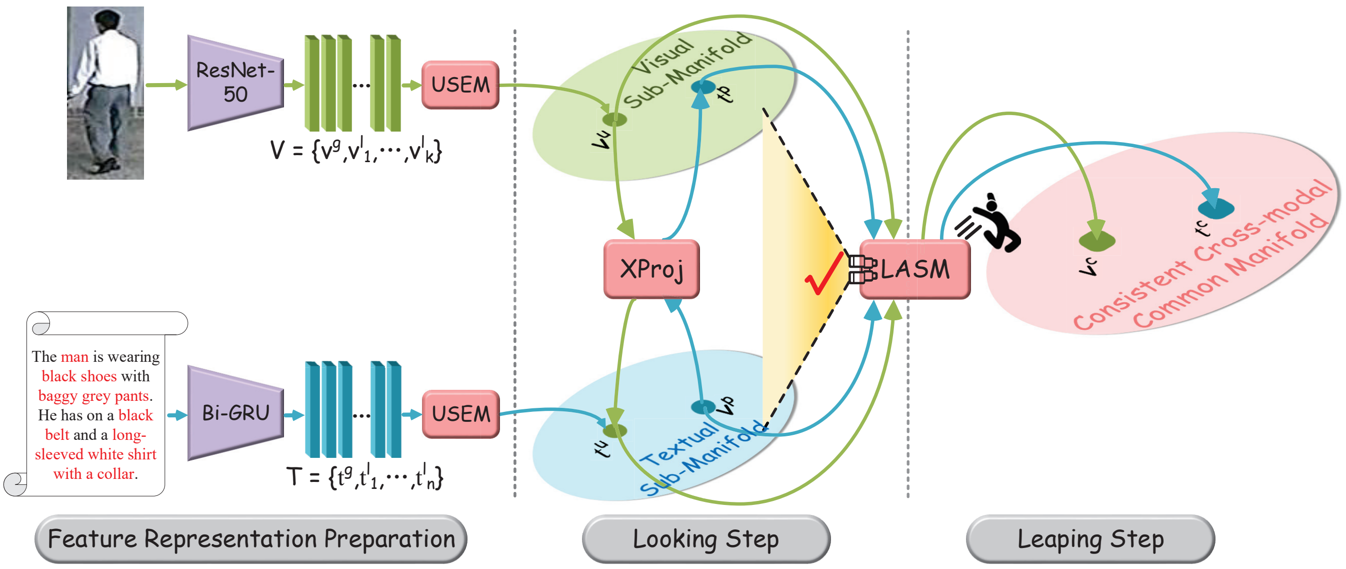

ResNet-50
USEM
$V = \{v^g, v^l_1, \cdots, v^l_k\}$
The man is wearing black shoes with baggy grey pants. He has on a black belt and a long-sleeved white shirt with a collar.
Bi-GRU
USEM
$T = \{t^g, t^l_1, \cdots, t^l_n\}$
Visual Sub-Manifold
v^u
t^p
XProj
Textual Sub-Manifold
t^u
v^p
LASM
v^c
t^c
Consistent Cross-modal Common Manifold
Feature Representation Preparation
Looking Step
Leaping Step

Figure 2: The overall framework of the proposed LBUL model.

3.2 Feature Representation Preparation

3.2.1 *Representation Extraction*.

Visual Representation Extraction. To extract multi-granular visual representations from a given image I, a pretrained ResNet-50 [7] backbone is utilized. To obtain the global representation $v^g \in \mathbb{R}^p$, the feature map before the last pooling layer of ResNet-50 is down-scaled to $1 \times 1 \times 2048$ with an average pooling layer and converted into a 2048-dim vector. Then it is passed through a group normalization (GN) layer followed by a fully-connected (FC) layer and transformed to p-dim. In the local branch, the same feature map is first horizontally k-partitioned by pooling it to $k \times 1 \times 2048$, and then the local strips are separately passed through a GN and two FCs with an ELU layer between them to form k p-dim vectors $V^l = \{v^l_1, v^l_2, \cdots, v^l_k\}$, which are finally concatenated with each other along with v^g to obtained the visual representation matrix $V = \{v^g, v^l_1, v^l_2, \cdots, v^l_k\} \in \mathbb{R}^{p \times (k+1)}$.

