Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval

Ding Jiang[1], Mang Ye[1,2*]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
School of Computer Science, Wuhan University, Wuhan, China
[2] Hubei Luojia Laboratory, Wuhan, China

https://github.com/anosorae/IRRA

Abstract

Text-to-image person retrieval aims to identify the target person based on a given textual description query. The primary challenge is to learn the mapping of visual and textual modalities into a common latent space. Prior works have attempted to address this challenge by leveraging separately pre-trained unimodal models to extract visual and textual features. However, these approaches lack the necessary underlying alignment capabilities required to match multimodal data effectively. Besides, these works use prior information to explore explicit part alignments, which may lead to the distortion of intra-modality information. To alleviate these issues, we present IRRA: a cross-modal Implicit Relation Reasoning and Aligning framework that learns relations between local visual-textual tokens and enhances global image-text matching without requiring additional prior supervision. Specifically, we first design an Implicit Relation Reasoning module in a masked language modeling paradigm. This achieves cross-modal interaction by integrating the visual cues into the textual tokens with a cross-modal multimodal interaction encoder. Secondly, to globally align the visual and textual embeddings, Similarity Distribution Matching is proposed to minimize the KL divergence between image-text similarity distributions and the normalized label matching distributions. The proposed method achieves new state-of-the-art results on all three public datasets, with a notable margin of about 3%-9% for Rank-1 accuracy compared to prior methods.

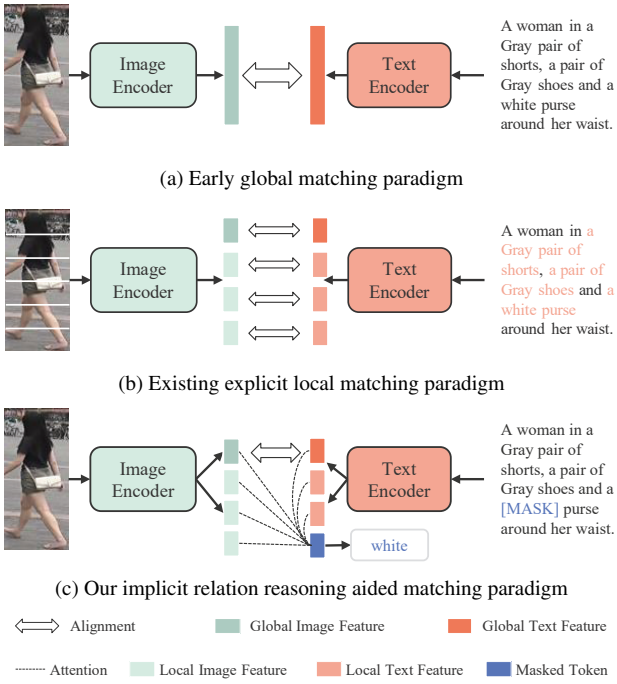


Figure 1. Evolution of text-to-image person retrieval paradigms. (a) Early global-matching method directly align global image and text embeddings. (b) Recent local-matching method, explicitly extract and align local image and text embeddings. (c) Our implicit relation reasoning method, implicitly reasoning the relation among all local tokens to better align global image and text embeddings.

1. Introduction

Text-to-image person retrieval aims to retrieve a person-of-interest from a large image gallery that best matches the text description query [30], which is a sub-task of both image-text retrieval [26, 33, 42] and image-based person re-identification (Re-ID) [15, 32, 45]. Textual descriptions provide a natural and relatively comprehensive way to describe a person's attributes, and are more easily accessible than images. Text-to-image person retrieval thus received increas-

*Corresponding Author: Mang Ye (yemang@whu.edu.cn)

ing attention in recent years, benefiting a variety of applications from personal photo album search to public security.

