Unified Pre-training with Pseudo Texts for Text-To-Image Person Re-identification

Zhiyin Shao[1,2*], Xinyu Zhang[2*], Changxing Ding[1,3†], Jian Wang[2], Jingdong Wang[2]

[1]South China University of Technology, China [2]Baidu VIS, China [3]Pazhou Lab, China

eezyshao@mail.scut.edu.cn chxding@scut.edu.cn

{zhangxinyu14,wangjian33,wangjingdong}@baidu.com

Abstract

*The pre-training task is indispensable for the text-to-image person re-identification (T2I-ReID) task. However, there are two underlying inconsistencies between these two tasks that may impact the performance: i) **Data** inconsistency. A large domain gap exists between the generic images/texts used in public pre-trained models and the specific person data in the T2I-ReID task. This gap is especially severe for texts, as general textual data are usually unable to describe specific people in fine-grained detail. ii) **Training** inconsistency. The processes of pre-training of images and texts are independent, despite cross-modality learning being critical to T2I-ReID. To address the above issues, we present a new unified pre-training pipeline (UniPT) designed specifically for the T2I-ReID task. We first build a large-scale text-labeled person dataset "LUPerson-T", in which pseudo-textual descriptions of images are automatically generated by the CLIP paradigm using a divide-conquer-combine strategy. Benefiting from this dataset, we then utilize a simple vision-and-language pre-training framework to explicitly align the feature space of the image and text modalities during pre-training. In this way, the pre-training task and the T2I-ReID task are made consistent with each other on both data and training levels. Without the need for any bells and whistles, our UniPT achieves competitive Rank-1 accuracy of, i.e., 68.50%, 60.09%, and 51.85% on CUHK-PEDES, ICFG-PEDES and RSTPReid, respectively. Both the LUPerson-T dataset and code are available at https://github.com/ZhiyinShao-H/UniPT.*

1. Introduction

Text-to-image person re-identification [23] (T2I-ReID) is a retrieval task that aims to search specific person im-
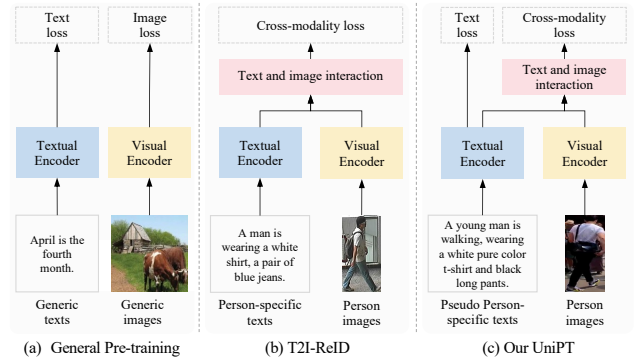
Figure 1. Pipeline comparison on (a) general pre-training in previous works, (b) T2I-ReID tasks, and (c) our unified pre-training (UniPT). Our UniPT shares a similar format with the T2I-ReID pipeline at both the data and training levels.

ages based on natural language descriptions [23]. Compared with large-scale datasets in general image-text retrieval tasks [34, 24, 51], existing T2I-ReID datasets suffer from limited scale and diversity. Therefore, the prior knowledge learned through large-scale pre-training is critical if good performance on the T2I-ReID task is to be achieved.

Previous works [60, 56, 44, 25, 30] simply utilize publicly released pre-trained models that have learned from generic data. In more detail, the visual encoders [14, 6] are usually pre-trained on ImageNet [33], while the textual encoders [16, 4] learn on hundreds of millions of words [65, 4]. These models naturally serve as the initial parameters of visual and textual encoders in T2I-ReID. The existing literature [48, 35, 8, 57, 42, 5] thus pays more attention to designing various cross-modality modules to mitigate the modality gap [48, 35, 8] and the part-alignment problems [57, 42, 5] instead. No extant research has yet investigated how the pre-training task affects the following T2I-ReID task.

In this paper, we make the first attempt to explicitly re-

veal the inconsistencies between the existing pre-training task and the T2I-ReID task. As shown in Figure 1, there are two main inconsistencies: *i.e.*, data and training. Regarding the *data inconsistency*, we observe that the pre-training data are obtained from various objects and scenes, while the T2I-ReID data are specific to pedestrians: a large domain gap exists between these two kinds of datasets. A similar observation can be made with respect to image-based person ReID [9, 28, 10]. Accordingly, [9] proposes a large-scale person dataset, *i.e.*, "LUPerson". Despite the possibility of initializing the visual encoder with LUPerson pre-trained models, this approach is not optimal due to the remaining textual gap. In fact, the generic text data fail to describe the person in the kind of fine-grained detail that is crucial to capturing an individual person's characteristics. This highlights the requirement of *a large-scale image-text person dataset for the pre-training*. Moreover, regarding *training inconsistency*, it is worth noting that the initial visual and textual models in previous methods are pre-trained independently. This approach might not be suitable for T2I-ReID, since the interaction between images and texts key to narrowing down the modality gap. We therefore consider *how to directly apply the modality interaction during the pre-training to match the process in T2I-ReID*.

To this end, we present a novel unified pre-training pipeline, **UniPT**, which is specifically designed for the T2I-ReID task. Enabled by LUPerson [9], we create a new text-labeled variant, *i.e.*, LUPerson-**T**, for the model pre-training. Removing the labor and cost associated with manual annotation, the pseudo-text descriptions are derived automatically using a novel divide-conquer-combine strategy based on CLIP [32]. Specifically, we first divide the person characteristics into attribute phrases, *e.g.*, "a blue striped shirt" and "long hair" (*divide* stage). Next, we convert these phrases to prompts, and input them into the frozen CLIP model to obtain the most matched attributes for specific person images (*conquer* stage). Subsequently, we insert the matched attributes into pre-defined templates to obtain complete pseudo-texts. Moreover, we adopt synonym substitution to enrich the diversity of descriptions (*combine* stage). Despite the noise contained in the pseudo-texts, LUPerson-T focuses describing individual persons in detail, effectively reducing the data domain gap between the pre-training and the T2I-ReID tasks.