Textual Representation Extraction. For textual representation extraction, a whole sentence along with n phrases extracted from it is taken as textual materials, which is processed by a bi-directional Gated Recurrent Unit (bi-GRU). The last hidden states of the forward and backward GRUs are concatenated to give global/local $2p$-dim feature vectors. And then the $2p$-dim vector got from the whole sentence is passed through a GN followed by an FC to form the global textual representation $t^g \in \mathbb{R}^p$. With each certain input phrase, the corresponding output p-dim vector is handled consecutively by a GN and two FCs with an ELU layer between them. The obtained local vectors $T^l = \{t^l_1, t^l_2, \cdots, t^l_n\}$ are then concatenated with each other along with t^g to form the final textual representation matrix $T = \{t^g, t^l_1, t^l_2, \cdots, t^l_n\} \in \mathbb{R}^{p \times (n+1)}$.

3.2.2 *Uni-modal Sub-manifold Learning*.

After the raw visual and textual representation matrices V and T are obtained, we adopt a Uni-modal Sub-manifold Embedding Module (USEM) based on the self-attention mechanism [23] to distill both global and fine-grained local discriminative information into a unified feature, which can be formulated as

$$v^u = USEM(V) = \sum_{i=1}^{k} \frac{exp(v^g v^l_i)}{\sum_{j=1}^{k} exp(v^g v^l_j)} v^l_i + v^g, \tag{1}$$

$$t^u = USEM(T) = \sum_{i=1}^{n} \frac{exp(t^g t^l_i)}{\sum_{j=1}^{n} exp(t^g t^l_j)} t^l_i + t^g, \tag{2}$$

where v^u or t^u is the distilled unified visual/textual feature and is embedded into the corresponding visual/textual sub-manifold, respectively.

3.3 Look Before You Leap

Now that the uni-modal representations are obtained and each uni-modal sub-manifold which reveals the latent distribution characteristics of the corresponding modality is constructed, a two-step common manifold mapping mechanism including a looking step and a leaping step is proposed, which aims to embed the multi-modal data into the consistent cross-modal common manifold C^{3}M with less information loss and higher accuracy.

3.3.1 *Looking Step: Cross-modal Projection Module*.

At the looking step, the prime target is to see distributions of both modalities before mapping data from one certain modality into C^{3}M. To achieve this goal, a Cross-modal Projection (XProj) module is proposed to project one certain uni-modal representation into the opposite sub-manifold, which can be generally formulated as

$$s^p = XProj(s^u, r^u), \tag{3}$$

where s^u or r^u denotes the input source or target modal representation, respectively. s^p is the projected representation, which is embedded into the target sub-manifold.

Specifically, a Distribution Shifting ($\mathcal{DS}$) mechanism is first adopted to conduct the statistical transformation of the source modal representation according to the target modality:

$$s^{ds} = \mathcal{DS}(s^u, r^u) = \sigma(r^u)(\frac{s^u - \mu(s^u)}{\sigma(s^u)}) + \mu(r^u), \tag{4}$$

where s^{ds} is the shifted feature. $\mu(\cdot)$ and $\sigma(\cdot)$ denote the calculation of mean and variance, respectively. Then a multi-layer perceptron ($\mathcal{MLP}$) with a tanh activation layer is employed to embed s^{ds} into the target sub-manifold:

$$s^p = \mathcal{MLP}(s^{ds}). \tag{5}$$

With XProj, representation lying in one certain uni-modal sub-manifold can be properly projected into the opposite sub-manifold:

$$v^p = XProj(v^u, t^u),\ t^p = XProj(t^u, v^u). \tag{6}$$

3.3.2 *Leaping Step: Leaping After Seeing Module.*

After the looking step, for data from one certain modality, there exists two feature representations embedded in both the visual and textual modalities, which give the distribution characteristics of both modalities. Then at the leaping step, the two representations can be utilized by LBUL with a proposed Leaping After Seeing Module (LASM) to conduct the common manifold mapping operation after seeing distributions of both modalities. The mechanism of LASM can be formulated as

$$x^c = LASM(x^u, x^p), \tag{7}$$

where x^u and x^p are visual/textual representations before and after processed by XProj, respectively, while x^c is the visual/textual consistent common manifold representation embedded in C^{3}M.

