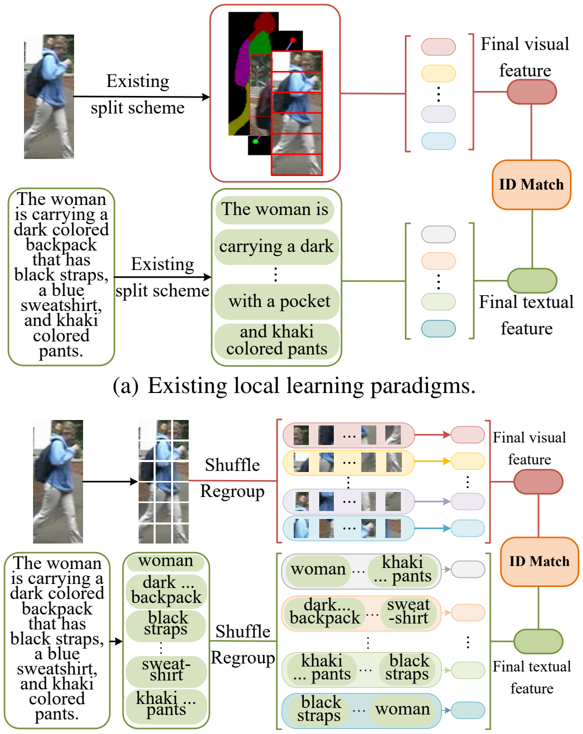
Multi-Modal Disordered Representation Learning Network for Description-Based Person Search

Fan Yang, Wei Li*, Menglong Yang, Binbin Liang, Jianwei Zhang

Sichuan University, Chengdu, China
yang_fan_s@163.com, {li.wei, mlyang, zhangjianwei}@scu.edu.cn, sculiang@126.com

Abstract

Description-based person search aims to retrieve images of the target identity via textual descriptions. One of the challenges for this task is to extract discriminative representation from images and descriptions. Most existing methods apply the part-based split method or external models to explore the fine-grained details of local features, which ignore the global relationship between partial information and cause network instability. To overcome these issues, we propose a Multi-modal Disordered Representation Learning Network (MDRL) for description-based person search to fully extract the visual and textual representations. Specifically, we design a Cross-modality Global Feature Learning Architecture to learn the global features from the two modalities and meet the demand of the task. Based on our global network, we introduce a Disorder Local Learning Module to explore local features by a disordered reorganization strategy from both visual and textual aspects and enhance the robustness of the whole network. Besides, we introduce a Cross-modality Interaction Module to guide the two streams to extract visual or textual representations considering the correlation between modalities. Extensive experiments are conducted on two public datasets, and the results show that our method outperforms the state-of-the-art methods on CUHK-PEDES and ICFG-PEDES datasets and achieves superior performance.


Existing
split scheme
Final visual
feature
ID Match
The woman is carrying a dark colored backpack that has black straps, a blue sweatshirt, and khaki colored pants.
Existing
split scheme
The woman is
carrying a dark
with a pocket
and khaki colored pants
Final textual
feature
Shuffle
Regroup
Final visual
feature
ID Match
The woman is carrying a dark colored backpack that has black straps, a blue sweatshirt, and khaki colored pants.
woman
dark ... backpack
black straps
sweat-shirt
khaki ... pants
Shuffle
Regroup
woman ... khaki ... pants
dark... backpack ... sweat -shirt
khaki ... pants ... black straps
black straps ... woman
Final textual
feature

(a) Existing local learning paradigms.

(b) The proposed disorder local learning paradigms.

Figure 1: Illustration of existing and our methods. We design a disorder strategy to enhance global correlation of local clues without extra models and enhance image/text features.

Introduction

Description-based person search is a critical task for cross-modality learning, aiming to retrieve images of the specific person according to a given descriptive sentence. It attracts increasing attention and plays an important role in public security and smart surveillance in recent years. The general process of the description-based person search is to match the same identity between visual and textual representation. The target of this task is pedestrians, and the appearance and language description of different persons are semantically similar in some cases. Therefore, one of the challenges is how to generate discriminative and robust features from multimodal data. To boost the learning capability of the cross-modality model, most methods (Niu et al. 2020; Wang et al. 2020; Ding et al. 2021; Shao et al. 2022; Farooq et al. 2022; Li, Cao, and Zhang 2022; Jing et al. 2020) utilize the local representation extract scheme to enrich the visual and textual information. However, most of these existing methods generally adopted the hard split strategy to divide the visual and textual representation by parts and extract the partial information, which ignores the global relationship between local features. If the local features of the two modalities are only extracted by region or phrase, the single details from each part will be explored, but the holistic relevance between them will be destroyed. For example, when two different pedestrians carry white backpacks simultaneously, the information about the white backpack

[*] Corresponding author

hardly fails to provide beneficial help for pedestrian distinction. However, when the backpack is associated with other partial clues (such as a blue coat or white trousers), as shown in Figure 1, the partial representation containing overall correlation will provide more powerful support for the distinguishability of final features. Therefore, it is necessary to explore the partial representation with global correlation. On the other hand, some of these methods (Wang et al. 2020; Jing et al. 2020) introduce extra models into the network, such as semantic segmentation, pose estimation, or attribute recognition, to guide the division of the region. The accuracy of the additional model directly affects the performance of the framework, and the whole network fails to be trained end-to-end.

