

Learning Granularity-Unified Representations for Text-to-Image Person Re-identification

Zhiyin Shao
South China University of Technology
Guangzhou, China
eezyshao@mail.scut.edu.cn

Xinyu Zhang
Baidu VIS
Beijing, China
zhangxinyu14@baidu.com

Meng Fang
University of Liverpool
Liverpool, United Kingdom
Meng.Fang@liverpool.ac.uk

Zhifeng Lin
South China University of Technology
Guangzhou, China
eezhifenglin@mail.scut.edu.cn

Jian Wang
Baidu VIS
Beijing, China
wangjian33@baidu.com

Changxing Ding*
South China University of Technology
Guangzhou, China
chxding@scut.edu.cn

ABSTRACT

Text-to-image person re-identification (ReID) aims to search for pedestrian images of an interested identity via textual descriptions. It is challenging due to both rich intra-modal variations and significant inter-modal gaps. Existing works usually ignore the difference in feature granularity between the two modalities, *i.e.*, the visual features are usually fine-grained while textual features are coarse, which is mainly responsible for the large inter-modal gaps. In this paper, we propose an end-to-end framework based on transformers to learn granularity-unified representations for both modalities, denoted as LGUR. LGUR framework contains two modules: a Dictionary-based Granularity Alignment (DGA) module and a Prototype-based Granularity Unification (PGU) module. In DGA, in order to align the granularities of two modalities, we introduce a Multi-modality Shared Dictionary (MSD) to reconstruct both visual and textual features. Besides, DGA has two important factors, *i.e.*, the cross-modality guidance and the foreground-centric reconstruction, to facilitate the optimization of MSD. In PGU, we adopt a set of shared and learnable prototypes as the queries to extract diverse and semantically aligned features for both modalities in the granularity-unified feature space, which further promotes the ReID performance. Comprehensive experiments show that our LGUR consistently outperforms state-of-the-arts by large margins on both CUHK-PEDES and ICFG-PEDES datasets. Code will be released at <https://github.com/ZhiyinShao-H/LGUR>.

CCS CONCEPTS

• **Information systems** → *Top-k retrieval in databases.*

KEYWORDS

Person Re-identification; Text-to-image Retrieval

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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/10.1145/3503161.3548028>

ACM Reference Format:

Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. 2022. Learning Granularity-Unified Representations for Text-to-Image Person Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3503161.3548028>

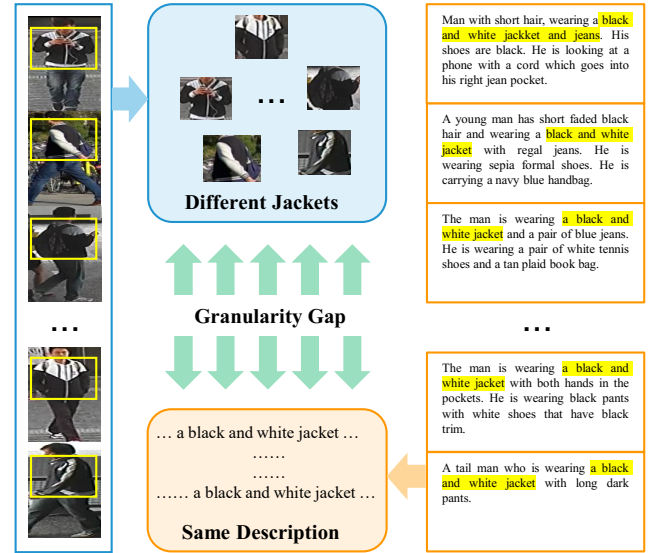


Figure 1: While textual descriptions on the jackets in the all above images are the same, these jackets do in fact differ in terms of their visual details. This example well reflects the granularity gap between the two modalities, *i.e.*, the visual information is fine-grained while the textual features are coarser.

1 INTRODUCTION

Text-to-image person re-identification (ReID) is a cross-modal retrieval task that searches for images of the target identity based on natural language descriptions [22]. Compared with images, natural language descriptions are more flexible and easier to obtain under certain circumstances. The text-to-image ReID task thus attracts much attention. However, text-to-image ReID is also significantly

more challenging than the image-based ReID [7, 11, 28, 32, 38–40, 44] due to the dramatic modality gap between vision and language.

One main aspect of the modality gap relates to the feature granularity. Generally, the visual feature contains *fine-grained* information, while the textual feature describes *coarse* attributes. This results in the same textual description being applicable to similar yet different image patches. For example, in Figure 1, there are several jackets that share the same description “black and white”; however, these jackets differ in their visual details. This difference in feature granularity enlarges the modality gap and makes the text-to-image ReID more challenging.

The modality gap caused by the feature granularity tends to be ignored by existing works. In fact, the “granularity gap” mentioned in the existing text-to-image ReID literature typically refers to the situation in which one word may correspond to image patches of dramatically different sizes [3, 16, 21, 22]. These existing approaches do not explicitly solve the modality gap in which ~~an image patch contains more fine-grained information than its corresponding words~~. Meanwhile, the common solution in these works is to apply cross-modal attention operations that build the correspondence between image patches and words, facilitating adaptation to the changing size of image patches.

In this paper, we focus on the true granularity gap brought by the fine-grained images and the coarse textual descriptions. We propose a novel **Learning Granularity-Unified Representations** (LGUR) framework, which maps both visual and textual features into a granularity-unified feature space. The LGUR framework contains a **Dictionary-based Granularity Alignment (DGA) module** and a **Prototype-based Granularity Unification (PGU) module**.

In the DGA module, we propose to reconstruct both textual and visual features via a **Multi-modality Shared Dictionary** (MSD) based on a transformer layer. In the dictionary, we store a set of **granularity-unified atoms**. The reconstruction operation aims to reduce the feature granularity gap between the two modalities based on these atoms. Intuitively, the information bottleneck lies in coarse textual features; therefore, the granularity of atoms in MSD should be as close as possible to the textual granularity.