Based on the LUPerson-T dataset, we simply use the vision-and-language pre-training framework in the pre-training process. Inspired by CLIP [32], we employ the contrastive loss on pairs of person images and their pseudo-texts. Meanwhile, we apply the masked language model objective [4] to prevent the underlying over-fitting. In this way, the features of images and texts can be made to align with each other during the pre-training, ensuring that the training processes of the pre-training and T2I-ReID tasks

Task	Dataset	Person Only	Image Size	Text	
				Size	Source
Pre-training	ImageNet [33]	×	1.3M	-	-
	Text data in [4]	×	-	3.3B Words	Collected
	LUPerson [9]	✓	4M	-	-
	LUPerson-**T** (Ours)	✓	1.3M	1.3M	Generated
T2I-ReID	CUHK-PEDES [23]	✓	40,206	80,412	Labeled
	ICFG-PEDES [5]	✓	54,522	54,522	Labeled
	RESPReid [63]	✓	20,505	40,110	Labeled

Table 1. The statistics comparison on the pre-training datasets and the T2I-ReID datasets.

are consistent. In summary, our key contributions are as follows:

- We first reveal the data and training inconsistencies between the pre-training and T2I-ReID tasks.
- We build a text-labeled dataset LUPerson-T, in which the pseudo text descriptions are automatically generated by the proposed divide-conquer-combine strategy.
- We propose a unified pre-training pipeline, namely UniPT, on LUPerson-T to make the pre-training and the T2I-ReID task consistent at both the data and training levels.
- We conduct comprehensive experiments and analyses to show the effectiveness of our UniPT. Our method outperforms current state-of-the-art methods on three benchmarks.

2. Related Work

2.1. Vision-Language Pre-training Models

The Transformer has attracted a lot of attention in the fields of computer vision and natural language processing. Many vision-language pre-training tasks apply the Transformer as their base structure [3, 20, 21, 27, 37, 38, 40, 62]. Depending on their model structure, the existing vision-language pre-trainin (VLPT) methods can be categorized into two-stream and single-stream models. Several two-stream models [37, 38, 40] fuse two modalities after extracting visual features and textual features respectively. Others, using an approach similar to CLIP, extract the features of two modalities completely separately and only employ contrastive loss at the end of the process. For their part, the single-stream models [3, 20, 21, 27, 62] process the image feature and the language feature together as a joint distribution. Approaches of this kind require the paired text and image to be fed into the network together. Recently, several works [50, 43] have attempted to explore the use of a shared backbone for both modalities. [50] explores a Modality-Shared CLIP architecture, in which the visual encoder and textual encoder share the parameters. Besides, [43] explores unified models for language, vision, and multimodal pre-training.
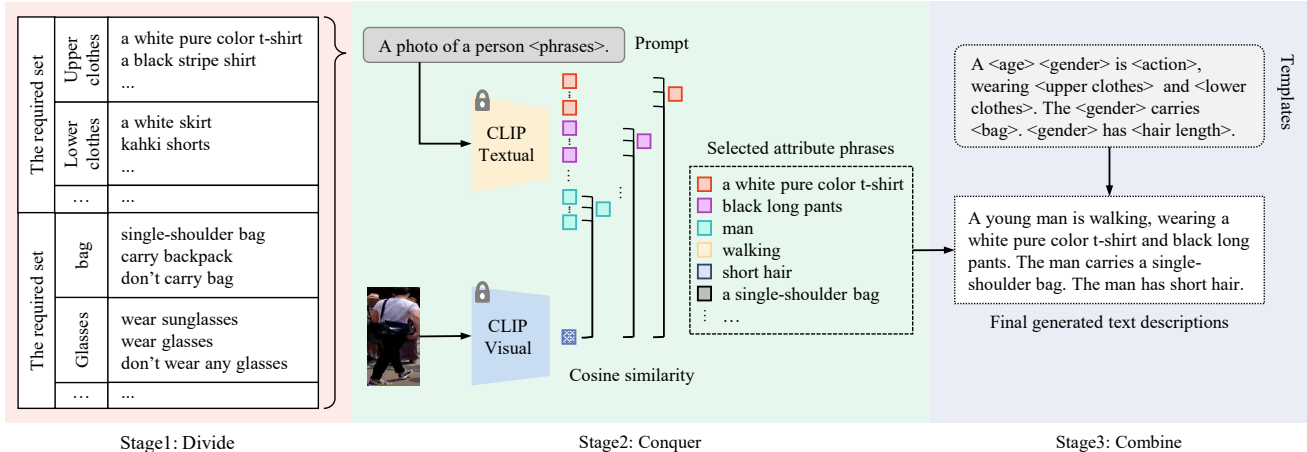
Figure 2. The overview of the pseudo text generation in LUPerson-T via the proposed divide-conquer-combine strategy. In stage1, we *divide* the person characters into attribute phrases. In stage2, we assemble these attribute phrases into prompts, and send them to the frozen CLIP model. According to the cosine similarity between text and image embeddings, we choose the most matched attribute phrases. In stage3, we *combine* the matched attributes into the pre-defined templates to obtain complete sentences.

2.2. Text-to-Image Person ReID

Recently, text-to-image ReID has begun to attract more research attention, despite being more challenging than the general image-based ReID tasks [58, 39, 52, 61, 53, 59, 54, 11, 10, 15, 28, 49, 64] and the general cross-modal retrieval tasks. Existing works of this kind can be divided into global-embedding methods and local-embedding methods. Methods in the former category [60, 56, 44] directly create one dual-stream network to extract the features of different modalities, which align the global feature vectors to a common space. Methods in the latter category [23, 22, 2, 17, 29, 42] design various kinds of model structures or objective functions to obtain and align the local features. For example, MGEL [42] used a coarse-to-fine approach to learn local visual features on different spatial scales for different words or phrases. These local-embedding methods are usually based on a two-stream backbone, to which several local alignment modules are added. IVT [36] utilized a single network to learn representation for both modalities. The network shares the same multi-head attention module but allots different FFNs to different modalities. However, it faces one significant problem: pre-training. To address this issue, in this paper, we automatically generate natural language description tags for LUPerson's pedestrian images and utilize this pseudo-labeled dataset for our pre-training.