To address the aforementioned problems, we propose a multi-modal disordered representation learning network for description-based person search to boost the learning capability of the cross-modality model. First of all, we build a strong cross-modality global feature learning architecture to facilitate the task, which contains a visual information learning branch and a textual information learning branch to generate the features of pictures and texts, respectively. To extract the partial information in images and descriptions effectively, we louse up the visual or textual token sequences and reorganize them into different groups. Different from the traditional segment method, each group includes random parts of the whole image or description. We concatenate the global representations of images or texts with each group to learn the local features and the relevance between partial information from these disordered token sequences. Furthermore, we combine the visual and textual token sequences and associate them employing identification loss in the training phase to optimize the two branches simultaneously.

Our main contributions can be summarized as follows:

- A Cross-modality Global Feature Learning Architecture is proposed to perform the description-based person search task and extract the visual and textual global features.
- We introduce a Disorder Local Learning Module is introduced to both explore the visual and textual fine-grained details, which strengthen the representation learning ability and the robustness of the network.
- We design a Cross-modality Interaction Module to learn the image-text interactive representation and to promote the two branches to extract information from images or descriptions considering the correlation between modalities.
- Extensive experiments verify the superiority of our model on CUHK-PEDES and ICFG-PEDES. The results demonstrate that the proposed method achieves state-of-the-art performance.

Related Work

Person Re-Identification

With the development of deep learning, person re-identification (Re-ID) has drawn increasing attention in recent years. This task aims to match the same pedestrian captured by different non-overlap cameras. Most methods have been proposed to meet this requirement. Sun et al. (Sun et al. 2018) propose a partitioning strategy that splits the image into several regions to mine the partial details. Although utilizing this method can boost the discriminative capability of the network, it ignores the important role of the global feature and breaks the structural characteristic of the person. Wei et al. (Wei et al. 2018) adopt an alignment scheme that utilizes human pose information to extract and align the holistic-local representations. The above methods require an extra human pose estimator, and the performance of additional models can influence the accuracy of the network. Wang et al. (Wang et al. 2018) employ a learning method combining holistic and local representations that exploit three branches with different granularities to learn fine-grained representation. Inspired by this method, Mao et al. (Mao et al. 2020) apply a multi-granularity region-based approach to mine discriminative partial representation. However, the above methods will produce enormous calculated volume while improving accuracy. He et al. (He et al. 2021) introduce the Transformer into Re-ID task and first utilize a patch shuffle method to mine the local features of the images. Park et al. (Park and Ham 2020) propose a region-level relation model to learn robust representations and introduce a holistic contrastive pooling strategy to explore discriminative information from human pictures. Chen et al. (Chen et al. 2020) employ partial representations of the first frame to acquire relationship between each frame. Zhang et al. (Zhang et al. 2020) introduce a guidance-based attention information fusion method, which obtains the reference as a guiding vector to learn relationship. However, possible poor-quality frames can impact the generation of the better relationship between frames. Inspired by this, Yang et al. (Yang et al. 2022) design a relation-based holistic-local representation extracting network, which applies the temporal attention module to obtain video reference and learn the discriminative features.

Description-Based Person Search

Description-based person search aims to retrieve pedestrian pictures from the visual database according to a textual query. This task is first proposed by Li et al. (Li et al. 2017b) , who collect a large-scale CUHK-PEDES dataset and design a novel recurrent neural network with a gated neural attention mechanism model to match the person image and the text description. Based on this work, many models is designed to accomplish the goal of this task in recent years. Li et al. (Li et al. 2017a) introduce a two-stage matching network that exploits the identity information to ease metric learning. Zhang et al. (Zhang and Lu 2018) utilize a cross-modality projection matching loss to learn relationship between images and texts. Aggarwal et al. (Aggarwal, Radhakrishnan, and Chakraborty 2020) utilize extra attribute annotation to strengthen the discriminative of the global features. Chen et al. (Chen et al. 2021) take the issue of information inequality between the two modalities into account and employ a cross-modal knowledge adaptation strategy to adapt the knowledge of images and texts at different levels. However, the above methods only focus on