However, without explicit guidance, it is hard to drive MSD to closely approximate the granularity of textual features. To address this problem, we introduce the following two strategies in DGA. One is that we guide the learning of MSD parameters using textual features. More specifically, we reconstruct the visual feature again, using its matched textual feature as value in the same transformer layer as above. By reducing the gap between the reconstructed visual features in these two ways, the MSD is forced to be optimized according to the granularity of the textual features. Note that we adopt this guidance during training only; therefore, it introduces no additional computational cost in the inference stage. Another one is that we enable MSD to focus on the foreground pedestrian body. It is because the background is typically ignored by linguistic descriptions and is therefore less relevant to the text-to-image ReID task. Based on the foreground reconstruction, the optimization difficulty of MSD is significantly reduced.

In the PGU module, we further project the textual and visual features into a unified format by a set of shared and learnable prototypes of one transformer layer. These prototypes extract discriminative and diverse features from the two modalities independently

via the cross-attention architecture. Through matching the paired textual and visual features produced by the same prototype, the granularity gap between the two modalities can be further reduced. Meanwhile, thanks to the use of shared prototypes as the query, the computational cost of LGUR is substantially diminished. In comparison, for methods adopting cross-modal attention operations, the visual and textual features are adopted as queries and values in turn; therefore, every image and text must be paired to get the retrieval features, resulting in a heavy computational cost.

We conduct extensive experiments on two existing large-scale benchmark datasets, *i.e.*, CUHK-PEDES [22] and ICFG-PEDES [6]. The results show that our simple LGUR framework consistently and significantly outperforms existing approaches. Compared with many existing methods [9, 18, 25], LGUR is also more efficient, since it does not require cross-modal attention operations between each image-text pair in the testing stage. More impressively, we find that LGUR performs well in domain generalization tasks due to the feature unification on the granularity level. The main contributions of the proposed method can be summarized as follows:

- We identify the difference in the feature granularity between the visual and textual modalities that results in the modality gap, which is an important element that is rarely considered in the text-to-image ReID literature.
- We propose a novel Learning Granularity-Unified Representations (LGUR) framework that efficiently extracts granularity-unified features from both modalities.
- Extensive experiments on two text-to-image ReID datasets, *i.e.*, CUHK-PEDES and ICFG-PEDES show that LGUR consistently outperforms the state-of-the-arts by large margins.

2 RELATED WORK

2.1 Vision-Language Models

Transformers have demonstrated their superiority on many vision and natural language processing tasks. Multiple works have also been developed that apply the transformer to the vision-language pre-training (VLPT) task [4, 19, 20, 24, 30, 31, 33, 48]. Depending on their model structure, existing VLPT methods can be categorized as either two-stream or single-stream models. Both types of methods extract vision-language joint features. The two-stream models [15, 30, 31, 33] extract features from the image and text modalities separately, then fuse them by means of the transformer structure. For their part, the single-stream models [4, 19, 20, 24, 48] adopt the BERT [5] model and process the image feature and the language feature together as a joint distribution. However, the above approaches require the text and image pair to be fed into the network. Specifically, in the testing stage of text-to-image retrieval tasks, every textual query needs to be paired with each image in the gallery, which introduces high computational complexities.

2.2 Text-to-Image Person ReID

Due to its fine-grained nature, text-to-image person ReID is more challenging than general cross-modal retrieval tasks. Depending on the alignment strategy utilized, existing works can be divided into cross-modal attention-based methods and cross-modal attention-free methods. The latter type of methods design various model