3. LUPerson-T: LUPerson with Pseudo-Texts

Existing T2I-ReID datasets cover very limited data scales. As shown in Table 1, the largest T2I-ReID dataset, ICFG-PEDES [5], only contains fewer than 0.06M image-text pairs. Therefore, current T2I-ReID methods [60, 56, 44, 25, 30] directly use models pre-trained on large-scale databases like ImageNet [33] or generic texts [4]. However, these approaches overlook the fact that a large domain gap exists in these contexts, since the pre-training data are inconsistent with the T2I-ReID data. In detail, generic images consist of various objects and scenes that are not images of specific individual people, while T2I-ReID datasets only contain pedestrians. Reviewing image-based ReID [11, 28, 10], pre-training on the large-scale unlabeled person dataset LUPerson [9] has demonstrated the effectiveness of pre-training on person images. It is practicable to replace the initial visual encoder with a model pre-trained on LUPerson; however, the textual inconsistency still exists under these circumstances. Moreover, the generic textual descriptions are usually coarse, while those in T2I-ReID datasets are fine-grained to describe each person's characteristics. We therefore naturally pose a question: *can we improve the pre-training feature representations with a large-scale text-image person dataset for T2I-ReID?*

In this paper, we create a text-labeled variant of LUPerson [9], named LUPerson-**T**. LUPerson-T consists of 1.3M person images (a subset taken from [28]) along with their corresponding textual descriptions. Dispensing with the need for manual annotations, we propose a *divide-conquer-combine* strategy to generate pseudo-texts for person images. In the following, we will carefully introduce this strategy. To the best of our knowledge, LUPerson-T is the first visual-and-language pre-training dataset aimed at pedestrians. We make LUPerson-T public and hope that this dataset can help to progress research in the area of person ReID.

3.1. Divide-Conquer-Combine Strategy: Pseudo Text Generation

Since the manual annotation of person descriptions is time-consuming, we opt to automatically generate pseudo-
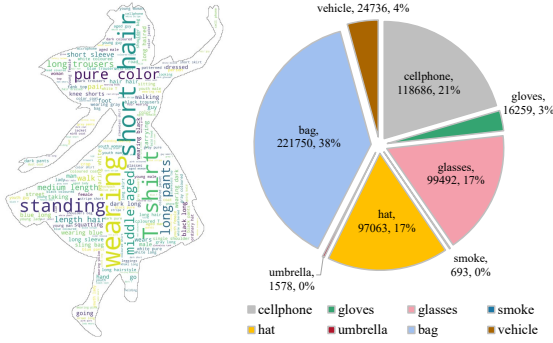
Figure 3. The WordCloud of high-frequency words (left) and the frequency of attributes in the optional set (right).

texts. Inspired by CLIP [32], a recent language-image pre-training approach that performs well on zero-shot tasks, we make an effort to obtain matched textual descriptions for a specific person image. However, it is challenging for CLIP to directly generate complete and fine-grained person descriptions. We thus propose a *divide-conquer-combine* strategy to split a difficult problem into sub-problems, then integrate the corresponding sub-solutions into a complete one. The overview is presented in Figure 2.

Divide. First of all, we examined most person images in the dataset and extract 14 kinds of attributes based on their frequency of occurrence. We intuitively group these attributes into two sets, *i.e.*, the required set and the optional set. The required set includes 6 attributes, namely <age>, <gender>, <upper clothes>, <lower clothes>, <action> and <hair length>, meaning that every person must possess all of these attributes[1]. The optional set contains attributes that a person is likely to have, including <bag>, <glasses>, <smoke>, <hat>, <cellphone>, <umbrella>, <gloves> and <vehicle>, for a total of eight attributes. These 14 attributes cover most basic aspects of pedestrian appearances and already facilitate good pre-training[2]. Each attribute additionally has its own phrases based on its classes. For example, <upper clothes> contains "a blue striped shirt", "a white pure color t-shirt", "a blue plaid wind coat" and so on. Please refer to the supplementary material for more details. These phrases are initial version in this stage that are artificially designed by ourselves, which will be then extended by synonym substitution in the combine stage.

Conquer. This is the most important step of the text generation. To make better use of the CLIP model, we convert attribute phrases into the standard prompt format, *i.e.*,

"A photo of a person <phrases>"[3]. We then put the person images into the frozen CLIP visual model to obtain visual representations, and the prompts into the frozen CLIP textual model to get textual features. Here, we use the ViT-B/16 model [1]. Subsequently, for each person image feature, we calculate cosine similarities between itself and all prompt features. Then, for each of the required attributes, we choose the most similar phrase. For each of the optional attributes, we choose the satisfied phrase with a softmax probability larger than 0.9. In this way, we can select the most relevant attribute phrases for person images.

Combine. Next, the selected attribute phrases should be combined together to produce complete sentences. In detail, we pre-defined 456 templates with "blank" positions for phrases, *e.g.*, "The <age> <gender> is <action>, wearing <upper clothes> and <lower clothes>. The <gender> wears <glasses> and carries <bag>. <gender> has <hair length>." Different templates have different patterns and diverse attribute blanks. For a person image, we randomly choose one template, then automatically fill its candidate phrases from the conquer stage into the corresponding blanks. It is common knowledge that the richness of vocabulary has a large impact on the quality of a dataset; we thus also use synonym substitution to enrich the diversity of the textual descriptions.

The proposed divide-conquer-combine strategy is automatic. Although we provide pre-defined templates, this effort is trivial compared to manual annotation. Based on this strategy, we can easily obtain a large-scale person-specific dataset with image and text pairs.