To be specific, the two input representations are first utilized to estimate a fusion gate $g \in \mathbb{R}^p$:

$$g = \sigma(W_2 ELU(W_1(x^u \oplus x^p))), \tag{8}$$

where $\oplus$ is the feature concatenation operation and can be implemented as several other methods (e.g. addition or concatenation), which will be further discussed along with some substitution variants of LASM in Sec. 4.2.3. Here, $W_1 \in \mathbb{R}^{2p \times 2p}$ and $W_2 \in \mathbb{R}^{2p \times p}$ denote linear transformations without bias while $\sigma(\cdot)$ stands for the sigmoid activation function. Then x^c is obtained through the weighted summation of x^u and x^p according to g:

$$x^c = gx^u + (1 - g)x^p. \tag{9}$$

To sum up, the corresponding visual and textual consistent cross-modal common manifold representations v^c and t^c can be obtained as

$$v^c = LASM(v^u, v^p),\ t^c = LASM(t^u, t^p). \tag{10}$$

3.4 Similarity for Inference

For test and inference, several kinds of similarity scores are calculated. First, the similarity sim^c between a pair of visual/textual representations embedded in C^{3}M is calculated:

$$sim^c = cos(v^c, t^c), \tag{11}$$

where $cos(\cdot, \cdot)$ denotes the cosine similarity between two feature vectors. To further improve the retrieval performance, the global and fine-grained similarities are also utilized as an auxiliary. The global similarity sim^g between a pair of global multi-modal representations is computed as

$$sim^g = cos(v^g, t^g). \tag{12}$$

To obtain the fine-grained similarity sim^f, a cross-modal attention ($\mathcal{CA}$) mechanism is adopted:

$$\alpha_i^X = \frac{exp(cos(x_i^l, y^g))}{\sum_j exp(cos(X_j^l, y^g))}, \tag{13}$$

$$x^f = \mathcal{CA}(y^g, X^l) = \sum_{\alpha_i^X > \gamma} \alpha_i^X x_i^l, \tag{14}$$

where (X, Y) can be (V, T) or (T, V) while (x, y) can be (v, t) or (t, v). γ is a threshold value. Thus, sim^f can be calculated by

$$v^f = \mathcal{CA}(t^g, V^l),\ t^f = \mathcal{CA}(v^g, T^l), \tag{15}$$

$$sim^f = \frac{cos(v^g, t^f) + cos(v^f, t^g)}{2}. \tag{16}$$

Eventually, we can get the overall similarity sim as

$$sim = sim^c + \lambda_1 sim^g + \lambda_2 sim^f. \tag{17}$$

3.5 Optimization

To optimize LBUL, the ranking loss is employed to constrain the matched pairs to be closer than the mismatched ones in a mini-batch with a margin β:

$$\begin{aligned} L_{rk}(x_1, x_2) = &\sum_{\widehat{x_2}} max\{\beta - cos(x_1, x_2) + cos(x_1, \widehat{x_2}), 0\} \\ &+ \sum_{\widehat{x_1}} max\{\beta - cos(x_1, x_2) + cos(\widehat{x_1}, x_2), 0\}, \end{aligned} \tag{18}$$

where $(x_1, \widehat{x_2})$ or $(\widehat{x_1}, x_2)$ denotes a mismatched pair while (x_1, x_2) is a matched pair. Besides, the identification (ID) loss is also adopted:

$$L_{id}(x) = -log(softmax(W_{id}x), \tag{19}$$

where $W_{id} \in \mathbb{R}^{Q \times p}$ is a shared FC layer without bias while Q is the number of different pedestrians.

The overall optimization process of LBUL includes two stages. Before conducting the common manifold mapping operation, it is necessary to ensure each learned sub-manifold is solid. Therefore, in the first stage, the ranking loss and ID loss are utilized to optimize the extracted global and fine-grained local features:

$$\mathcal{L}^g = L_{id}(v^g) + L_{id}(t^g) + L_{rk}(v^g, t^g), \tag{20}$$

$$\mathcal{L}^f = L_{id}(v^f) + L_{id}(t^f) + L_{rk}(v^g, t^f) + L_{rk}(v^f, t^g), \tag{21}$$

$$\mathcal{L}^{Stage1} = \mathcal{L}^g + \lambda_3 \mathcal{L}^f. \tag{22}$$

In the second stage, first the ranking loss between the projected feature and the unified feature in the target modality is computed to ensure a reliable projection. Besides, the ID loss is also employed:

$$\begin{aligned} \mathcal{L}^p = &L_{id}(v^u) + L_{id}(t^u) + L_{id}(v^p) + L_{id}(t^p) \\ &+ L_{rk}(v^p, t^u) + L_{rk}(v^u, t^p), \end{aligned} \tag{23}$$