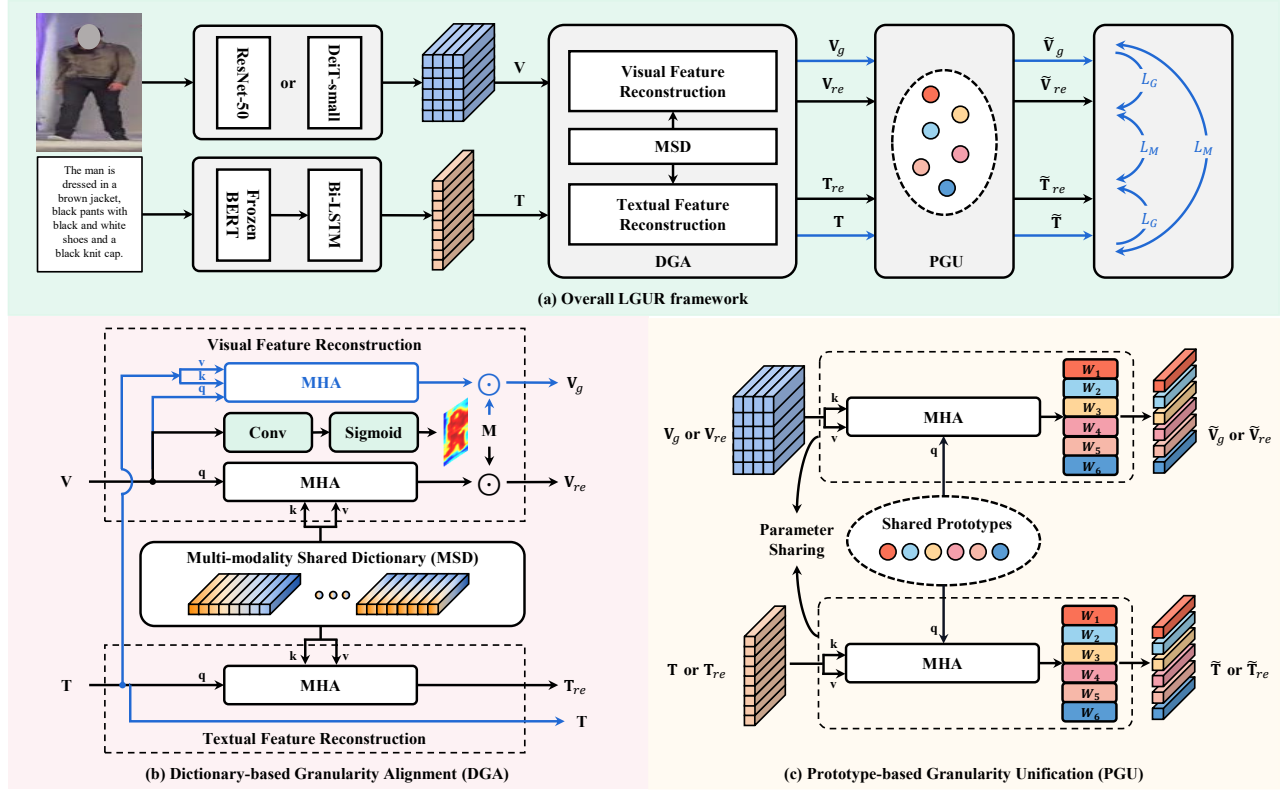


Figure 2: Overview of the LGUR framework (shown in a), which includes a Dictionary-based Granularity Alignment (DGA) module (shown in b) and a Prototype-based Granularity Unification (PGU) module (shown in c) to achieve feature extraction that is both efficient and granularity-unified. DGA reconstructs both visual and textual features via a Multi-modality Shared Dictionary (MSD). Moreover, we propose a cross-modal guidance strategy to optimize the MSD parameters according to the granularity of the textual features. In addition, a foreground mask is utilized to enable MSD to focus on the reconstruction of the pedestrian body. PGU projects both textual and visual features into a unified format via a set of shared and learnable prototypes. LGUR does not need to implement any cross-modal attention operations in the testing stage and is therefore computationally efficient. The blue arrows represent operations that are discarded during the testing stage. MHA represents the multi-head attention module. Best viewed in color.

structures [41, 47] or objective functions [8, 10, 21, 23, 27, 45] to align the features from both modalities in a shared feature space.

In comparison, cross-modal attention-based methods focus on establishing region-word [3, 21, 22] or region-phrase [16, 25] correspondences. These methods have their own advantages as well as disadvantages. Cross-modal attention-free methods are usually more efficient. Specifically, given N images and M sentences, their complexities are $O(M+N)$. By contrast, the complexity of the cross-modal attention-based methods increases to $O(MN)$ [37]. However, these methods usually achieve significantly better retrieval performance as they better reduce the modality gaps.

The granularity gap mentioned in existing cross-modal attention-based methods [16, 25, 36] usually refers to the situation in which each word may correspond to image patches of dramatically different sizes. However, these methods rarely pay attention to the modality gap in feature granularity, *i.e.*, similar but different image regions may share the same textual description. In this paper, we handle this new problem by means of a novel LGUR framework. As

LGUR avoids cross-modal attention operations between image and text through the use of a modality-shared dictionary, it has great advantages in computational efficiency.

3 METHODOLOGY

The overview of our LGUR framework is illustrated in Figure 2. LGUR comprises three modules: the Feature Extraction Backbones (see Section 3.1), the Dictionary-based Granularity Alignment Module (see Section 3.2) and the Prototype-based Granularity Unification Module (see Section 3.3). The latter two modules are used to enhance the granularity unification of the image and text modalities. The optimization of the overall framework is described in Section 3.4.

3.1 Feature Extraction Backbones

Visual modality. Let $V \in \mathbb{R}^{HW \times d}$ represents the visual feature produced by the visual backbone, while d denotes the feature dimension. We consider two backbones, namely DeiT-Small [34]

and ResNet50 [13]. For DeiT-Small, we split the image into $H \times W$ patches, so that HW denotes the number of patch tokens. As for ResNet50, H and W are the height and width respectively of its output feature maps. These two backbones have a similar number of parameters.

Textual modality. Let $\mathbf{T} \in \mathbb{R}^{L \times d}$ represents the output of the textual backbone. We adopt a light-weight bidirectional long short-term memory network (Bi-LSTM) [14] to extract textual features. The input embeddings of Bi-LSTM are obtained from a pretrained BERT model [5]. L stands for the number of words in the description. The backbones are introduced in more details in Section 4.2.

3.2 Dictionary-based Granularity Alignment

As analyzed in Section 1, the textual granularity is coarse while the visual granularity is fine-grained. This large granularity gap creates inconsistency between the two types of feature representations. We conjecture that the visual and textual features will match with each other more tightly if the former are made to be more abstract. To this end, we introduce a **Dictionary-based Granularity Alignment (DGA) module** to reconstruct the representations of both modalities, such that these features can be made consistent in a granularity-unified feature space. Our DGA utilizes a **Multi-modality Shared Dictionary (MSD)** to conduct the textual and visual feature reconstruction.

Multi-modality shared dictionary. We build the MSD as $\mathbf{D} \in \mathbb{R}^{s \times d}$, which is randomly initialized. Here, s indicates the number of atoms in \mathbf{D} . Each atom in \mathbf{D} is a d -dimensional vector, which has the same dimension as \mathbf{V} and \mathbf{T} . We expect \mathbf{D} to possess similar granularity to the textual features. In the following, we describe the way in which granularity-aligned visual and textual features can be obtained via reconstruction using \mathbf{D} .

Textual feature reconstruction. We first apply \mathbf{D} to reconstruct textual features. The textual features before and after reconstruction are expected to be similar with each other. To this end, we minimise the similarity of these two textual features using a ranking loss, which will be described in Section 3.4. This strategy ~~drives the atoms in \mathbf{D} to possess similar granularity to that of the text~~.

Formally, we utilize a transformer's **cross-attention operation** as the reconstruction process [35], in which \mathbf{T} is utilized as the query while \mathbf{D} acts as the key and value. The reconstructed textual feature \mathbf{T}_{re} can be expressed as follows:

$$\mathbf{T}_{re} = MHA_1(\mathbf{T}, \mathbf{D}, \mathbf{D}), \quad (1)$$

where $MHA_1(\cdot)$ denotes a transformer block, which consists of a multi-head attention and a feed-forward network [35]. More formally, $MHA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = FFN(MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}))$, where \mathbf{Q} , \mathbf{K} , and \mathbf{V} are abbreviations for query, key, and value, respectively.

Visual feature reconstruction. We also reconstruct the visual features via $MHA_1(\cdot)$ with \mathbf{V} as the query and \mathbf{D} as the key and value. The reconstructed visual feature is denoted as $\mathbf{V}_{re} \in \mathbb{R}^{HW \times d}$.