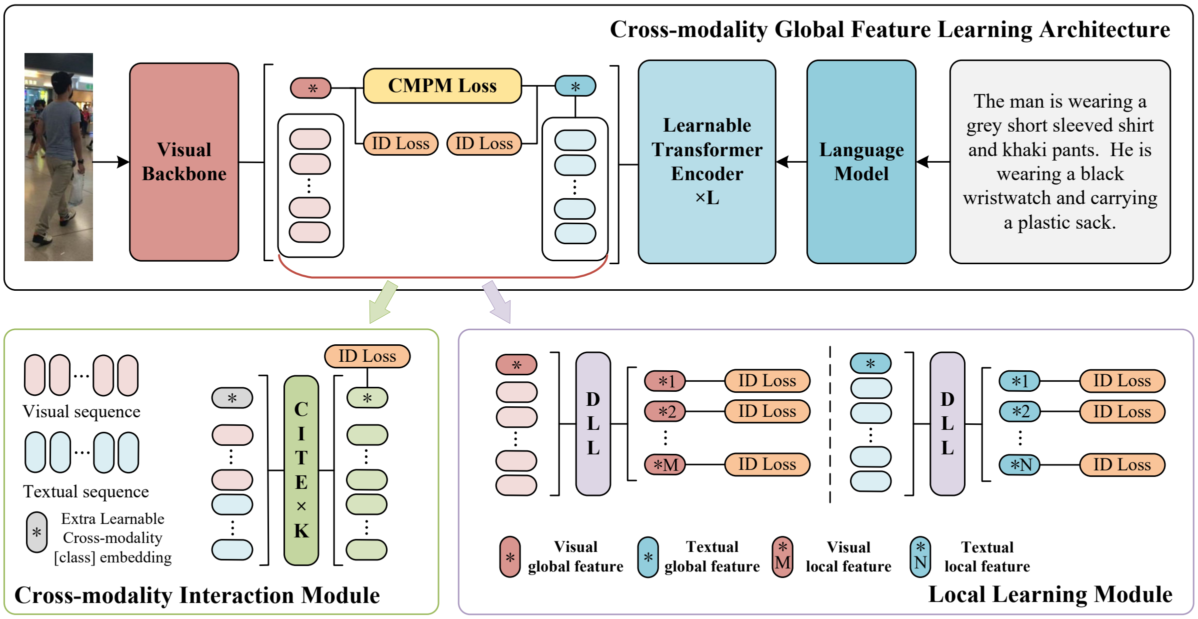

Cross-modality Global Feature Learning Architecture
Visual Backbone
CMPM Loss
ID Loss
ID Loss
Learnable Transformer Encoder ×L
Language Model
The man is wearing a grey short sleeved shirt and khaki pants. He is wearing a black wristwatch and carrying a plastic sack.
Visual sequence
Textual sequence
Extra Learnable Cross-modality [class] embedding
CITE × K
ID Loss
Cross-modality Interaction Module
DLL
*1
*2
*M
ID Loss
*1
*2
*N
Visual global feature
Textual global feature
Visual local feature
Textual local feature
Local Learning Module

Figure 2: Illustration of the proposed method.

the global level and neglect the fine-grained details of the visual and textual information. To learn partial discriminative information, some researchers attempt to explore the local feature both from images and texts. Wang et al. (Wang et al. 2020) employ the semantic attribute information of the person as a reference to extract the partial attribute representation and design a k-reciprocal sampling align loss to learn the relationship between visual and textual clues. Jing et al. (Jing et al. 2020) apply the pose information to align the human part of the images and phrases of captions. These methods boost the performance, but they utilize additional models whose performance can affect the accuracy of the whole framework. Niu et al. (Niu et al. 2020) introduce a multi-granularity cross-modality align approach from three types including holistic-holistic, holistic-partial, and partial-partial, and separately match them from the three aspects. Ding et al. (Ding et al. 2021) introduce a semantic self-aligned model to automatically extract the local textual features to search the related visual patch and build another description-based person search dataset named ICFG-PEDES. Wu et al. (Wu et al. 2021) take color reasoning into account and introduce a fine-grained cross-model association explicitly to capture the color information from the image and description. Shao et al. (Shao et al. 2022) design a Learning Granularity-Unified Representations framework to learn equal granularity between visual and textual modality, which boosts the accuracy of the text-to-image retrieval model. Farooq et al. (Farooq et al. 2022) utilize a cross-modal semantic alignment block to align the both modalities and strength the region-based information. However, the above works obtain the visual and textual parts by hand-crafted split or additional models, which neglect the holistic corresponding between each local feature, and the performance of the extra models can affect the accuracy of the whole framework.

Proposed Method

Framework Overview

The overall network of the proposed mothed is shown in Figure 2. The framework contains a Cross-modality Global Feature Learning Architecture, a Local Learning Module (LLM), and a Cross-modality Interaction Module (CIM). In the training process, the training data is assumed as $D = \{I_r, T_r\}_{r=1}^{G}$ where G denotes the image-text pair number of each batch. Both input the images and descriptions of the pedestrian into the global learning architecture to learn visual holistic representation f_g^I and textual global representation f_g^T utilizing the visual model and the language model. Then, the visual patches token sequence $\{f_i^I(i \in [1, N])\}$ and the textual words token sequence $\{f_j^T(j \in [1, M])\}$ is respectively processed by LLM and CIM to learn the visual/textual local feature and cross-modality representation (i and j are integers). Finally, we combine the identification (ID) loss and the Cross-model Projection Matching (CMPM) loss (Zhang and Lu 2018) supervise and optimize the whole model. The details of the modules are elaborated in the following sections.