And then the loss between a pair of common manifold features is calculated as

$$\mathcal{L}^c = L_{id}(v^c) + L_{id}(t^c) + L_{rk}(v^c, t^c). \tag{24}$$

The entire loss for the second stage is

$$\mathcal{L}^{Stage2} = \mathcal{L}^{Stage1} + \lambda_4 \mathcal{L}^{p} + \lambda_5 \mathcal{L}^{c}. \tag{25}$$

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 ***Dataset and Metrics.*** Our approach is evaluated on two challenging Text-based Person Retrieval datasets including CUHK-PEDES [13] and RSTPReid [40].

(1) **CUHK-PEDES**: Following the official data split approach [13], the training set of CUHK-PEDES contains 34054 images, 11003 persons and 68126 textual descriptions. The validation set contains 3078 images, 1000 persons and 6158 textual descriptions while the test set has 3074 images, 1000 persons and 6156 descriptions. Every image generally has two descriptions, and each sentence is commonly no shorter than 23 words. After dropping words that appear less than twice, the word number is 4984.

(2) **RSTPReid**: The RSTPReid dataset [40] is constructed based on MSMT17 [32], which contains 20505 images of 4,101 persons from 15 cameras. Each person has 5 corresponding images taken by different cameras and each image is annotated with 2 textual descriptions. For data division, 3701, 200 and 200 identities are utilized for training, validation and test, respectively. Each sentence is no shorter than 23 words.

Evaluation Metrics. The performance is evaluated by the rank-k accuracy. All images in the test set are ranked by their similarities with a given query natural language sentence. If any image of the corresponding person is contained in the top-k images, we call this a successful search. We report the rank-1, rank-5, and rank-10 accuracies for all experiments.

4.1.2 ***Implementation Details.*** In our experiments, we set the representation dimensionality $p = 2048$. The dimensionality of embedded word vectors is set to 500. The pretrained ResNet-50 [7] is utilized as the visual CNN backbone and a pretrained BERT language model [24] is used to better handle the textual input. The total number of noun phrases obtained from each sentence is kept flexible while k is set to 6. For both the CUHK-PEDES and RSTPReid dataset, the input images are resized to $384 \times 128 \times 3$. The random horizontal flipping strategy is employed for data augmentation. The threshold value γ in $\mathcal{CA}$ can be $\frac{1}{k}$ or $\frac{1}{n}$ for visual or textual data, respectively. The margin β of ranking losses is set to 0.2 while the λ's are empirically set to 1 in this paper. An Adam optimizer [11] is adopted to train the model with a batch size of 64 for 100 epochs.

4.2 Ablation Analysis

4.2.1 ***Comparison with CDCP-based Paradigms.*** The core idea of this paper is the 'Look Before You Leap (LBUL)' mechanism. To demonstrate its effectiveness, a series of experiments on CUHK-PEDES and RSTPReid are carried out to compare LBUL with CDCP-based paradigms:

(1) **CDCP-Sep (glo)**: directly using the global features v^g and t^g extracted by modality-specific sub-models to calculate s^g for matching;

(2) **CDCP-Sha (glo)**: calculating s^g after processing v^g and t^g with a shared mapping block;

Table 1: Performance comparisons of common manifold mapping paradigms on CUHK-PEDES and RSTPReid.

	Method	Rank-1	Rank-5	Rank-10
CUHK-PEDES	CDCP-Sep (glo)	56.19	76.43	83.63
	CDCP-Sha (glo)	56.78	77.19	84.07
	LBUL (glo)	**57.20**	**77.78**	**84.23**
	CDCP-Sep (USEM)	59.62	79.39	85.83
	CDCP-Sha (USEM)	59.88	79.56	85.91
	LBUL (USEM)	**61.95**	**81.16**	**87.19**
RSTPReid	CDCP-Sep (glo)	37.05	61.75	71.10
	CDCP-Sha (glo)	37.85	62.05	71.50
	LBUL (glo)	**38.65**	**64.70**	**73.20**
	CDCP-Sep (USEM)	40.70	65.55	75.10
	CDCP-Sha (USEM)	41.40	65.85	75.40
	LBUL (USEM)	**43.35**	**66.85**	**76.50**