3.2. Properties of LUPerson-T

Table 1 provides the statistics of LUPerson-T and existing T2I-ReID datasets up to now. LUPerson-T contains about 1.3M image–text person pairs, making it the largest T2I-ReID dataset in existence, even though the text descriptions are pseudo-texts. The optional set appears 580,257 times in total, in which the occurrence frequency of four attributes accounts for 94% of all optional attributes (namely <bag>, <cellphone>, <glasses> and <hat>). Among them, the <bag> attribute appears most frequently (about a 38% chance).

4. Unified Pre-Training for T2I-ReID Task

Based on LUPerson-T, we can apply vision-and-language pre-training frameworks. Unlike solely visual or solely textual pre-training, these frameworks can perform interactive learning between images and texts. In other words, the pre-training shares a unified behavior as T2I-ReID tasks. For simplicity, we refer to this unified pre-

---

[1]To be more precise, each person has a high probability of exhibiting each of the required attributes, since there may exist blur, occlusion, truncation, and so on in person images. For simplicity, we do not consider these problems in this paper.

[2]We conjecture that introducing more attributes is potential to bring further improvement that we leave for the future work.

[3]There are minor modifications based on different phrases, *e.g.*, "A photo of a <gender>".
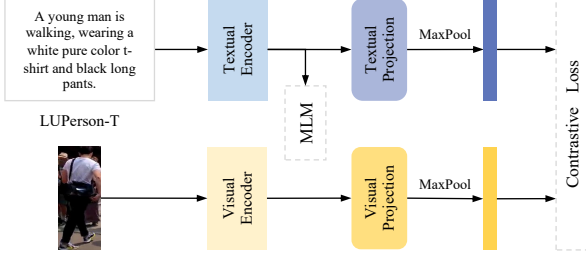
Figure 4. Overview of our UniPT pipeline.

training pipeline on LUPerson-T as **UniPT**.

4.1. UniPT: Unified Pre-training on LUPerson-T

We follow CLIP [32] as our basic framework. We feed images and texts into the image encoder and text encoder to obtain visual and textual features, respectively. These two modality features are forced into a multi-modal embedding space by jointly training the visual and textual encoders. The objective function is the contrastive loss [55, 31, 32] on visual and textual feature representations. Besides, we also introduce the mask language model (MLM) pre-training objective in [4] to prevent the underlying over-fitting.

Contrastive loss on image-text pairs. Given a mini-batch of N image-text pairs, we aim to predict which of the $N \times N$ possible image-text pairings across a batch are truly matched (or not). The cosine similarity matrix of $N \times N$ image-text pairings is calculated on visual embeddings $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_N\}$ and textual embeddings $\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_N\}$. The contrastive loss acts to maximize the cosine similarity of the real matched pairs of image and text embeddings while minimizing that of the N^2 - N incorrect pairings. We apply the symmetric contrastive loss function used in [32], which is formulated as follows:

$$
\begin{aligned}
L_{\text{I2T}} &= \frac{1}{N} \sum_{n=1}^{N} -\log\left(\frac{\exp(\text{sim}(\boldsymbol{v}_n, \boldsymbol{t}_+))}{\sum_{i=1}^{N} \exp(\text{sim}(\boldsymbol{v}_n, \boldsymbol{t}_i))}\right), \\
L_{\text{T2I}} &= \frac{1}{N} \sum_{n=1}^{N} -\log\left(\frac{\exp(\text{sim}(\boldsymbol{t}_n, \boldsymbol{v}_+))}{\sum_{i=1}^{N} \exp(\text{sim}(\boldsymbol{t}_n, \boldsymbol{v}_i))}\right), \\
L_{\text{con}} &= (L_{\text{I2T}} + L_{\text{T2I}})/2,
\end{aligned}
\tag{1}
$$

where $\boldsymbol{v}_i/\boldsymbol{t}_i$ is the i-th image/text representation of a batch. $\boldsymbol{t}_+$ is the corresponding textual description of $\boldsymbol{v}_n$. Similarly, $\boldsymbol{v}_+$ and $\boldsymbol{t}_n$ are a matched image-text pair. $\text{sim}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^{\text{T}}\boldsymbol{b}/ ||\boldsymbol{a}|| \, ||\boldsymbol{b}||$ denotes the cosine similarity between the two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

Mask language model objective on texts. Recall that CLIP contains 400M image–text pairs, representing adequate textual diversity. In contrast, the text descriptions in our LUPerson-T are specific for each person, constructed by a relatively small scale of attribute phrases. Therefore, using only the contrastive loss may result in over-fitting. Ac-

cordingly, in order to learn more robust representations, we also employ the mask language model (MLM) [4] objective in our pre-training. Specifically, we randomly mask 15% of the tokens in a sentence, following [4]. The objective of MLM is to predict the masked words based on their context. Let x denote a person's textual description, while $\mathcal{M}$ is the set of the randomly masked positions of x. The formula of the MLM loss is defined as follows:

$$
L_{\text{mlm}} = - \sum_{i \in \mathcal{M}} \log p(x_i | x_{\backslash \mathcal{M}}),
\tag{2}
$$

where $x_{\backslash \mathcal{M}}$ is the masked version of x. The target of the masked word is its corresponding vocabulary id [4]. With MLM loss, the optimization of the pre-training task is made more difficult to prevent the over-fitting problem.

Total optimization. We use the contrastive loss and the MLM loss together to optimize the overall UniPT framework. The final loss function is as follows:

$$
L_{\text{pre}} = L_{\text{con}} + \beta L_{\text{mlm}},
\tag{3}
$$

where $\beta = \{0, 1\}$ controls whether or not the MLM loss is used. When $\beta = 0$, the optimization is reduced to using only contrastive loss. By default, we set $\beta = 1$ so that L_{con} and L_{mlm} are used together.

Model architecture. The model architecture is illustrated in Figure 4. We attach a 1×1 convolution layer to the visual and textual encoder respectively to project the image and text features into the same dimension. Subsequently, we conduct a max-pooling operation on both the visual and textual features to obtain the global embeddings of $\boldsymbol{V}$ and $\boldsymbol{T}$ for model training.