However, text-to-image person retrieval remains a challenging task due to significant intra-identity variations and modality heterogeneity between vision and language. The former challenge stems from the fact that visual appearances of an identity differ based on pose, viewpoint, illumination, and other factors, while textual description varies by arbitrary descriptive order and textual ambiguity. The latter challenge is the primary issue in cross-modal tasks and is caused by inherent representation discrepancies between vision and language. To tackle above two challenges, the core research problem in text-to-image person retrieval is to explore better ways to extract discriminative feature representations and to design better cross-modal matching methods to align images and texts into a joint embedding space. Early *global-matching* methods [53, 54] aligned images and texts into a joint embedding space by designing cross-modal matching loss functions (Fig. 1 (a)). Typically, these approaches learned cross-modal alignments by using matching losses only at the end of the network, failing to achieve sufficient modality interaction in middle-level layers, which are crucial to bridge the feature-level modality gap. Therefore, some later methods [5, 7, 21, 46] introduced the practice of *local-matching* by building the correspondence between the body parts and the textual entities (Fig. 1 (b)). Although this local matching strategy benefits retrieval performance, it introduces unavoidable noise and uncertainty in the retrieval process. Besides, the strategy requires extracting and storing multiple local part representations of images and texts, computing pairwise similarity between all those representations during inference. These resource-demanding properties limit their applicability for practical large-scale scenarios.

In this paper, we present IRRA: a *cross-modal Implicit Relation Reasoning and Aligning* framework, which performs global alignment with the aid of cross-modal implicit local relation learning. Unlike previous methods that heavily rely on explicit fine-grained local alignment, our approach implicitly utilizes fine-grained information to enhance global alignment without requiring any additional supervision and inference costs (Fig. 1 (c)). Specifically, we design an Implicit Relation Reasoning module that effectively builds relations between visual and textual representations through self- and cross-attention mechanisms. This fused representation is then utilized to perform masked language modeling (MLM) task to achieve effective implicit inter-modal and intra-modal fine-grained relation learning. MLM is generally utilized during the pre-training stage of vision-language pre-training (VLP) [6, 9, 27, 31, 41]. In this work, we make the first attempt to demonstrate the effectiveness of MLM in downstream fine-tuning tasks. Our main innovation is the design of a multimodal interaction encoder that can efficiently fuse visual and textual representations, align cross-modal fine-grained features through the MLM task. This design helps the backbone network to extract more discriminative global image-text representations without requiring additional supervision.

To guide the image-text matching, commonly used loss functions include ranking loss and cross-modal projection matching (CMPM) [53] loss. Compared to ranking loss, the CMPM loss does not require the selection of specific triplets or margin parameter tuning. It exhibits great stability with varying batch sizes, making it widely used in text-to-image person retrieval [5, 39, 50]. However, we found that the projection in CMPM can be regarded as a variable weight that adjusts the distribution of softmax output logits, similar to the temperature parameter [17] for knowledge distillation. Nevertheless, limited by the varying projection length, CMPM therefore cannot precisely control the projection probability distribution, making it difficult to focus on hard-negative samples during model updates. To explore more effective cross-modal matching objective, we further propose an image-text similarity distribution matching (SDM) loss. The SDM loss minimizes the KL divergence between the normalized image-text similarity score distributions and the normalized ground truth label matching distributions. Additionally, we introduce a temperature hyperparameter to precisely control the similarity distribution compactness, which enables the model updates focus on hard-negative samples and effectively enlarges the variance between non-matching pairs and the correlation between matching pairs.

To address the limitations of separate pre-trained models on unimodal datasets, we leverage the Contrastive Language-Image Pre-training (CLIP) [35] as the initialization of our model. CLIP is pre-trained with abundant image-text pairs and has powerful underlying cross-modal alignment capabilities. Some previous approaches [13, 50] have either frozen some part of parameters or introduced only CLIP's image encoder, which resulted in their inability to fully exploit CLIP's powerful capabilities in image-text matching. With the proposed IRRA, we successfully transfer the powerful knowledge directly from the pre-trained full CLIP model and continue to learn fine-grained cross-modal implicit local relations on text-to-image person retrieval datasets. In addition, compared to many recent methods [5, 38, 50], IRRA is more efficient as it computes only one global image-text pair similarity score in the inference stage. The main contributions can be summarized as follows:

- We propose IRRA to implicitly utilize fine-grained interaction to enhance the global alignment without requiring any additional supervision and inference cost.

- We introduce a new cross-modal matching loss named

image-text similarity distribution matching (SDM) loss. It directly minimizes the KL divergence between image-text similarity distributions and the normalized label matching distributions.