(3) **LBUL (glo)**: processing global features with LBUL mechanism (XProj + LASM) to obtain s^g;

(4) **CDCP-Sep (USEM)**: calculating a similarity s^u with output of USEM v^u and t^u, and then matching multi-modal samples with s^g, s^f and s^u;

(5) **CDCP-Sha (USEM)**: calculating s^c after processing v^u and t^u with a shared mapping block and then matching multi-modal samples with s^g, s^f and s^c;

(6) **LBUL (USEM)**: just equivalent to the complete LBUL method, which conducts the LBUL-based operation on v^u and t^u to obtain s^c and matches multi-modal samples with s^g, s^f and s^c.

The experimental results are reported in Tab. 1. As can be observed from the table, for CDCP-based methods, approaches using a shared common manifold mapping block outperform ones without to some extent. Furthermore, with the idea of LBUL for common manifold mapping, an obvious performance gain can be achieved. For instance, for methods using v^u and t^u, the LBUL-based method outperforms the CDCP-based method with a shared mapping block by 2.07%, 1.60%, 1.28% and 1.95%, 1.00%, 1.10% on CUHK-PEDES and RSTPReid, respectively. The experimental results demonstrate that by conducting the common manifold mapping operation after seeing distribution characteristics of both modalities, LBUL is more capable of learning consistent common representations with less information loss and higher retrieval accuracy.

4.2.2 ***Impact of Uni-modal Sub-manifold Embedding Module (USEM).*** We enumerate some variants for the Uni-modal Sub-manifold Embedding Module (USEM) and compare them with our proposed USEM:

(1) **Glo**: $v^u,\ t^u = v^g,\ t^g$;

(2) **Avg**: $v^u,\ t^u = avg(V),\ avg(T)$;

(3) **AvgLoc + Glo**: $v^u,\ t^u = v^g + avg(V^l),\ t^g + avg(T^l)$, where $avg(\cdot)$ denotes the feature averaging operation.

We can observe from Tab. 2 that the performance of the model with USEM is obviously better than the other substitutions, which indicate the effectiveness of USEM.

4.2.3 ***Impact of Leaping After Seeing Module (LASM).*** How to properly take advantage of the two feature representations x^u

Table 2: Ablation study on proposed components of LBUL on CUHK-PEDES and RSTPReid.

Component							CUHK-PEDES			RSTPReid		
s^g	s^f	s^c	2ST	USEM	LASM	BERT	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
✓	✓	✓	×	✓	✓	×	53.50	75.05	82.68	34.85	60.15	70.95
✓	×	×	-	-	-	×	54.81	75.65	83.43	37.00	61.75	71.10
✓	✓	×	-	-	-	×	57.37	77.83	85.00	40.15	63.95	74.10
✓	✓	✓	✓	glo	✓	×	60.81	80.79	86.72	41.60	65.55	75.30
✓	✓	✓	✓	avg	✓	×	60.37	80.17	86.62	42.40	66.15	76.10
✓	✓	✓	✓	avgloc + glo	✓	×	61.04	80.95	86.85	42.15	66.00	76.15
✓	✓	✓	✓	✓	w/o $\mathcal{DS}$	×	60.24	80.12	86.32	41.25	65.15	75.55
✓	✓	✓	✓	✓	add	×	60.57	80.56	86.56	40.45	64.35	73.70
✓	✓	✓	✓	✓	add + $\mathcal{MLP}$	×	60.32	80.05	86.53	41.15	65.25	75.35
✓	✓	✓	✓	✓	concat	×	60.31	80.41	86.93	42.25	66.30	76.15
✓	✓	✓	✓	✓	concat + $\mathcal{MLP}$	×	60.39	80.36	86.19	42.05	65.70	75.95
✓	✓	✓	✓	✓	scalar gate	×	61.27	80.70	86.89	42.65	66.55	76.30
✓	✓	✓	✓	✓	⊕=add	×	61.74	80.93	86.97	43.20	66.60	76.55
✓	✓	✓	✓	✓	✓	×	61.95	81.16	87.19	43.35	66.85	76.50
✓	✓	✓	✓	✓	✓	✓	**64.04**	**82.66**	**87.22**	**45.55**	**68.20**	**77.85**

Table 3: Comparison with SOTA on CUHK-PEDES.