Cross-Modality Global Feature Learning Architecture

We build a Cross-modality Global Feature Learning Architecture for description-based person search, which contains a visual information learning branch and a textual information learning branch. The pipeline is shown in Figure 2.

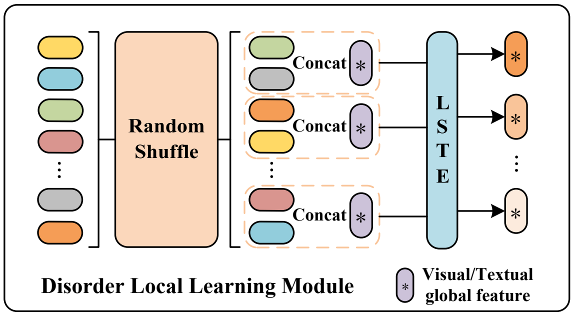


Figure 3: Illustration of disorder local learning module.

Visual information learning branch. We utilize the Vision Transformer (ViT) (Dosovitskiy et al. 2020) visual backbone to extract visual representation. Given an image $I \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width of the image, we split it into N fixed-sized patches $\{I_i | i \in [1, N]\}$. Then, these patches are fed into the visual backbone to learn the visual global representation f^I_g of the image. Specifically, we utilize a linear projection embedding layer to map these patches to D dimensions, which are denoted as "tokens" $x_i(i \in [1, N])$. The embedding function can be expressed as:

$$x_i = E(I_i) \tag{1}$$

where $E(\cdot)$ represents the linear projection embedding layer.

To extract the visual representation, an extra learnable [cls] embedding token x_{cls} is employed before the image patch sequence. However, due to the characteristics of the Transformer, the patches are flattened into one-dimensional vectors in the process of linear mapping and the original location information is lost. Therefore, we utilize a learnable position embedding $P \in R^{(n+1)\times D}$ to learn the position information for each patch. The input sequence X of the visual backbone is shown as follows:

$$X = (x_{cls}; x_1; x_2; \ldots; x_N) + P \tag{2}$$

The output of the visual backbone is:

$$F^I = VB(X) = (f^I_g; f^I_1; f^I_2; \ldots; f^I_N) \tag{3}$$

where F^I represents the output visual sequence, $VB(\cdot)$ is the visual backbone, f^I_g denotes the global feature of the image.

Textual information leaning branch. For textual representation learning, we first employ the superior pre-trained language model BERT (Devlin et al. 2018) the word representation. Specifically, given a sentence of a person with M words $\{T_j | j \in [1, M]\}$, we combine the extra learnable [cls] embedding token y_{cls} and the [sep] embedding token y_{sep} into the beginning and end of each sentence. The textual token sequence can be expressed as:

$$T = (y_{cls}; T_1; T_2; \ldots; T_M; y_{sep}) \tag{4}$$

Similar to the visual token sequence, the location information is also important for each word. Thus, we add the learnable position embedding P into the textual token sequence. The input of the textual branch can be expressed as:

$$Y = (y_{cls}; T_1; T_2; \ldots; T_M; y_{sep}) + P \tag{5}$$

Then, we input the token sequence of the description into the BERT model to learn the textual embedding. It is worth noting that, due to the robust language representation ability, we fix the parameter of the BERT and add L Learnable Transformer Encoders (LTE) after the BERT to learn the training parameters from the caption of the pedestrian. Thus, the output of the textual branch is:

$$F^T = TB(Y) = (f^T_g; f^T_1; f^T_2; \ldots; f^T_M) \tag{6}$$

where F^T denotes the output textual embedding sequence, $TB(\cdot)$ represents the textual branch, f^T_g is the global representation of the description.

Disorder Local Learning Module

To explore the local details of the person from the visual and textual information, we propose a Disorder Local Learning Module (DLL), as shown in Figure 3. Different from the traditional hard split methods for the token sequences, we utilize a random disturbing strategy to shuffle the visual embedding token sequence $\{f^I_i | i \in [1, N]\}$ or textual embedding token sequence $\{f^T_j | j \in [1, M]\}$ disorderly. Then, we reorganize the new disorder visual and textual embedding token sequence into P and Q groups (N or M is divisible by P or Q respectively). With the random shuffling strategy, each visual or textual group contains some random embedding tokens of the whole image or description. We combine the global feature f^*_g with each new group to learn the local features. The new visual and textual local sequences are shown as follows:

$$Z^I_{l^a} = (f^I_g; f^I_{n^a_1}; f^I_{n^a_2}; \ldots; f^I_{n^a_{N/P}}), a \in [1, P], n^a_* \in [1, N] \tag{7}$$

$$Z^T_{l^b} = (f^T_g; f^T_{m^b_1}; f^T_{m^b_2}; \ldots; f^T_{m^b_{M/Q}}), b \in [1, Q], m^b_* \in [1, M] \tag{8}$$

where a, b, n^a_*, and m^b_* are integers, $Z^I_{l^a}$ and $Z^T_{l^b}$ denote the a-th visual local sequence and the b-th textual local sequence. Then, we input the new sequence into the shared self-attention layer named Local Spatial Transformer Encoder (LSTE) to explore visual and textual partial features, respectively. The function of LSTE is shown as follows:

$$q = Z^*_{l^\cdot} W_q \tag{9}$$

$$k = Z^*_{l^\cdot} W_k \tag{10}$$

$$v = Z^*_{l^\cdot} W_v \tag{11}$$

$$A = softmax(qk^T / \sqrt{C/h}) \tag{12}$$

$$F^*_{l^\cdot} = Av + MLP(LN(Av)) \tag{13}$$

where $W_q, W_k, W_v \in \mathbb{R}^{D\times(D/h)}$ are the learnable parameters, h denotes the number of heads, A is the attention map, $F^*_{l^\cdot}$ is the output of the LSTE, and $LN(\cdot)$ represents the layer normalization.

Therefore, the visual and textual outputs of the LSTE are:

$$F^I_{l^a} = LSTE(Z^I_{l^a}) = (f^I_{l^a}; f^I_{l^a_1}; f^I_{l^a_2}; \ldots; f^I_{l^a_{N/P}}), a \in [1, P] \tag{14}$$

$$F^T_{l^b} = LSTE(Z^T_{l^b}) = (f^T_{l^b}; f^T_{l^b_1}; f^T_{l^b_2}; \ldots; f^T_{l^b_{M/Q}}), b \in [1, Q] \tag{15}$$

where $F^I_{l^a}$ and $F^T_{l^b}$ is the output of the LSTE, $LSTE(\cdot)$ denotes the LSTE layer, $f^I_{l^a}$ represents the $a-th$ visual local feature, $f^T_{l^b}$ represents the $b-th$ textual local representation.

Cross-Modality Interaction Module

Although the images and texts are different modalities, the visual and textual information describing the same person have strong correlations. Therefore, we propose a Cross-modality Interaction Module (CIM) to learn the interactive representations of images and texts, which are beneficial for the model to classify different people and learn the visual and textual representation of the two branches. The CIM consists of K Cross-modality Interaction Transformer Encoders (CITE), as shown in Figure 2. We concatenate the visual token sequence $\{f^I_i | i \in [1, N]\}$ and textual token sequence $\{f^T_j | j \in [1, M]\}$ and add a learnable cross-modality [cls] embedding token x^c_{cls} before the regroup token sequence to learn the cross-modality interaction representation. The input of the CIM can be expressed as:

$$W = (x^c_{cls}; f^I_1; f^I_2; \ldots; f^I_N; f^T_1; f^T_2; \ldots; f^T_M) \tag{16}$$

The output of the CITE is:

$$F_c = CITE(W) = (f_c; f^c_1; \ldots; f^c_N; f^c_{N+1}; \ldots; f^c_{N+M}) \tag{17}$$

where F_c respects the cross-modality sequence, $CITE(\cdot)$ denotes the cross-modality interaction transformer encoder, f_c is the cross-modality interaction feature. The cross-modality interaction feature f_c is processed by the identification loss in the training phase to optimize the whole network well and promote the two branches to take into account the correlation between modalities in the process of extracting text or image information respectively.

Loss Function

We adopt the identification (ID) loss for classification and the Cross-model Projection Matching (CMPM) loss for eliminating the modality gap between picture and textual caption to optimize the whole network.

ID loss. The ID loss L_{ID} is the cross-entropy loss, which is implemented for discriminative learning. The ID loss can be formulated as follows:

$$L_{ID} = \sum_{i=1}^{N} -p_i log(q_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases} \tag{18}$$

where p_i is the prediction of the class i, y is the ground truth label, N is the number of persons.

CMPM loss. The CMPM loss is utilized to associate the visual and textual modalities and close the gap between the two modalities. In CMPM loss, the similarity of the two modalities is calculated as follows:

$$p^{V2T}_{i,j} = \frac{exp(v_i^\top \bar{t}_j)}{\sum_{k=1}^{N} exp(v_i^\top \bar{t}_k)} \tag{19}$$

where v_i is the visual feature, t_j is the textual feature, $\bar{t}_j$ is the normalized t_j , N is the number of persons, $p^{V2T}_{i,j}$ denotes the similarity from a visual representation to a textual representation.