Different from the textual feature reconstruction, we propose the following two strategies in the visual reconstruction to further reduce the modality gap between \mathbf{V}_{re} and \mathbf{T}_{re} . First, we enable MSD to focus on the reconstruction of the image foreground, i.e., the pedestrian body. This is based on the consideration that pedestrian images are usually characterized by rich occlusion and background clutters, while textual descriptions are identity-centric and tend to

ignore the background noises. Therefore, we generate a foreground mask $\mathbf{M} \in \mathbb{R}^{HW \times 1}$ via the spatial attention mechanism [29]. Specifically, we attach a 1×1 convolutional layer with a sigmoid function to \mathbf{V} to obtain \mathbf{M} . As illustrated in Figure 2, the reconstructed visual features from MSD based on the foreground restriction is shown as follows:

$$\mathbf{V}_{re} = MHA_1(\mathbf{V}, \mathbf{D}, \mathbf{D}) \odot \mathbf{M}, \quad (2)$$

where \odot represents the Hadamard product between \mathbf{M} and each column of $MHA_1(\mathbf{V}, \mathbf{D}, \mathbf{D})$. With the help of \mathbf{M} , the reconstructed background clutters are suppressed and the optimization difficulty of MSD is alleviated.

Second, we guide the granularity of \mathbf{V}_{re} to be abstract under the help of the textual features. More specifically, we reconstruct \mathbf{V} again using \mathbf{T} from the paired language description with the input image. \mathbf{T} acts as both the key and value in the transformer block:

$$\mathbf{V}_g = MHA_1(\mathbf{V}, \mathbf{T}, \mathbf{T}) \odot \mathbf{M}, \quad (3)$$

where $\mathbf{V}_g \in \mathbb{R}^{HW \times d}$.

Compared with \mathbf{V}_{re} , the granularity of \mathbf{V}_g is closer to that of textual features. Therefore, we impose a ranking loss to penalize the difference between \mathbf{V}_{re} and \mathbf{V}_g , as detailed in Section 3.4.

3.3 Prototype-based Granularity Unification

In Section 3.2, we align the granularity between image and text via MSD. Despite this, the model's ability to accurately match texts and images of a specific identity remains limited. In fact, what \mathbf{D} has learned is general semantic knowledge. In this subsection, we aim to extract more powerful features for the ReID purpose via a Prototype-based Granularity Unification (PGU) module. PGU projects both textual and visual features into a unified format, which further aligns the granularity of both modalities. More specifically, we design a set of prototypes $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_K] \in \mathbb{R}^{d \times K}$, which are randomly initialized. The K prototypes contain diverse semantic information. To enable these prototypes to capture both textual and visual features, we let each prototype act as the query in the transformer layer, while textual or visual feature acts as both the key and value. For simplicity, we define \mathbf{F} as an example to represent a textual or visual feature. As shown in Figure 2, the refined example $\tilde{\mathbf{F}} \in \mathbb{R}^{K \times d'}$ after PGU is defined as follows:

$$\begin{aligned} \tilde{\mathbf{F}} &= PGU(\mathbf{P}, \mathbf{F}) = Concat(f_1(\mathbf{p}_1, \mathbf{F}), \dots, f_K(\mathbf{p}_K, \mathbf{F})), \\ f_i(\mathbf{p}_i, \mathbf{F}) &= \mathbf{W}_k(MHA_2(\mathbf{p}_i, \mathbf{F}, \mathbf{F})), \end{aligned} \quad (4)$$

where $\mathbf{W}_k \in \mathbb{R}^{d' \times d}$ denotes an FC layer for the k -th query \mathbf{p} . We apply independent FC layers for the queries in order to produce semantically diverse features. Meanwhile, each query adopts the same FC layer for both modalities, which further aligns their feature granularity. d' is the output dimension after FC. *Concat* denotes the concatenation operation. MHA_2 is of the same structure as MHA_1 in Eq. (1) albeit with independent parameters. Based on Eq. (4), we obtain the granularity-unified feature $\tilde{\mathbf{T}}_{re}$ of \mathbf{T}_{re} in Eq. (1), $\tilde{\mathbf{V}}_{re}$ of \mathbf{V}_{re} in Eq. (2), $\tilde{\mathbf{T}}$ of \mathbf{T} , and $\tilde{\mathbf{V}}_g$ of \mathbf{V}_g in Eq. (3). More formally,

$$\begin{aligned} \tilde{\mathbf{T}}_{re} &= PGU(\mathbf{P}, \mathbf{T}_{re}), \tilde{\mathbf{V}}_{re} = PGU(\mathbf{P}, \mathbf{V}_{re}), \\ \tilde{\mathbf{T}} &= PGU(\mathbf{P}, \mathbf{T}), \tilde{\mathbf{V}}_g = PGU(\mathbf{P}, \mathbf{V}_g). \end{aligned} \quad (5)$$

Here, we use a single shared MHA_2 in Eq. (5). After PGU, the granularities of the visual and textual features are aligned to a unified space, which substantially reduces the granularity gap.

Discussion. Previous works [3, 9, 16, 21, 22, 25] usually rely on a cross-model attention between two modalities to reduce modality gaps. However, all image-text pairs need to be fed into this models of this kind, which is very computationally costly. In comparison, the prototypes in our PGU are unified representations for both modalities. For a single piece of text or image, we directly obtain the feature \tilde{T}_{re} or \tilde{V}_{re} rather than first composing text-to-image pairs; thus, our PGU is more computationally efficient. Meanwhile, the format of our prototype is similar to the object query in [2]. However, in [2], the output corresponding to each object query represents a potential object instance, while the output corresponding to each prototype in our PGU denotes a discriminative region of one pedestrian. The prototypes in PGU thus contain more detailed information.

3.4 Optimization & Inference

Optimization. Inspired by [32, 45], we adopt cross-entropy loss as the identification loss for each prototype. For a specific feature \tilde{F} in Eq. (4), we denote the predicted identity probabilities by the k -th prototype as \hat{y}_k . The identification loss can thus be written as:

$$L_{ID}(\tilde{F}) = \frac{1}{K} \sum_{k=1}^K -y \odot \log(\hat{y}_k), \quad (6)$$

where y is the ground-truth label vector.