4.2. Supervised Learning on T2I-ReID datasets

Based on our UniPT models, we next conduct supervised fine-tuning on the downstream T2I-ReID datasets. We use the identification loss [39, 56] and ranking loss [7] on global and granularity-unified features [35] for optimization.

Identification loss. Inspired by [39, 56], we adopt cross-entropy loss as the identification (ID) loss. For a specific visual feature $\boldsymbol{v}$ and textual feature $\boldsymbol{t}$, we denote the predicted identity probability as $\hat{y}_v$ and $\hat{y}_t$, respectively. The identification loss is formulated as follows:

$$
L_{\text{id}}(\boldsymbol{v}, \boldsymbol{t}) = -(y \log \hat{y}_v + y \log \hat{y}_t),
\tag{4}
$$

where y is the ground-truth person ID label for the $\boldsymbol{v}$ and $\boldsymbol{t}$ pair. Following [5], we use a shared classifier to get image and text probabilities.

Ranking loss. The ranking loss [7] is commonly used to align visual and textual modalities. It is based on the similarity of the text-image triplets, which is formulated as follows:

$$
\begin{aligned}
L_{\text{rk}}(\boldsymbol{v}, \boldsymbol{t}) = &\max(\alpha - \text{sim}(\boldsymbol{v}, \boldsymbol{t}_+) + \text{sim}(\boldsymbol{v}, \boldsymbol{t}_-), 0) \\
&+ \max(\alpha - \text{sim}(\boldsymbol{t}, \boldsymbol{v}_+) + \text{sim}(\boldsymbol{t}, \boldsymbol{v}_-), 0),
\end{aligned}
\tag{5}
$$

Pre-training data		Backbone	CUHK-PEDES		ICFG-PEDES	
Image	Text		Rank-1	Rank-5	Rank-1	Rank-5
IN [33]	Text [4]	DeiT-small [41]	65.30	83.19	57.60	74.53
		ViT-B/16 [6]	66.59	84.40	58.62	75.31
LUP-T	LUP-T	DeiT-small [41]	66.83	84.16	59.08	75.92
		ViT-B/16 [6]	**68.50**	**84.67**	**60.09**	**76.19**

Table 2. LUPerson-T boosts the T2I-ReID performance. Previous works [60, 56, 44, 25, 30] use the pre-trained models on ImageNet [33] and generic texts [4] as the visual and textual models (in the top block). Differently, our UniPT utilizes the pre-trained models on the proposed LUPerson-T (in the bottom block). The abbreviations {"IN", 'Text', 'LUP-T'} represent ImageNet [33], generic texts in [4] and our LUPerson-T respectively.

Type	Pre-training models		CUHK-PEDES		ICFG-PEDES	
	Visual	Textual	Rank-1	Rank-5	Rank-1	Rank-5
Single	SSL [28]	BERT [4]	60.14	77.33	47.25	66.58
Single	Ours	BERT [4]	65.01	82.60	58.15	75.05
UniPT	Ours	Ours	**66.83**	**84.16**	**59.08**	**75.92**

Table 3. Our unified pre-training pipeline (UniPT) improves the feature representations of both visual and textual modalities. "Single" means that the visual and textual models are pre-trained independently. For a fair comparison, all experiments use small-size ViT as the visual backbone, and BERT as the textual backbone.

where t_+/v_+ and t_-/v_- denote the positive sample and the semi-hard negative sample [13] of v/t respectively. α is the hyper-parameter of the margin.

Granularity-unified loss. In the T2I-ReID task, the granularity alignment is key to reducing the modality gap between text and image [35]. Therefore, we also employ the prototype-based granularity unification (PGU) module presented in [35] to extract the granularity-unified features $\widetilde{v}$ and $\widetilde{t}$. Following [35], we apply the identification loss and the ranking loss on $\widetilde{v}$ and $\widetilde{t}$, which is formulated as follows:

$$L_{\mathrm{pgu}} = L_{\mathrm{id}}(\widetilde{v},\widetilde{t}) + L_{\mathrm{rk}}(\widetilde{v},\widetilde{t}). \tag{6}$$

Total optimization. The overall loss function in the T2I-ReID supervised fine-tuning stage is as follows:

$$L_{\mathrm{ft}} = L_{\mathrm{id}} + L_{\mathrm{rk}} + \gamma L_{\mathrm{pgu}} \tag{7}$$

where $\gamma = \{0, 1\}$ controls whether or not the granularity-unified loss is used. When $\gamma = 0$, we only employ L_{id} and L_{rk} on v and t: when $\gamma = 1$, the additional L_{pgu} is added for the supervised learning (our default setting).

5. Experiments

5.1. Datasets and Evaluation Metrics

Datasets. We evaluate three T2I-ReID datasets, *i.e.*, CUHK-PEDES [23], ICFG-PEDES [5], and RST-PReid [63]. The statistics of these three datasets are listed in Table 1.
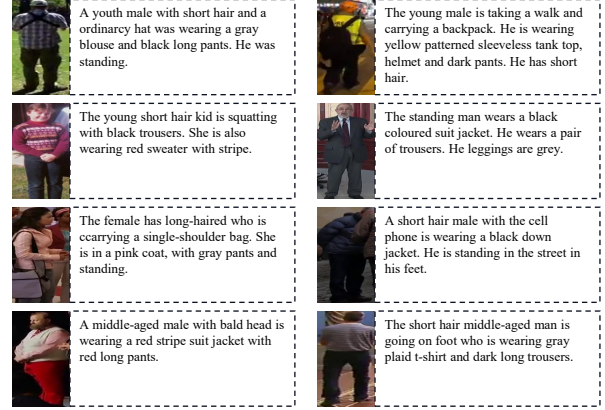

Figure 5. Visualization of some examples in LUPerson-T. These sentences can fairly accurately describe the appearance of the human body in a fluent sentence pattern.