The CMPM loss can be formulated as follows:

$$\begin{aligned} L_{CMPM} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} (p^{V2T}_{i,j} log(\frac{p^{V2T}_{i,j}}{\bar{y}_{i,j} + \varepsilon}) \\ + p^{T2V}_{j,i} log(\frac{p^{T2V}_{j,i}}{\bar{y}_{i,j} + \varepsilon}) \end{aligned} \tag{20}$$

where $p^{V2T}_{i,j}$ denotes the similarity from textual representation to visual representation, ε is a small value to avoid numerical problems, $y_{i,j} = 1$ represents v_i and t_j are a matching pair, otherwise not. $\bar{y}_{i,j}$ represents the normalized true matching probability between the two modalities for the situation of more than one matched item, which is calculated as $\bar{y}_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{N} y_{i,k}}$. The final loss function of our network is the sum of CMPM and ID losses from different branches.

Experiment

Experiment Settings

Datasets. We evaluate the proposed method on CUHK-PEDES (Li et al. 2017b) and ICFG-PEDES (Ding et al. 2021) , two public description-based person Re-ID datasets. The CUHK-PEDES dataset contains 40,206 images of 13,003 identities collected from several Re-ID datasets. Each image is annotated by two language descriptions, and each language description consists of more than 23 words. These language descriptions in this dataset cover 9,408 various words. We utilize the same data splitting method as (Li et al. 2017b) , which is split into 34,54 images with 11,003 identities for training, 3,078 images with 1,000 identities for validation, and 3,074 images with 1,000 pedestrians for testing. The ICFG-PEDES dataset contains 54,522 images with 4,102 identities. Each image is described by one caption. The vocabulary in this dataset includes more than 5,000 different words. For training and testing, the dataset is split into 3,102 and 1,000 pedestrians.

Evaluation metrics. As the standard evaluation metrics, the Cumulative Matching Characteristic curve (CMC) is adopted to evaluate the performance of the proposed method. We utilize Rank-1, Rank-5, and Rank-10 accuracies to express the CMC curve.

Implementation detail. We resize all input images to 384×128 and unify the length of input texts to 64. We train the whole network for 150 epochs and apply SGD to optimize model with the weight decay of 0.01 and a momentum of 0.9. The learning rate is set to 7×10^{-5}. The learning rate is initialized by the warm-up trick in the first 10 epochs.

Comparison with the State-of-the-Art

In order to explore the performance of the proposed model, we compare it with the state-of-the-art methods on CUHK-PEDES and ICFG-PEDES datasets, respectively. The results are summarized in Table 1.

Method	Source	CUHK-PEDES			ICFG-PEDES		
		Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
GNA-RNN (Wei et al. 2018)	CVPR17	18.05	—	53.64	18.05	—	53.64
IATV (Li et al. 2017a)	ICCV17	25.94	—	60.49	—	—	—
CMPM+CMPC (Zhang and Lu 2018)	ECCV18	49.37	—	79.27	18.05	—	53.64
GLA (Chen et al. 2018)	ECCV18	43.58	66.93	76.26	—	—	—
TIMAM (Sarafianos, Xu, and Kakadiaris 2019)	ICCV19	54.51	77.56	84.78	—	—	—
PMA (Jing et al. 2020)	AAAI20	47.02	68.54	78.06	—	—	—
MIA (Niu et al. 2020)	TIP20	48.00	70.70	79.30	46.49	67.14	75.18
TDE (Niu, Huang, and Wang 2020)	MM20	53.94	75.26	83.31	—	—	—
ViTAA (Wang et al. 2020)	ECCV20	55.97	75.84	83.52	43.58	66.93	76.26
MGEL (Wang et al. 2021)	IJCAI21	60.27	80.01	86.74	—	—	—
CMKA (Chen et al. 2021)	TIP21	54.69	73.65	81.86	—	—	—
DSSL (Zhu et al. 2021)	MM21	59.98	80.41	87.56	—	—	—
NAFS+LapsCore (Wu et al. 2021)	ICCV21	63.40	—	87.80	—	—	—
LGUR (Shao et al. 2022)	MM22	65.25	83.12	89.00	59.02	75.32	81.56
AXM-Net (Farooq et al. 2022)	AAAI22	64.44	80.52	86.77	—	—	—
IRRA (Jiang and Ye 2023)	CVPR23	73.38	89.93	93.71	63.46	80.25	85.82
MDRL(Ours)	AAAI24	**74.56**	**92.56**	**96.30**	**65.88**	**85.25**	**90.38**

Table 1: Comparison with the state-of-the-art methods on CUHK-PEDES and ICFG-PEDES datasets.