Method	Rank-1	Rank-5	Rank-10
CNN-RNN [20]	8.07	-	32.47
Neural Talk [25]	13.66	-	41.72
GNA-RNN [13]	19.05	-	53.64
IATV [12]	25.94	-	60.48
PWM-ATH [3]	27.14	49.45	61.02
Dual Path [39]	44.40	66.26	75.07
GLA [2]	43.58	66.93	76.26
CMPM-CMPC [36]	49.37	71.69	79.27
MIA [18]	53.10	75.00	82.90
A-GANet [16]	53.14	74.03	81.95
PMA [10]	54.12	75.45	82.97
TIMAM [21]	54.51	77.56	84.78
CMAAM [1]	56.68	77.18	84.86
AMEN [29]	57.16	78.64	86.22
HGAN [38]	59.00	79.49	86.62
DSSL [40]	59.98	80.41	87.56
LBUL (Ours)	61.95	81.16	87.19
LBUL + BERT (Ours)	**64.04**	**82.66**	**87.22**

and x^p in the Leaping After Seeing Module (LASM) is crucial to the performance of LBUL. Therefore, we conduct experiments with several substitution variants of LASM on both the CUHK-PEDES and RSTPReid dataset to see the effectiveness of our proposed method:

(1) **Add**: $x^c = x^u + x^p$;

(2) **Add +** $\mathcal{MLP}$: $x^c = tanh(W(x^u + x^p) + b)$;

(3) **Concat**: $x^c = \begin{bmatrix} x^u \\ x^p \end{bmatrix}$;

Table 4: Comparison with SOTA on RSTPReid.

Method	Rank-1	Rank-5	Rank-10
IMG-Net [31]	37.60	61.15	73.55
AMEN [29]	38.45	62.40	73.80
DSSL [40]	39.05	62.60	73.95
SSAN [4]	43.50	67.80	77.15
LBUL (ours)	43.35	66.85	76.50
LBUL + BERT (Ours)	**45.55**	**68.20**	**77.85**

(4) **Concat +** $\mathcal{MLP}$: $x^c = tanh(W \begin{bmatrix} x^u \\ x^p \end{bmatrix} + b)$;

(5) **Scalar Gate**: $x^c = ax^u + (1 - a)x^p$ where a is a real-valued number parameterized by x^u and x^p;

(6) ⊕=**add**: substitute the concatenation operation in Eq. 8 for addition.

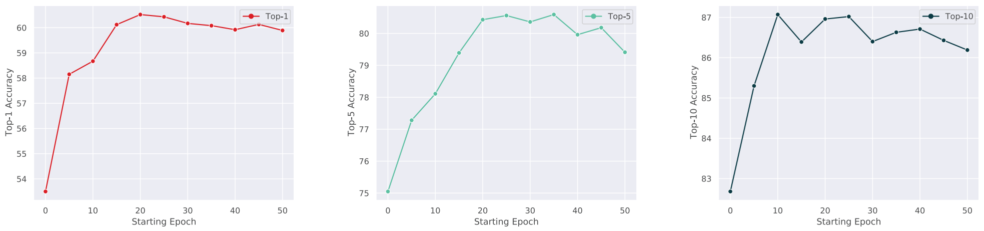
As can be seen from Tab. 2, the feature aggregation paradigm proposed in LASM (i.e. Eq. 8 and Eq. 9) achieves substantially better performance than all the basic variants, while implementing ⊕ in Eq. 8 as addition or concatenation give similar retrieval accuracies, with the concatenation method slightly better.

4.2.4 *Impact of Distribution Shifting Mechanism.*

As shown in Tab. 2, the Distribution Shifting ($\mathcal{DS}$) mechanism is a key component in XProj, without which the performance drops by 1.71%, 1.04%, 0.87% and 2.10%, 1.70%, 0.95% on CUHK-PEDES and RSTPReid, respectively. The decrease in retrieval accuracy indicates the significance of the proposed $\mathcal{DS}$ mechanism.

4.2.5 *Impact of the Two-stage Training Strategy.*

As described in Sec. 3.5, a two-stage strategy is proposed to train LBUL. To prove the validity of this optimization strategy, we conduct experiments



(a) Rank-1 Accuracy (b) Rank-5 Accuracy (c) Rank-10 Accuracy

Figure 3: Illustration of the impact of the starting epoch for the second training stage on CUHK-PEDES.