- We demonstrate that the full CLIP model can be applied to text-to-image person retrieval and can outperform existing state-of-the-art methods with straightforward fine-tuning. Moreover, our proposed IRR module enables fine-grained image-text relation learning, allowing IRRA to learn more discriminative image-text representations.
- Extensive experiments on three public benchmark datasets, *i.e*., CUHK-PEDES [30], ICFG-PEDES [7] and RSTPReid [55] show that IRRA consistently outperforms the state-of-the-arts by a large margin.

2. Related work

Text-to-image Person Retrieval was first introduced by Li *et al*. [30], who proposed the first benchmark dataset, CUHK-PEDES [30]. The main challenge is how to efficiently align image and text features into a joint embedding space for fast retrieval. Early works [2, 29, 30] utilized VGG [40] and LSTM [18] to learn representations for visual-textual modalities and then aligned them using a matching loss. Later works [4, 36, 53] improved the feature extraction backbone with ResNet50/101 [14] and BERT [22], as well as designed novel cross-modal matching losses to align global image-text features in a joint embedding space. More recent works [5, 46, 47, 49, 55] extensively employs additional local feature learning branches that explicitly exploit human segmentation, body parts, color information, and text phrases. There is also some works [7, 10, 38, 51] that implicitly performs local feature learning through attentional mechanisms. However, while these approaches have been shown to provide better retrieval results than using only global features, they also introduce additional computational complexity during inference when computing image-text similarity. The aforementioned works all use backbones pre-trained separately with unimodal data to extract visual and textual features, and then perform cross-modal alignment without exploiting the great cross-modal alignment capabilities of recently promising vision-language pre-training models. Han *et al*. [13] first introduced a CLIP model for text-to-image person retrieval using a momentum contrastive learning framework to transfer the knowledge learned from large-scale generic image-text pairs. Later, Yan *et al*. [50] proposed a CLIP-driven fine-grain information excavation framework to transfer the knowledge of CLIP. However, they failed in directly transferring the original aligned CLIP dual-encoder to text-to-image person retrieval. In this work, we demonstrate that the CLIP model can be easily transferred to text-to-image person retrieval and propose the IRRA to learn more discriminative image-text embeddings.

Vision-Language Pre-training aims to learn the semantic correspondence between vision and language modalities by pre-training on large-scale image-text pairs. Inspired by the success of Transformer-based [44] language model pre-training (such as BERT) [22] and Vision Transformer (ViT) [8], Vision-Language Pre-training (VLP) has emerged as the prevailing paradigm in learning multimodal representations, demonstrating strong results on downstream tasks such as image captioning [3], image-text retrieval [25] and visual question answering [1]. Existing work on VLP can be categorized into two types: single-stream and dual-stream, depending on their model structure. In single-stream models [6, 23, 41], text and visual features are concatenated and then fed into a single transformer encoder. Although this architecture is more parameter-efficient as it uses the same set of parameters for both modalities, it has a slow retrieval speed during the inference stage because it needs to predict the similarity score of all possible image-text pairs. On the other hand, dual-stream models [9, 20, 35] use two separate encoders to extract the text and visual features independently. These two transformer encoders do not share parameters. While achieving remarkable performance on image-text retrieval tasks, dual-stream modals lack the ability to model complex interactions between vision and language for other vision-language understanding tasks.

3. Method

In this section, we present our proposed IRRA framework. The overview of IRRA is illustrated in Fig. 2 and the details are discussed in the following subsections.

3.1. Feature Extraction Dual-Encoder

Previous works in text-to-image person retrieval typically utilize image and text encoders that are pre-trained separately on unimodal datasets. Inspired by the partial success of transferring knowledge from CLIP to text-image person retrieval [13], we directly initialize our IRRA with the full CLIP image and text encoder to enhance its underlying cross-modal alignment capabilities.

Image Encoder. Given an input image $I \in R^{H \times W \times C}$, a CLIP pre-trained ViT model is adopted to obtain the image embedding. We first split I into a sequence of $N = H \times W/P^2$ fixed-sized non-overlapping patches, where P denotes the patch size, and then map the patch sequence to 1D tokens $\{f_i^v\}|_{i=1}^N$ by a trainable linear projection. With injection of positional embedding and extra [CLS] token, the sequence of tokens $\{f_{cls}^v, f_1^v, ..., f_N^v\}$ are input into L-layer transformer blocks to model correlations of each patch. Finally, a linear projection is adopted to map f_{cls}^v to the joint image-text embedding space, which serves as global image representation.