Method	Backbone	CUHK-PEDES		ICFG-PEDES	
		Rank-1	Rank-5	Rank-1	Rank-5
CLIP [32]	ResNet50	61.25	81.34	52.34	71.19
	ViT-B/16	66.34	84.18	59.01	75.96
UniPT	DeiT-small	66.83	84.16	59.08	75.92
	ViT-B/16	**68.50**	**84.67**	**60.09**	**76.19**

Table 4. Our UniPT *vs*. CLIP [32]. CLIP is also a vision-and-language pre-training framework. Our UniPT pre-train on LUPerson-T, and CLIP uses the generic image-text pairs.

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 a query, at least one matching person image is found in the top-k candidate list.

5.2. Implementation Details

Model structure. In our experiments, we opt to use DeiT-Small [41] and ViT-B/16 [6] as the visual backbones and BERT [4] as the textual backbone, respectively.

Pre-training. The input image resolution is set to 384×128 and the text token length is 100. For all experiments, we use the AdamW [26] optimizer with a base learning rate of 1e-5. The learning rate is warmed up for 10% of the total steps and the batch size is set to 512. The total number of epochs is 15. The model is pre-trained on 8 Nvidia Tesla V100 GPUs.

Supervised learning. All person images are resized to 384×128 and augmented with random horizontal flipping. The margin α of the ranking loss is set to 0.2 in both L_{rk} and L_{pgu}. The batch size is 64, and the optimizer used is Adam [18]. Except for the PGU module, with a learning rate of 1e-4, the initial learning rate is set to 1e-5 for the pre-trained textual encoder and 1e-4 for the visual encoder. The training is conducted on a single Nvidia Tesla V100 GPU.

Method	CUHK-PEDES		ICFG-PEDES	
	Rank-1	Rank-5	Rank-1	Rank-5
UniPT w/o MLM	65.77	83.53	57.35	74.14
UniPT w MLM (UniPT)	**66.83**	**84.16**	**59.08**	**75.92**

Table 5. Effectiveness of the masked language modeling (MLM) objective in our UniPT. The visual backbone is DeiT-small [41].

Pre-training Type	Method	CUHK-PEDES		ICFG-PEDES	
		Rank-1	Rank-5	Rank-1	Rank-5
Single	w/o L_{pgu}	60.56	79.09	51.66	69.06
	w/ L_{pgu}	65.30	83.19	57.30	74.18
UniPT	w/o L_{pgu}	61.99	79.86	54.41	71.10
	w/ L_{pgu}	**66.83**	**84.16**	**59.08**	**75.92**

Table 6. Effectiveness of the granularity-unified loss L_{pgu} in the T2I-ReID supervised learning. "Single" means that the visual and textual models are pre-trained independently. DeiT-small [41] is the backbone.

5.3. Effectiveness of UniPT on LUPerson-T

LUPerson-T boosts the T2I-ReID performance. We compare our LUPerson-T with other generic datasets [33, 4] in Table 2. Our model, pre-trained on LUPerson-T, achieves +1.53% and +1.48% improvement in terms of Rank-1 accuracy on CUHK-PEDES and ICFG-PEDES with a DeiT-small backbone. Furthermore, our pre-trained model also outperforms the generic pre-trained model on ViT-B/16 by +1.91% and +1.47%. The underlying reason is that the previous generic pre-training process, under which visual and textual models are trained separately, can lead to the misalignment of the visual and textual feature representations, which harms the performance of the downstream T2I-ReID tasks. Our UniPT, based on LUPerson-T and containing pairs of images and pseudo-texts, helps to alleviate this issue to some extent.

Figure 5 presents the visualization of some examples in LUPerson-T. The pseudo-texts that we generate contain basic information about each person's characteristics. The pre-defined templates and the use of synonym replacement guarantee the fluidity and diversity of the generated pseudo-texts.

UniPT improves the visual and textual representations. Table 3 shows the effectiveness of our unified vision-and-language pre-training pipeline (UniPT). In the first row in Table 3, we equip the image-based pre-trained model [28] on LUPerson [9] as the visual encoder. Compared with the top two rows, our visual pre-trained model from UniPT is largely superior than that from [28] by +4.87% and +10.90% on CUHK-PEDES and ICFG-PEDES. These results show that our UniPT can boost the visual representation. Meanwhile, our UniPT can also improve the textual feature representation. From the bottom two rows in Table 3, we can see that the textual encoder from UniPT
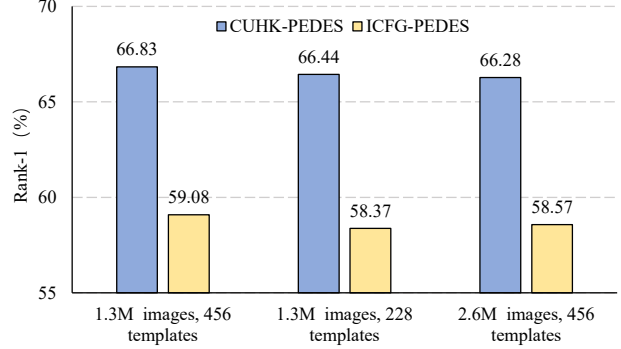

Figure 6. Comparison on different data scales of LUPerson-T.

achieves +1.82% and +0.93% improvement compared with that from BERT [4] on CUHK-PEDES and ICFG-PEDES. Therefore, our UniPT pipeline has great merit for both images and texts.

Our UniPT *vs*. CLIP [32]. CLIP [32] is also a vision-and-language pre-training framework, which is trained on a significant number of generic image-text pairs (400M private data pairs). As shown in Table 4, directly using the pre-trained CLIP visual and textual models as the initial visual and textual encoders can achieve comparable results. It is because that the generic image-text pre-training is able to gain benefit from the training consistency with T2I-ReID, *i.e.*, both tasks align the visual and textual features. However, the data in CLIP still suffer due to the existence of a domain gap between the pre-training and the T2I-ReID task. Compared on the ViT-B/16 backbone, our UniPT achieves 68.50% Rank-1 accuracy while CLIP achieves only 66.34% on CUHK-PEDES. It is noted that even though the scale of data in UniPT is significantly smaller than that of the data in CLIP, the image–text pairs in the former are specific for persons. Therefore, the pre-trained models in our UniPT are better than those in CLIP.