Meanwhile, ranking loss is commonly applied to the text-to-image ReID task. For two features \tilde{F}_1 and \tilde{F}_2 from one matched image-text pair, the ranking loss is formulated as follows:

$$L_{RK}(\tilde{F}_1, \tilde{F}_2) = \max(\alpha - S(\tilde{F}_1, \tilde{F}_2^+) + S(\tilde{F}_1, \tilde{F}_2^-), 0) + \max(\alpha - S(\tilde{F}_2, \tilde{F}_1^+) + S(\tilde{F}_2, \tilde{F}_1^-), 0), \quad (7)$$

where $\tilde{F}_1^+/\tilde{F}_2^+$ and $\tilde{F}_1^-/\tilde{F}_2^-$ are one positive sample and one semi-hard negative sample of \tilde{F}_1/\tilde{F}_2 in a mini-batch, respectively. In addition, α is a margin hyper-parameter, while S denotes the cosine similarity metric.

By applying Eq. (6) and Eq. (7) to LGUR, we obtain the following loss:

$$L_M = L_{ID}(\tilde{T}_{re}) + L_{ID}(\tilde{V}_{re}) + L_{ID}(\tilde{T}) + L_{ID}(\tilde{V}_g) + L_{RK}(\tilde{T}_{re}, \tilde{V}_{re}) + L_{RK}(\tilde{T}, \tilde{V}_g). \quad (8)$$

Moreover, to achieve tighter granularity alignment, we impose another loss function based on the guidance features, \tilde{V}_g and \tilde{T} , as described in Section 3.2. Specifically, we adopt the ranking loss to pull the reconstructed features closer to the guidance features when they refer to the same person, or push them away when they refer to different identities. The guidance loss can be represented as follows:

$$L_G = L_{RK}(\tilde{T}_{re}, \tilde{T}) + L_{RK}(\tilde{V}_{re}, \tilde{V}_g). \quad (9)$$

The overall loss function can thus be expressed as:

$$L = L_M + L_G. \quad (10)$$

Inference. We separately extract the textual feature \tilde{T}_{re} and the visual feature \tilde{V}_{re} for the text-to-image retrieval. \tilde{T} and \tilde{V}_g are

Table 1: Performance comparisons on CUHK-PEDES.

Methods		Rank-1	Rank-5	Rank-10
ResNet-50	Dual Path [47]	44.40	66.26	75.07
	CMPM/C [45]	49.37	-	79.27
	MIA [25]	53.10	75.00	82.90
	A-GANet [23]	53.14	74.03	82.95
	GALM [16]	54.12	75.45	82.97
	TIMAM [27]	54.51	77.56	84.78
	TDE [26]	55.25	77.46	84.56
	VTa [10]	55.32	77.00	84.26
	SCAN [18]	55.86	75.97	83.69
	ViTAA [41]	55.97	75.84	83.52
	CMAAM [1]	56.68	77.18	84.86
	HGAN [46]	59.00	79.49	86.62
	NAFS [9]	59.94	79.86	86.70
	DSSL [49]	59.98	80.41	87.56
	MGEL [36]	60.27	80.01	86.74
	SSAN [6]	61.37	80.15	86.73
	Han <i>et al.</i> [12]	61.65	80.98	86.78
	LapsCore [43]	63.40	-	87.80
	LGUR	64.21	81.94	87.93
	LGUR (DeiT-Small)	65.25	83.12	89.00

abandoned during inference. The cosine similarity is adopted as the metric for retrieval.

4 EXPERIMENTS

We evaluate the LGUR framework on two datasets, namely CUHK-PEDES and ICFG-PEDES. We further adopt the Rank-1, Rank-5, and Rank-10 accuracies as metrics to evaluate performance on both databases.

4.1 Datasets and Evaluation Metrics

CUHK-PEDES. CUHK-PEDES [22] contains 40,206 images and 80,412 textual descriptions for 13,003 identities. The training set comprises 34,050 images and 68,120 textual descriptions of 11,000 pedestrians. The testing set includes 3,074 images and 6,156 textual descriptions of the rest 1,000 pedestrians. Each image contains at least two textual descriptions, each of which is made up of 23.5 words on average.

ICFG-PEDES. ICFG-PEDES [6] contains 54,522 pedestrian images of 4,102 different identities, all of which were collected from the MSMT17 database [42]. The training set includes 34,674 images of 3,102 pedestrians. The testing set comprises 19,848 images of 1,000 pedestrians. Each image is associated with only one textual description; these descriptions contain 37.2 words on average.

Evaluation metrics. We adopt the popular Rank- k metrics ($k=1,5,10$) as the evaluation metrics. Rank- k reveals the probability that, when given a textual description as query, we can find at least one matching person image in the top- k candidate list.

4.2 Implementation Details

In our experiments, we choose to use DeiT-Small [34] with a patch size of 16 and ResNet-50 [13] as the visual backbones, respectively.

Table 2: Performance comparisons on ICFG-PEDES.

Methods		Rank-1	Rank-5	Rank-10
ResNet-50	Dual Path [47]	38.99	59.44	68.41
	CMPM/C [45]	43.51	65.44	74.26
	MIA [25]	46.49	67.14	75.18
	SCAN [18]	50.05	69.65	77.21
	ViTAA [41]	50.98	68.79	75.78
	SSAN [6]	54.23	72.63	79.53
	LGUR	57.42	74.97	81.45
LGUR (DeiT-Small)		59.02	75.32	81.56

Table 3: Performance comparisons on the domain generalization task. “C” denotes CUHK-PEDES, while “I” represents ICFG-PEDES.

Methods		Rank-1	Rank-5	Rank-10
$C \rightarrow I$	Dual Path [47]	15.41	29.80	38.19
	MIA [25]	19.35	36.78	46.42
	SCAN [18]	21.27	39.26	48.83
	SSAN [6]	24.72	43.43	53.01
	SSAN(w/ BERT) [6]	29.24	49.00	58.53
	LGUR	34.25	52.58	60.85
$I \rightarrow C$	Dual Path [47]	7.63	17.14	23.52
	MIA [25]	10.93	23.77	32.39
	SCAN [18]	13.63	28.61	37.05
	SSAN [6]	16.68	33.84	43.00
	SSAN(w/ BERT) [6]	21.07	38.94	48.54
	LGUR	25.44	44.48	54.39

We attach one 1×1 convolutional layer to the visual backbone to project its output to d -dim. For the textual modality, the sequence of word embeddings extracted from BERT is then fed to a Bi-LSTM. Note that we “freeze” the weights of BERT, similar to [27], and only fine-tune Bi-LSTM. We resize all images to 384×128 pixels and use only random horizontal flipping as the data augmentation. We set the feature dimension d for both the image and text to 384. d' is set to 512. The dictionary size s is 400 and the margin α is set to 0.3. The number of shared prototypes K is set to 6. During training, we adopt the Adam optimizer [17]. The batch size is 64, and the number of epochs is 60. The initial learning rate of DeiT-Small is set to 0.0001, while the others are set to 0.001.