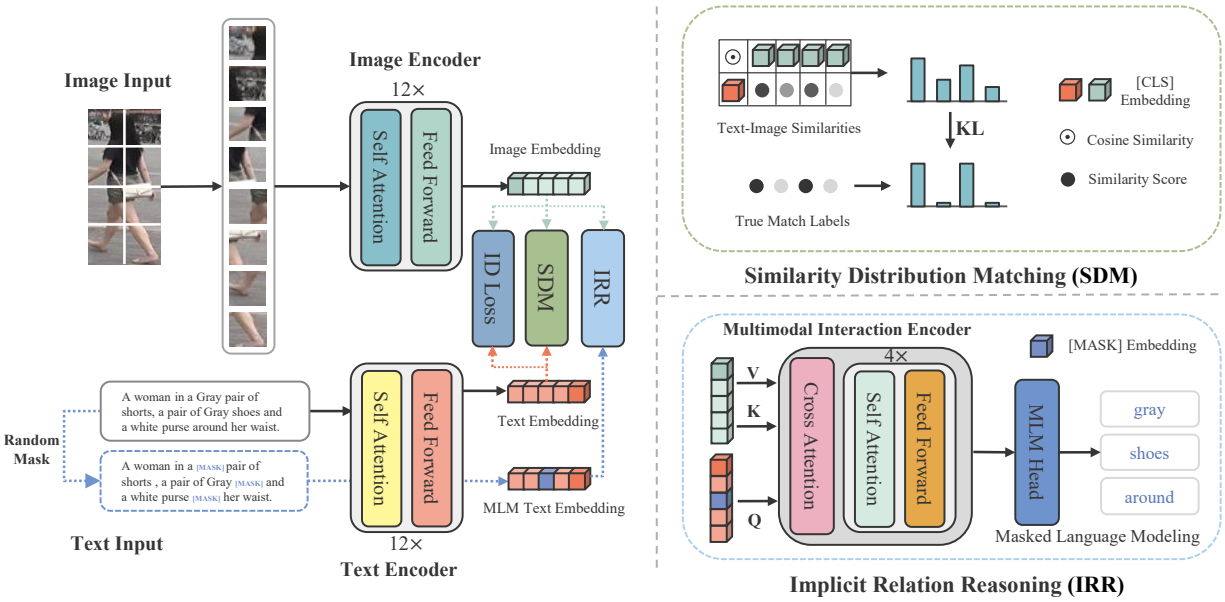


Figure 2. **Overview of the proposed IRRA framework.** It consists of a dual-stream feature extraction backbone and three representation learning branches, *i.e*. Implicit Relation Reasoning (IRR), Similarity Distribution Matching (SDM) and Identity Identification (ID loss). IRR aims to implicitly utilize fine-grained information to learn a discriminative global representation. SDM minimizes the KL divergence between image-text similarity score distributions and true label matching distributions, which can effectively enlarges the variance between non-matching pairs and the correlation between matching pairs. Additionally, we adopt ID loss to aggregate the feature representations of the same identity, further improving the retrieval performance. IRRA is trained end-to-end with these three tasks, and it computes only one global image-text similarity score, making it computationally efficient. Modules connected by dashed lines will be removed during inference stage.

Text Encoder. For an input text T, we directly use the CLIP text encoder to extract the text representation, which is a Transformer [44] modified by Radford *et al*. [35]. Following CLIP, the lower-cased byte pair encoding (BPE) with a 49152 vocab size [37] is firstly employed to tokenize the input text description. The text description is bracketed with [SOS] and [EOS] tokens to indicate the start and end of sequence. Then the tokenized text $\{f^t_{sos}, f^t_1, ...f^t_{eos}\}$ are fed into the transformer and exploit correlations of each patch by masked self-attention. Finally, the highest layer of the transformer at the [EOS] token f^t_{eos} is linearly projected into the image-text joint embedding space to obtain the global text representation.

3.2. Implicit Relation Reasoning

To fully exploit fine-grained information, it is crucial to bridge the significant modality gap between vision and language. While most existing methods do so by explicitly aligning local features between images and text, this paper introduces a novel approach. Specifically, we use MLM to implicitly mine fine-grained relations and learn discriminative global features.