5.4. Ablation Study

Effectiveness of MLM loss in the pre-training. Our UniPT adopts not only the contrastive loss but also the mask language modeling (MLM) objective during the pre-training process. We evaluate the effectiveness of MLM loss in Table 5. When MLM is removed (*i.e.*, $\beta = 0$ in Eq. (7)), the performance drops by 1.06% and 1.73% compared with using MLM on CUHK-PEDES and ICFG-PEDES, respectively. The reason may be that the text descriptions are not sufficiently diverse. The MLM makes the training harder by randomly masking some text tokens, which helps to prevent over-fitting.

Effectiveness of the granularity-unified loss in the T2I-ReID supervised learning. The granularity-unified loss L_{pgu} with the prototype-based granularity unification (PGU) module is proposed by [35] to extract granularity-

Methods	Backbone	CUHK-PEDES			ICFG-PEDES			RSTPReid		
		Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
Dual Path [60]	ResNet50	44.40	66.26	75.07	38.99	59.44	68.41	-	-	-
CMPM/C [56]	ResNet50	49.37	-	79.27	43.51	65.44	74.26	-	-	-
A-GANet [25]	ResNet50	53.14	74.03	82.95	-	-	-	-	-	-
TDE [30]	ResNet50	55.25	77.46	84.56	-	-	-	-	-	-
VTA [13]	ResNet50	55.32	77.00	84.26	-	-	-	-	-	-
MIA [29]	ResNet50	53.10	75.00	82.90	46.49	67.14	75.18	-	-	-
SCAN [19]	ResNet50	55.86	75.97	83.69	50.05	69.65	77.21	-	-	-
ViTAA [45]	ResNet50	55.97	75.84	83.52	50.98	68.79	75.78	-	-	-
HGAN [57]	ResNet50	59.00	79.49	86.62	-	-	-	-	-	-
NAFS [12]	ResNet50	59.94	79.86	86.70	-	-	-	-	-	-
DSSL [63]	ResNet50	59.98	80.41	87.56	-	-	-	32.43	55.08	63.19
MGEL [42]	ResNet50	60.27	80.01	86.74	-	-	-	-	-	-
SSAN [5]	ResNet50	61.37	80.15	86.73	54.23	72.63	79.53	43.50	67.80	77.15
LapsCore [48]	ResNet50	63.40	-	87.80	-	-	-	-	-	-
LBUL [47]	ResNet50	64.04	82.66	87.22	-	-	-	45.55	68.20	77.85
CAIBC [46]	ResNet50	64.43	82.87	88.37	-	-	-	47.35	69.55	79.00
LGUR [35]	ResNet50	64.21	81.94	87.93	57.42	74.97	81.45	46.95	69.90	79.12
AXM-Net [8]	-	64.44	80.52	86.77	-	-	-	-	-	-
LGUR [35]	DeiT-small	65.25	83.12	89.00	59.02	75.32	81.56	47.95	71.85	80.25
IVT [36]	ViT-B/16	65.59	83.11	89.21	56.04	73.60	80.22	46.70	70.00	78.80
Our UniPT	DeiT-small	**66.83**	**84.16**	**89.42**	**59.08**	**75.92**	**82.08**	**49.45**	**72.75**	**80.35**
Our UniPT	ViT-B/16	**68.50**	**84.67**	**90.38**	**60.09**	**76.19**	**82.46**	**51.85**	**74.85**	**82.85**

Table 7. Performance comparisons on supervised T2I-ReID tasks on CUHK-PEDES [23], ICFG-PEDES [5] and RSTPReid [63].

	Methods	Rank-1	Rank-5	Rank-10
C→I	Dual Path [60]	15.41	29.80	38.19
	MIA [29]	19.35	36.78	46.42
	SCAN [19]	21.27	39.26	48.83
	SSAN [5]	24.72	43.43	53.01
	SSAN(w/ BERT) [5]	29.24	49.00	58.53
	LGUR [35]	34.25	52.58	60.85
	Our UniPT (DeiT-small)	**35.51**	**54.25**	**62.69**
I→C	Dual Path [60]	7.63	17.14	23.52
	MIA [29]	10.93	23.77	32.39
	SCAN [19]	13.63	28.61	37.05
	SSAN [5]	16.68	33.84	43.00
	SSAN(w/ BERT) [5]	21.07	38.94	48.54
	LGUR [35]	25.44	44.48	54.39
	Our UniPT (DeiT-small)	**28.48**	**48.02**	**57.65**

Table 8. Performance comparison on the cross-domain setting.

unified features. In Eq. (7), when $\gamma = 0$, the model does not use L_{pgu}. We evaluate the effectiveness of L_{pgu} on the previous separate pre-training and our UniPT. Table 6 shows that L_{pgu} can effectively improve the feature representation.

Data scale. We provide the performance on various scales of data in Figure 6. As can be seen from the left two bars, a higher number of templates improves the performance due to the increase in the diversity of text descriptions. In contrast, including more images with the same number of templates (right two bars) produces the opposite effect. There are two key reasons for this. The first is that the templates

and attributes are not sufficiently abundant, meaning that many images might share the same description. The second is that selected attribute phrases are not exactly accurate; thus, more images will result in more noisy text descriptions. We thus use 1.3M image–text pairs with 456 templates by default.