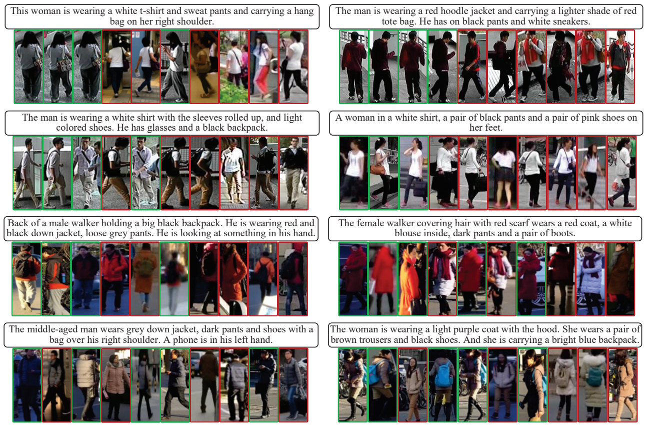

This woman is wearing a white t-shirt and sweat pants and carrying a hang bag on her right shoulder.
The man is wearing a red hoodle jacket and carrying a lighter shade of red tote bag. He has on black pants and white sneakers.
The man is wearing a white shirt with the sleeves rolled up, and light colored shoes. He has glasses and a black backpack.
A woman in a white shirt, a pair of black pants and a pair of pink shoes on her feet.
Back of a male walker holding a big black backpack. He is wearing red and black down jacket, loose grey pants. He is looking at something in his hand.
The female walker covering hair with red scarf wears a red coat, a white blouse inside, dark pants and a pair of boots.
The middle-aged man wears grey down jacket, dark pants and shoes with a bag over his right shoulder. A phone is in his left hand.
The woman is wearing a light purple coat with the hood. She wears a pair of brown trousers and black shoes. And she is carrying a bright blue backpack.

Figure 4: Illustration of top-10 text-based person retrieval results by LBUL. The matched pedestrian images are marked by green rectangles, while the mismatched person images are marked by red rectangles.

which train LBUL with a one-stage strategy, namely, utilize $\mathcal{L}^{Stage2}$ since the beginning. As can be seen from Tab. 2, the retrieval performance decrease by 8.45%, 6.11%, 4.51% and 8.50%, 6.70%, 5.55% on CUHK-PEDES and RSTPReid, respectively, which demonstrate that after reliable uni-modal sub-manifolds are learned, a more consistent cross-modal common manifold can be constructed by LBUL. To further analysis the impact of the starting epoch for the second training stage, we conduct extensive experiments on CUHK-PEDES and the results are illustrated in Fig. 3. It can be observed that initially the performance keeps increasing with the growth of the starting epoch for the second stage. Then after the value of the starting epoch passes 15, the performance gradually stabilizes. And the performance drop slightly after the value of the starting epoch gets too large. Note that when the starting epoch is 0, it is just equivalent to adopting a one-stage training strategy.

4.3 Comparison with SOTA on Text-based Person Retrieval

We compare the proposed LBUL with previous methods on CUHK-PEDES and RSTPReid. It can be observed from Tab. 3 and Tab. 4 that without BERT, our proposed LBUL achieves 61.95%, 81.16% and 87.19% of rank-1, rank-5 and rank-10 accuracies respectively on CUHK-PEDES and 43.35%, 66.85% and 76.50% on RSTPReid. By encouraging the proposed model to look before it leaps with the proposed two-step common manifold mapping mechanism, LBUL outperforms existing methods and achieves the state-of-the-art performance on the text-based person retrieval task. For instance, TIMAM [21] is one of the typical CDCP-based approaches. It aims to learn modality-invariant feature representations using adversarial and cross-modal matching objectives and utilizes a pretrained ResNet-101 as the visual backbone. With a ResNet-50 backbone, LBUL outperforms TIMAM by 7.44%, 3.60% and 2.38% on CUHK-PEDES of rank-1, rank-5 and rank-10 accuracies, respectively, which further proves the effectiveness of our proposed method. We display some examples of the top-10 text-based person retrieval results by LBUL in Fig. 4. The matched/mismatched pedestrian images are marked by green/red rectangles.