4.3 Comparisons with State-of-the-Art Methods

To facilitate fair comparison, we evaluate the performance of LGUR with DeiT-Small and ResNet-50 as the visual backbone, respectively. **Comparisons on CUHK-PEDES.** In Table 1, our LGUR outperforms all state-of-the-art methods, achieving Rank-1 accuracy of 64.21% and 65.25% based on ResNet-50 and DeiT-Small, respectively. In particular, our LGUR outperforms NAFS [9] (which also adopts BERT for textual feature extraction) by as much as 4.27% in terms of Rank-1 accuracy. Moreover, NAFS requires cross-modality attention operations, which are computationally expensive. By contrast, LGUR extracts textual and visual features independently and

Table 4: Performance comparisons in terms of time complexity. “CAM” refers to the cross-modal attention mechanism [18].

	CAM	Methods	Train	Inference	Rank-1
CUHK-PEDES	✓	MIA [25]	680ms	42ms	53.10
	✓	SCAN [18]	718ms	46ms	55.86
	✓	NAFS [9]	1,284ms	42ms	59.94
	×	Dual Path [47]	321ms	10ms	44.40
	×	CMPM/C [45]	338ms	27ms	49.37
	×	SSAN [6]	901ms	76ms	61.37
	×	LGUR (Ours)	886ms	26ms	64.21
ICFG-PEDES	✓	MIA [25]	711ms	113ms	46.49
	✓	SCAN [18]	738ms	114ms	50.05
	✓	NAFS [9]	1,304ms	116ms	-
	×	Dual Path [47]	342ms	11ms	38.99
	×	CMPM/C [45]	356ms	31ms	43.51
	×	SSAN [6]	973ms	77ms	54.23
	×	LGUR (Ours)	910ms	31ms	57.42

thereby substantially reduces the computational cost (as discussed in Section 3.3). LGUR also achieves higher performance than one most recent method, named LapsCore [43]. It is worth noting that LapsCore is based on the NAFS [9] model and therefore also has a much higher computational cost than LGUR. Furthermore, LapsCore focuses on designing auxiliary tasks for regularization and does not consider the granularity gap between the two modalities; therefore, the contributions of LGUR and LapsCore complement each other.

Comparisons on ICFG-PEDES. Comparison results are summarized in Table 2. Since ICFG-PEDES is a new database, we directly cite the performance of existing approaches evaluated in [6]. LGUR consistently achieves the best performance. Specifically, it achieves 57.42% and 59.02% Rank-1 accuracies with the ResNet-50 and DeiT-Small backbones, respectively. SSAN [6] achieves superior performance since it extracts fine-grained part-level textual and visual features. However, this method still ignores the granularity gap between the two modalities. By bridging the granularity gap, LGUR outperforms SSAN by 3.19% in terms of Rank-1 accuracy.

Comparisons on the domain generalization (DG) task. Our LGUR effectively narrows the granularity gap between the textual and visual features. Due to the feature unification on a coarse granularity level, it could be naturally assumed that the model is able to generalize well to the other domains. To this end, we conduct experiments on DG tasks. Here, we directly deploy the model pretrained on the source domain to the target dataset. As shown in Table 3, our LGUR outperforms all other comparison methods. In particular, LGUR achieves Rank-1 improvements of 9.53% and 8.76% on the $C \rightarrow I$ and $I \rightarrow C$ settings respectively when compared to SSAN [6]. To exclude the differences arising from BERT [5], we also equip SSAN with the same textual feature extraction backbone as LGUR. The performance of SSAN increases, yet is still lower than that of LGUR by 5.01% and 4.37% in terms of Rank-1 accuracy. This experiment demonstrates that the granularity-unified representations of text and image have good generalization ability.

Table 5: Ablation study on key modules of LGUR. DGA has three important components: the multi-modality shared dictionary (D), the foreground mask (M), the guidance (G) for reconstruction (T and V_g).

No.	Methods	Components				CUHK-PEDES			ICFG-PEDES		
		PGU	DGA			Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
			D	M	G						
0	baseline	-	-	-	-	58.67	79.08	85.82	52.09	70.91	78.07
1	+ PGU	✓	-	-	-	63.26	81.17	87.69	56.34	73.58	80.23
2	+ DGA	-	✓	✓	✓	61.86	80.43	87.20	55.83	73.65	80.48
3	+ PGU + DGA (only D)	✓	✓	-	-	64.28	81.95	88.11	57.52	74.84	81.14
4	+ PGU + DGA (D + M)	✓	✓	✓	-	64.64	82.91	88.52	57.82	74.99	81.17
5	+ PGU + DGA (D + G)	✓	✓	-	✓	64.80	82.29	88.32	58.17	75.83	81.51
6	LGUR	✓	✓	✓	✓	65.25	83.12	89.00	59.02	75.32	81.56

Table 6: Comparisons with variants of MSD, including without reconstruction (w/o reconstruction), reconstruction with a self-attention layer (w/ SA) [35], reconstruction with modality unshared dictionary (w/ unshared D).

Type of reconstruction	CUHK-PEDES		ICFG-PEDES	
	Rank-1	Rank-5	Rank-1	Rank-5
w/o reconstruction	63.26	81.17	56.34	73.58
w/ SA	63.39	82.28	56.34	74.09
w/ unshared D	61.55	80.69	55.96	73.23
w/ shared D (Ours)	64.28	81.95	57.52	74.84

Table 7: Comparisons between different types of prototypes P in PGU, including the modality shared prototypes (w/ shared P) and the modality unshared prototypes (w/ unshared P).