5.5. Comparison with State-of-the-Art Methods

Comparison on supervised T2I-ReID tasks. As illustrated in Table 7, our UniPT outperforms the state-of-the-art methods on CUHK-PEDES [23], ICFG-PEDES [5] and RSTPReid [63]. We first compare the performance of our method on CUHK-PEDES, which achieves 66.83% and 68.50% Rank-1 accuracy based on DeiT-small and ViT-B/16. LGUR [35] proposed two modules, *i.e.*, a dictionary-based granularity alignment module and a prototype-based granularity unification (PGU) module, to learn granularity-unified representations for both modalities. Here, we only equip the PGU module on the supervised learning on downstream tasks to improve the feature representation. Our method outperforms LGUR [35] by +1.58% in terms of Rank-1 accuracy. IVT [36] proposed multi-level alignmen (MLA) and bidirectional mask modeling (BMM), which enable the model to mine finer and more semantic alignments. Our method outperforms IVT by +2.91% in terms of Rank-1 accuracy.

We further evaluate our proposed UniPT on ICFG-PEDES [5] and RSTPReid [63]. On ICFG-PEDES, our

method achieves 59.08% and 60.09% Rank-1 accuracy respectively based on DeiT-small and ViT-B/16. On RST-PReid, in a fair comparison, our method outperforms LGUR [35] by +1.5% with a DeiT-small backbone and surpasses IVT [36] by +5.15% with a ViT-B/16 backbone. In addition, the performance of our method is superior to that of the previous works LBUL [47] and CAIBC [46], by a large margin.

Comparison on the domain generalization. Our model is pre-trained on the proposed LUPerson-T, which contains a large amount of knowledge derived from various images of and texts describing pedestrians. Based on this, we conjecture that the model might have a good generalization ability. To verify this, we conduct domain generalization experiments on two cross-domain settings, following [35]: CUHK-PEDES to ICFG-PEDES (C→I) and ICFG-PEDES to CUHK-PEDES (I→C). As shown in Table 8, our architecture, pre-trained on LUPerson-T, achieves a substantial improvement compared with other methods. For example, our model outperforms SSAN (w/ BERT) [5] by +6.27%, and +7.41% Rank-1 accuracy on C→I and I→C, respectively. Moreover, compared with LGUR [35] which is dedicated to narrowing the granularity gap between the textual and visual feature, we achieve better accuracy overall: +1.26% and +3.04% on C→I and I→C. These experiments accordingly demonstrate the domain generalization ability of our methods.

6. Conclusion

This paper first reveals the data and training inconsistencies between the previous pre-training task and the T2I-ReID task. To solve this problem, we propose a unified pre-training paradigm named UniPT for the T2I-ReID task. Specifically, we build a text-labeled person dataset named LUPerson-T, in which pseudo-textual descriptions are automatically generated. Moreover, we apply the vision-and-language pre-training paradigm for the model pre-training. Our UniPT shares similar format with the T2I-ReID task from both data and training scheme perspectives. Extensive experiments on three benchmarks verify the effectiveness of our proposed UniPT.

References

[1] https://github.com/openai/clip.

[2] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *WACV*, pages 1879–1887. IEEE, 2018.

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[8] Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *AAAI*, volume 36, pages 4477–4485, 2022.

[9] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *CVPR*, pages 14750–14759, 2021.

[10] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. Large-scale pre-training for person re-identification with noisy labels. In *CVPR*, pages 2476–2486, 2022.

[11] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, volume 33, pages 8295–8302, 2019.

[12] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*, 2021.

[13] Jing Ge, Guangyu Gao, and Zhen Liu. Visual-textual association with hardest and semi-hard negative pairs mining for person search. *arXiv preprint arXiv:1912.03083*, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[15] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. In *AAAI*, 2020.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[20] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, volume 34, pages 11336–11344, 2020.

[21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[22] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, pages 1890–1899, 2017.

[23] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, pages 1970–1979, 2017.

[24] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10921–10930, 2020.

[25] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. Deep adversarial graph attention convolution network for text-based person search. In *ACMMM*, pages 665–673, 2019.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[28] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021.

[29] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *TIP*, 29:5542–5556, 2020.

[30] Kai Niu, Yan Huang, and Liang Wang. Textual dependency embedding for person search by language. In *ACMMM*, pages 4032–4040, 2020.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[34] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, pages 5814–5824, 2019.

[35] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *ACMMM*, pages 5566–5574, 2022.

[36] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *ECCVW*, pages 624–641. Springer, 2023.

[37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[38] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019.

[39] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

[40] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021.

[42] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. Text-based person search via multi-granularity embedding learning. In *IJCAI*, pages 1068–1074, 2021.

[43] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[44] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. Language person search with mutually connected classification loss. In *ICASSP*, pages 2057–2061. IEEE, 2019.

[45] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *ECCV*, pages 402–420. Springer, 2020.

[46] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *ACMMM*, pages 5314–5322, 2022.

[47] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *ACMMM*, pages 1984–1992, 2022.

[48] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: Language-guided person search via color reasoning. In *ICCV*, pages 1624–1633, 2021.

[49] Zizheng Yang, Xin Jin, Kecheng Zheng, and Feng Zhao. Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification. In *CVPR*, pages 14298–14307, 2022.

[50] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *ECCV*, pages 69–87. Springer, 2022.

[51] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *CVPR*, pages 15661–15670, 2022.

[52] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV*, pages 8222–8231, 2019.

[53] Xinyu Zhang, Dong Gong, Jiewei Cao, and Chunhua Shen. Memorizing comprehensively to learn adaptively: Unsupervised cross-domain person re-id with multi-level memory. *arXiv preprint arXiv:2001.04123*, 2020.

[54] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *CVPR*, pages 7369–7378, 2022.

[55] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *MLHC*, pages 2–25. PMLR, 2022.

[56] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, pages 686–701, 2018.

[57] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. Hierarchical gumbel attention network for text-based person search. In *ACM MM*, pages 3441–3449, 2020.

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[59] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019.

[60] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *TOMM*, 16(2):1–23, 2020.

[61] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *TPAMI*, 2020.

[62] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, volume 34, pages 13041–13049, 2020.

[63] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *ACMMM*, pages 209–217, 2021.

[64] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *ECCV*, pages 198–214, 2022.

[65] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pages 19–27, 2015.