Comparisons on CUHK-PEDES. The results of the comparison on CUHK-PEDES are shown in Table 1. From the table, we can observe that our model outperform all existing advanced methods, achieving 74.56%, respectively. Specifically, the two proposed methods have 1.18% improvements compared with the best existing method IRRA in terms of Rank-1 accuracy. It is noteworthy that the proposed model obtains excellent performance, whose architecture is simple but effective. Consequently, we set a new SOTA result for description-based person search on the CUHK-PEDES dataset.

Comparisons on ICFG-PEDES. In Table 1, we summarize the accuracy of comparison with existing advanced models on the ICFG-PEDES dataset. It can be observed that our model achieves the best performance. Specifically, the proposed method obtains 65.88% Rank-1 accuracy, which outperforms all SOTA methods. Especially, compared with IRRA which has excellent performance, our model outperforms IRRA by 2.42% in terms of Rank-1 accuracy. The experimental results demonstrate the effectiveness of our method on description-based person search.

Ablation Studies

Effectiveness of each component. To evaluate the effectiveness of each component, we gradually combine them with the proposed global architecture. Table 2 lists the results of the ablation studies on the CUHK-PEDES dataset. “Global” denotes our cross-modality global model for description-based person search. “DDL (V)” and “DDL (V&T)” represent utilizing the DDL module to learn the visual local or visual and textual local features. From the table, we can observe that both DDL (V) and DDL (T) obtain excellent performance and achieve 3.27% and 2.20% Rank-1 accuracy improvements compared with the initial global model, respectively. When combining the DLL (V&T) with the global modal, the performance is boosted by 3.71% Rank-1 accuracy, which demonstrates that the DDL (V) and DDL (T) can prompt the global architecture to explore more discriminative visual and textual local representation. In addition, we add the CIM module to the global network and obtain 1.77% improvement. The results mean the multimodal interaction feature is beneficial for the feature extraction of multimodal data. With all modules adopted together, our model achieves 74.56% (+4.37%) Rank-1 accuracy. The experimental results prove the superiority of the proposed method.

Method	Rank-1	Rank-5	Rank-10
Global	70.19	90.55	95.13
Global + DLL(V)	73.46	92.24	95.94
Global + DLL(T)	72.39	91.36	95.58
Global + DLL(V&T)	73.90	92.37	96.00
Global + CIM	71.96	91.42	95.26
Global + CIM+DLL(V&T)	**74.56**	**92.56**	**96.30**

Table 2: Comparison of various proposed components on the CUHK-PEDES dataset.

Exploring the numbers of the learnable transformer encoder for the performance. We conduct the following ablation studies to evaluate the numbers of the learnable Transformer encoder (LTE) for the performance. As presented in Table 3, we can find that the capability of the network boosts as the number of LTEs increases and achieves excellent results. However, when $L > 4$, the performance tends to decline. The reason is that when $L < 4$, the parameters of LTE layers are not sufficient to completely fit the training textual data. When $L > 4$, the parameters of LTE

Method	Rank-1	Rank-5	Rank-10
$L = 1$	68.45	89.05	94.38
$L = 2$	72.06	91.13	95.45
$L = 3$	74.27	92.11	96.10
$L = 4$	**74.56**	**92.56**	**96.30**
$L = 5$	73.26	91.72	95.52
$L = 6$	74.07	91.91	95.97

Table 3: Comparison of different numbers of LTE on the CUHK-PEDES dataset.

Method	Number	Rank-1	Rank-5	Rank-10
Horizontal Split	$P = 2$	72.71	91.72	95.39
	$P = 3$	73.91	91.72	95.71
	$P = 4$	73.33	91.91	95.81
	$P = 6$	73.59	91.78	96.07
	$P = 8$	73.26	92.11	**96.43**
Ours	$P = 2$	73.26	92.07	95.91
	$P = 3$	73.75	91.88	96.07
	$P = 4$	73.23	91.88	95.97
	$P = 6$	**74.56**	**92.56**	96.30
	$P = 8$	73.68	91.98	96.13

Table 4: Comparison of different visual split methods and different numbers of visual split groups on the CUHK-PEDES dataset. ($Q = 4$)

Method	Number	Rank-1	Rank-5	Rank-10
Equidistant Split	$Q = 2$	73.39	91.62	95.65
	$Q = 4$	73.20	92.07	95.91
	$Q = 8$	72.58	91.88	95.52
Ours	$Q = 2$	73.23	91.88	**96.49**
	$Q = 4$	**74.56**	**92.56**	96.30
	$Q = 8$	73.49	92.50	95.68

Table 5: Comparison of different visual backbones on the CUHK-PEDES dataset. ($P = 6$)

Method	Rank-1	Rank-5	Rank-10
Without CIM	72.90	92.37	95.94
$K = 1$	73.78	91.52	95.97
$K = 2$	**74.56**	92.56	**96.30**
$K = 3$	73.88	**93.11**	96.20
$K = 4$	72.90	91.62	95.78
$K = 5$	73.39	91.88	95.97

Table 6: Comparison of various numbers of CIM on the CUHK-PEDES dataset.

layers are too much and the training data fail to train it well.