Type of P	CUHK-PEDES		ICFG-PEDES	
	Rank-1	Rank-5	Rank-1	Rank-5
w/ unshared P	63.78	82.15	57.73	75.11
w/ shared P (Ours)	65.25	83.10	59.02	75.32

4.4 Comparisons on Time Complexity

As discussed in Section 3.3, our LGUR has advantages in time complexity. In this subsection, we evaluate the training time¹, inference time² and Rank-1 accuracy of LGUR, three works that do not implement the cross-modal attention mechanism (*i.e.*, Dual Path [47], CMPM/C [45], and SSAN [6]), and another three methods that adopt cross-attention mechanism (*i.e.*, MIA [25], SCAN [18], and NAFS [9]). To facilitate fair comparison, we set the image size of input images to 384×128 pixels and the batch size to 64 for all methods. All experiments are conducted on a Titan X GPU. As shown in Table 4, LGUR is dramatically more efficient than all methods that rely on cross-modal attention operations. This advantage in efficiency benefits from the disentanglement of the textual and visual feature extraction in LGUR. In addition, the computational cost of LGUR is competitive with the works that do not incorporate cross-modal

¹The training time refers to the average time taken to process one mini-batch of images and descriptions.

²The inference time includes the feature extraction time for a given query and time for similarity computation with all gallery images.

attention mechanisms. For example, LGUR costs 26ms per query on the CUHK-PEDES database, which is faster than SSAN at 76ms. Considering the trade-off between accuracy and efficiency, LGUR is thus superior to other methods.

4.5 Ablation Study

In this subsection, we analyze the effectiveness of each key component in the LGUR framework. Here, we adopt the DeiT-Small as the visual backbone.

Effectiveness of PGU. In Table 5, the efficacy of PGU is revealed via the experimental results of No.0 *v.s.* No.1. Adding PGU on the baseline promotes the Rank-1 accuracy of the baseline by 4.59% on CUHK-PEDES. When comparing the results of No.2 and No.6 experiments, PGU can improve the Rank-1 accuracy from 61.86% to 65.25% on baseline+DGA. The above results clearly show that the unified textual and visual feature representations from PGU are beneficial for improving the performance.

Effectiveness of DGA. The experimental results of No.0 *v.s.* No.2, and No.1 *v.s.* No.6 in Table 5 demonstrate the efficacy of DGA. In particular, when adding DGA to the baseline, the Rank-1 accuracy is promoted by 3.19% and 3.74% on CUHK-PEDES and ICFG-PEDES, respectively. These results justify that DGA well aligns the feature granularity of the two modalities and therefore promotes the retrieval accuracy.

Meanwhile, we provide comprehensive experimental analysis to further explore the impact of each component in DGA.

First, the most important part in DGA is the multi-modality shared dictionary (D). In Table 5, when adopting D on baseline+PGU in experiment No.3, the performance is promoted by 1.02% on CUHK-PEDES, showing that the reconstructed features from D is more effective. This is because via the shared D for the visual and textural features, their granularity is roughly aligned.

Second, we also enable DGA to focus on the reconstruction of foreground visual features. As shown in experiments No.3 and No.4 in Table 5, the foreground mask M further promotes the Rank-1 accuracy from 64.28% to 64.64% on CUHK-PEDES. This result reveals that foreground-oriented reconstruction helps to further reduce the modality gap and relieves the optimization burden of D, as analyzed in Section 3.2. Meanwhile, we also visualize the foreground mask M in Figure 3. It clearly shows that M makes the model to focus on the meaningful pedestrian body instead of the useless background.

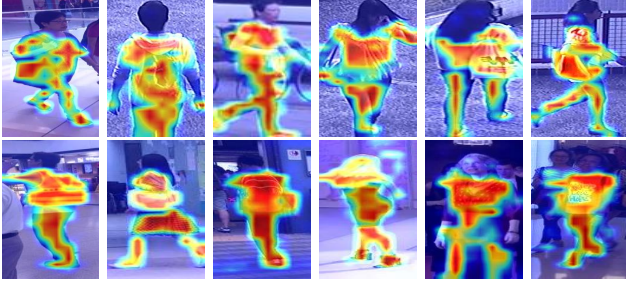


Figure 3: Visualization of the foreground masks M . The pedestrian body areas are highlighted while the cluttered background is suppressed.

Third, we adopt V_g in Eq. (3) without M to guide D to learn atoms of coarse granularities as the same as that in text. As shown by the experimental results No.5 v.s. No.3, the Rank-1 accuracy improves by 0.65% on ICFG-PEDES when V_g is adopted, verifying the effectiveness of the guidance from V_g .

Finally, our LGUR combines D , M , G and PGU together, achieving the best result. For example, LGUR outperforms the baseline by as much as 6.58% and 6.93% in Rank-1 accuracy on CUHK-PEDES and ICFG-PEDES, respectively.

Comparisons with variants of MSD. We adopt the multi-modality shared dictionary (D) to reconstruct both the textual and the visual features. The experiment No.3 in Table 5 has justified the effectiveness of D . In Table 6, we further verify its design via comparison with several possible variants.

First, MSD significantly outperforms the naïve self-attentions layer (SA) in transformer [35]. The result in Table 6 reveals that reconstruction via D is more effective since it aligns the granularity of both modalities. In comparison, the self-attention layer cannot explicitly reduce the granularity gap between the two modalities, and thus is inferior to our MSD. Second, the shared D between the two modalities significantly outperforms its unshared counterpart, as shown in the last two rows in Table 6. Actually, the separate D for the two modalities may enlarge the modality gap, which hurts the ReID performance. Overall, the design of our shared D in MSD is a good choice to narrow down the modality gap.

Comparisons with variants of PGU. The prototypes P in PGU are shared between the image and text modalities. To validate the advantage of this design, we also evaluate the unshared P , i.e., using different P for text and image. In Table 7, it is clear that adopting shared P outperforms the unshared P by large margins. It is due to the superiority of the shared P , which can improve the granularity unification between the two modalities.