5 CONCLUSION

In this paper, we propose a novel algorithm termed LBUL to learn a Consistent Cross-modal Common Manifold (C^3M) for text-based person retrieval to overcome the CDCP dilemma. The core idea of our method, just as a Chinese saying goes, is 'san si er hou xing', namely, to Look Before yoU Leap (LBUL). The common manifold mapping mechanism of LBUL includes two steps, namely, a looking step and a leaping step. Compared to CDCP-based common manifold mapping paradigms, LBUL considers distribution characteristics of both the visual and textual modalities before embedding data from one certain modality into C^3M to achieve a more solid cross-modal distribution consensus, and hence achieve a superior retrieval accuracy. We evaluate our proposed method on two text-based person retrieval datasets CUHK-PEDES and RSTPReid. Experimental results demonstrate that the proposed LBUL outperforms previous methods and achieves the state-of-the-art performance.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (Grant No. 62101245, 61972016), China Postdoctoral Science Foundation (Grant No.2019M661999) , Natural Science Research of Jiangsu Higher Education Institutions of China (19KJB520009) and Future Network Scientific Research Fund Project (Grant No. FNSRFP-2021-YB-21).

REFERENCES

[1] Surbhi Aggarwal, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. 2020. Text-based person search via attribute-aided matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2617–2625.

[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 54–70.

[3] T. Chen, C. Xu, and J. Luo. 2018. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1879–1887.

[4] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification. *arXiv preprint arXiv:2107.12666* (2021).

[5] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 4 (2018), 1–18.

[6] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. 2019. 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*. 6112–6121.

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

[8] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9317–9326.

[9] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2020. Global distance-distributions separation for unsupervised person re-identification. In *European Conference on Computer Vision*. Springer, 735–751.

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

[11] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

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

[13] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1970–1979.

[14] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2197–2206.

[15] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. 2018. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440* (2018).

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

[17] Xin Ning, Ke Gong, Weijun Li, and Liping Zhang. 2021. JWSAA: joint weak saliency and attention aware for person re-identification. *Neurocomputing* 453 (2021), 801–811.

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

[19] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. 2021. Adaptive Label-aware Graph Convolutional Networks for Cross-Modal Retrieval. *IEEE Transactions on Multimedia* (2021).

[20] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–58.

[21] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5814–5824.

[22] Changchang Sun, Xuemeng Song, Fuli Feng, Wayne Xin Zhao, Hao Zhang, and Liqiang Nie. 2019. Supervised hierarchical cross-modal hashing. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 725–734.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

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

[25] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[26] Dongkai Wang and Shiliang Zhang. 2020. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10981–10990.

[27] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and CH Steven Hoi. 2021. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Transactions on Multimedia* (2021).

[28] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2275–2284.

[29] Zijie Wang, Jingyi Xue, Aichun Zhu, Yifeng Li, Mingyi Zhang, and Chongliang Zhong. 2021. AMEN: Adversarial Multi-space Embedding Network for Text-Based Person Re-identification. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 462–473.

[30] Zijie Wang, Aichun Zhu, Jingyi Xue, Daihong Jiang, Chao Liu, Yifeng Li, and Fangqiang Hu. 2022. SUM: Serialized Updating and Matching for text-based person retrieval. *Knowledge-Based Systems* 248 (2022), 108891.

[31] Zijie Wang, Aichun Zhu, Zhe Zheng, Jing Jin, Zhouxin Xue, and Gang Hua. 2020. IMG-Net: inner-cross-modal attentional multigranular network for description-based person re-identification. *Journal of Electronic Imaging* 29, 4 (2020), 043028.

[32] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 79–88.

[33] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. 2019. Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3760–3769.

[34] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 34–39.

[35] Yuan Yuan, Jian'an Zhang, and Qi Wang. 2020. Deep Gabor convolution network for person re-identification. *Neurocomputing* 378 (2020), 387–398.

[36] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 686–701.

[37] Mingbo Zhao, Jiao Liu, Zhao Zhang, and Jicong Fan. 2021. A scalable sub-graph regularization for efficient content based image retrieval with long-term relevance feedback enhancement. *Knowledge-based systems* 212 (2021), 106505.

[38] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. 2020. Hierarchical Gumbel Attention Network for Text-based Person Search. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3441–3449.

[39] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-Path Convolutional Image-Text Embeddings with Instance Loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.

[40] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 209–217.