Comparison with different numbers of split groups and different split methods. The performance of the proposed DLL module is summarized in Table 4 and Table 5, respectively. To explore the influence of various visual and textual split groups, we conduct experiments and compare the accuracy. P and Q denote the number of visual split regions and textual split regions, respectively. With the increase of split regions, the performance shows an upward trend in the beginning. When $P = 6$ and $Q = 4$, our model achieves the best Rank-1 accuracy. At the later stage, the performance begins to decrease after the best one. Compared with the result of $P = 6$, the accuracy of $P = 8$ declined by 0.88%. The possible reason is that when there are too many groups divided, the patch tokens obtained by each group decrease and the local information learned by the global feature token also reduces. In this case, it is difficult for the model to extract the discriminative fine-grained details. Moreover, to explore the effectiveness of the DDL module, we perform various experiments based on the different split methods of visual and textual information. We select the classical horizontal split method for images and the equidistant split method for descriptions. Comparing the results with the above methods, we can observe that our disordered grouping approach improves the Rank-1 accuracy with 0.97% in terms of vision and 1.36% in terms of text. Accordingly, the proposed visual and textual DDL module is superior to the existing local feature learning method. The experiments suggest that our method is excellent for extracting partial information and is effective for description-based person search.

Exploring the numbers of CIM for the performance. In Table 6, we investigate the effect of the numbers of CIM on the performance of the proposed method. From the results, we can observe that the best Rank-1 accuracy is achieved when $K = 2$. When $K < 2$, the information of the two modalities cannot be fully interactively integrated. However, with the increase of CIM numbers, the performance shows a downtrend. An explanation is that as the layers of CIM increase, the limited training data will lead to the CIM module overfitting and fail to be efficiently learned by the model.

Conclusion

In this paper, we propose a multi-modal disordered representation learning network for description-based person search to fully extract the discriminative visual and textual features. Through this network, we extract the discriminative local features by a disorder local learning strategy and obtain the interactive features by combining the visual and textual information to impel the parallel branches to perceive the correlation between modalities in the training phase. The overall framework has strong discriminative representation learning ability without any other extra auxiliary model. Extensive experiments are performed on CUHK-PEDES and ICFG-PEDES datasets, and the results confirm that our method outperforms the existing advanced methods and achieves state-of-the-art performance.

Acknowledgments

This work is supported by the Project of Sichuan Science and Technology Department (No. 2022YFG0153 and No.

2022YFG0261), Opening Project of State Key Laboratory of Explosion Science and Technology from Beijing Institute of Technology (No. KFJJ23-06M), and the National Nature Science Foundation of China (No. 62271334).

References

Aggarwal, S.; Radhakrishnan, V. B.; and Chakraborty, A. 2020. Text-based person search via attribute-aided matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2617–2625.

Chen, D.; Li, H.; Liu, X.; Shen, Y.; Shao, J.; Yuan, Z.; and Wang, X. 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.

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.

Chen, Z.; Zhou, Z.; Huang, J.; Zhang, P.; and Li, B. 2020. Frame-guided region-aligned representation for video person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10591–10598.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4477–4485.

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.

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.; Si, C.; Wang, J.; Wang, W.; Wang, L.; and Tan, T. 2020. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11189–11196.

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.; Yang, W.; and Wang, X. 2017a. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*, 1890–1899.

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

Mao, X.; Cao, J.; Li, D.; Jia, X.; and Zheng, Q. 2020. Integrating coarse granularity part-level features with supervised global-level features for person re-identification. *arXiv preprint arXiv:2010.07675*.

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.

Niu, K.; Huang, Y.; and Wang, L. 2020. Textual dependency embedding for person search by language. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4032–4040.

Park, H.; and Ham, B. 2020. Relation network for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11839–11847.

Sarafianos, N.; Xu, X.; and Kakadiaris, I. A. 2019. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5814–5824.

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.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, 480–496.

Wang, C.; Luo, Z.; Lin, Y.; and Li, S. 2021. Text-based Person Search via Multi-Granularity Embedding Learning. In *IJCAI*, 1068–1074.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, 274–282.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 402–420. Springer.

Wei, L.; Zhang, S.; Yao, H.; Gao, W.; and Tian, Q. 2018. GLAD: Global–local-alignment descriptor for scalable person re-identification. *IEEE Transactions on Multimedia*, 21(4): 986–999.

Wu, Y.; Yan, Z.; Han, X.; Li, G.; Zou, C.; and Cui, S. 2021. LapsCore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1624–1633.

Yang, F.; Wang, X.; Zhu, X.; Liang, B.; and Li, W. 2022. Relation-based global-partial feature learning network for video-based person re-identification. *Neurocomputing*, 488: 424–435.

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.

Zhang, Z.; Lan, C.; Zeng, W.; and Chen, Z. 2020. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10407–10416.

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.