4.6 Qualitative Results

Here, we provide qualitative results to demonstrate MSD’s reconstruction ability in Figure 4. Ideally, if one phrase describes an image patch, their cross-attention maps with D should be similar. Otherwise, their cross-attention maps should be different. Considering that the whole attention map is large, here we only show 15 selected attention scores.

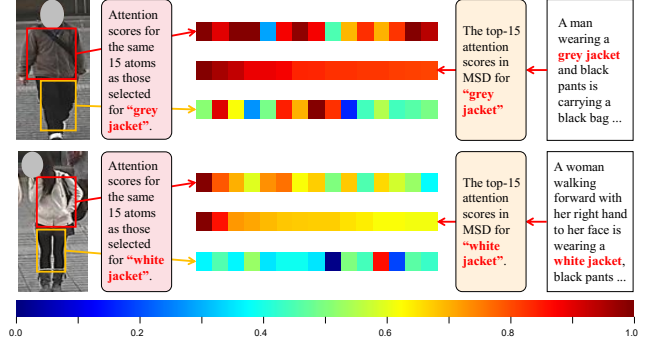


Figure 4: Visualization of the attentions scores between one selected phrase with D (the right part) and the image patch with D (the left part) in the DGA module. The red rectangle denotes the text-matched image patch, while the yellow one represents the text-irrelevant image patch.

Specifically, we first choose one phrase and its corresponding image patch, which are highlighted in red text colour and red rectangle in Figure 4, respectively. We also select one irrelevant image patch to the phrase and frame the patch in yellow. We first draw the top-15 attention scores that represent the highest similarities between the phrase and D . Then we show the attention scores on the same 15 atoms as above between each of the two patches and D . Results in Figure 4 show that the scores for the matched phrase-patch pair are similar, which indicates that the atoms in D can well represent features for both text and image. In contrast, the attention scores for the irrelevant phrase-patch pair are different. This phenomenon indicates that the granularity gap of the two modalities can be reduced after the dictionary-based reconstruction.

5 CONCLUSION

In this paper, we have introduced a novel framework named LGUR to learn granularity-unified representations for the text-to-image ReID task. This framework includes a Dictionary-based Granularity Alignment (DGA) module and a Prototype-based Granularity Unification (PGU) module. In the DGA module, we build a Multi-modality Shared Dictionary (MSD) to reconstruct both visual and textual features, such that their granularity can be unified. We further provide a cross-modal guidance strategy and a foreground mask to facilitate the optimization of MSD parameters. In the PGU module, we adopt a set of shared prototypes for diverse textual and visual feature extraction, which further aligns the granularity of both modalities. Extensive experiments on two large-scale databases demonstrate the effectiveness of our LGUR.

ACKNOWLEDGMENTS

This work was supported by the CCF-Baidu Open Fund, the National Natural Science Foundation of China under Grant 62076101 and 61702193, the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X183, Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011549, and Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004).

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] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.
- [3] Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1879–1887.
- [4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [6] 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).
- [7] Ding, Changxing and Wang, Kan and Wang, Pengfei and Tao, Dacheng. 2022. Multi-Task Learning With Coarse Priors for Robust Part-Aware Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2022), 1474 – 1488.
- [8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. (2018).
- [9] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. 2021. Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search. *arXiv preprint arXiv:2101.03036* (2021).
- [10] Jing Ge, Guangyu Gao, and Zhen Liu. 2019. Visual-Textual Association with Hardest and Semi-Hard Negative Pairs Mining for Person Search. *arXiv preprint arXiv:1912.03083* (2019).
- [11] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. 2019. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3642–3651.
- [12] Xiao Han, Sen He, Li Zhang, and Tao Xiang. 2021. Text-Based Person Search with Limited Data. *arXiv preprint arXiv:2110.10807* (2021).
- [13] 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.
- [14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [15] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing Out of the box: End-to-End Pre-training for Vision-Language Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12976–12985.
- [16] 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 *AAAI*.
- [17] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR* (2015).
- [18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.
- [19] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.
- [20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).
- [21] 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.
- [22] 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.
- [23] 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 *ACMMM*. 665–673.
- [24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265* (2019).
- [25] 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.
- [26] Kai Niu, Yan Huang, and Liang Wang. 2020. Textual dependency embedding for person search by language. In *ACMMM*. 4032–4040.
- [27] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial Representation Learning for Text-to-Image Matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 5814–5824.
- [28] Walter J Scheirer, Patrick J Flynn, Changxing Ding, Guodong Guo, Vitomir Struc, Mohamad Al Jazaery, Klemen Grm, Simon Dobrisek, Dacheng Tao, Yu Zhu, et al. 2016. Report on the BTAS 2016 video person recognition evaluation. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–8.
- [29] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1179–1188.
- [30] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VI-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).
- [31] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7464–7473.
- [32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 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.
- [33] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).
- [34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [36] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. 2021. Text-based Person Search via Multi-Granularity Embedding Learning. *IJCAI*.
- [37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*. 274–282.
- [38] Kan Wang, Pengfei Wang, Changxing Ding, and Dacheng Tao. 2021. Batch coherence-driven network for part-aware person re-identification. *IEEE Transactions on Image Processing* 30 (2021), 3405–3418.
- [39] Pengfei Wang, Changxing Ding, Zhiyin Shao, Zhibin Hong, Shengli Zhang, and Dacheng Tao. 2022. Quality-aware part models for occluded person re-identification. *IEEE Transactions on Multimedia* (2022).
- [40] Pengfei Wang, Changxing Ding, Wentao Tan, Mingming Gong, Kui Jia, and Dacheng Tao. 2022. Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. *IEEE Transactions on Multimedia* (2022).
- [41] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. 2020. ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language. (2020).
- [42] 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.
- [43] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. 2021. LapsCore: Language-Guided Person Search via Color Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1624–1633.
- [44] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. 2019. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing* 28, 6 (2019), 2860–2871.
- [45] 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.
- [46] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. 2020. Hierarchical gumbel attention network for text-based person search. In *ACMMM*. 3441–3449.
- [47] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-Path Convolutional Image-Text Embeddings with Instance Loss. *TOMM* 16, 2 (2020), 1–23.
- [48] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13041–13049.
- [49] 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 *ACMMM*. 209–217.