Masked Language Modeling. Masked language modeling (MLM) was initially proposed by Taylor [43] in 1953, it became widely known when the BERT model adapted it as a novel pre-training task. In this work, We utilize MLM to predict masked textual tokens not only by the rest of unmasked textual tokens but also by the visual tokens. Similar to the analysis of Fu *et al*. [11] in pure language pre-training, MLM optimizes two properties: (1) the alignment of image and text contextualized representations with the static embeddings of masked textual tokens, and (2) the uniformity of static embeddings in the joint embedding space. In the alignment property, sampled embeddings of masked textual tokens serve as an anchor to align images and text contextualized representations, as illustrated in Fig. 3. We find that such a local anchor is essential for modeling local dependencies and can implicitly utilize fine-grained local information for global feature alignment.

Multimodal Interaction Encoder. To achieve full interaction between image and text modalities, We design an efficient multimodal interaction encoder to fuse the image and text embeddings, compared to two other popular multimodal interaction modules [9, 16], our design is more computationally efficient, as illustrated in Fig. 4. The multimodal interaction encoder consists of a multi-head cross attention (MCA) layer and 4-layer transformer blocks. Given an input text description T, we randomly mask out the text

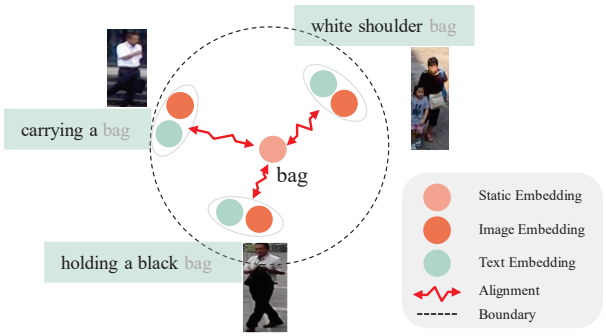


Figure 3. Illustration of the MLM objective. MLM uses static embedding of masked textual tokens as local fine-grained keys to align image and text contextualized representations in the same context.

tokens with a probability of 15% and replace them with the special token [MASK]. Following BERT, the replacements are 10% random tokens, 10% unchanged, and 80% [MASK]. The masked text is defined as $\hat{T}$, and fed into the Text Transformer as described in Sec. 3.1. Then the last hidden states $\{h_i^{\hat{t}}\}_{i=1}^{L}$ and $\{h_i^{v}\}_{i=1}^{N}$ of the text transformer and the vision transformer are fed into the multimodal interaction encoder jointly. In order to fuse image and masked text representations more effectively, the masked text representation $\{h_i^{\hat{t}}\}_{i=1}^{L}$ served as query($\mathcal{Q}$), and the image representation $\{h_i^{v}\}_{i=1}^{N}$ are served as key($\mathcal{K}$) and value($\mathcal{V}$). The full interaction between image and masked text representations can be achieved by:

$$\{h_i^m\}_{i=1}^{L} = Tansformer(MCA(LN(\mathcal{Q},\mathcal{K},\mathcal{V}))), \quad (1)$$

where $\{h_i^m\}_{i=1}^{L}$ is the fused image and masked text contextualized representations, L is the length of input textual tokens, $LN(\cdot)$ denotes Layer Normalization, the $MCA(\cdot)$ is the multi-head cross attention and can be realized by:

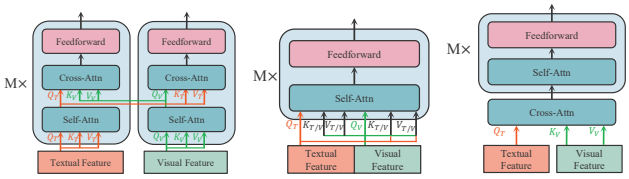
$$MCA(\mathcal{Q},\mathcal{K},\mathcal{V}) = softmax(\frac{\mathcal{Q}\mathcal{K}^\top}{\sqrt{d}})\mathcal{V}, \quad (2)$$

where d is the embedding dimension of masked tokens.

For each masked position $\{h_i^m : i \in \mathcal{M}\}_{i=1}^{L}$, we use a multi-layer perception (MLP) classifier to predict the probability of the corresponding original tokens $\{m_j^i\}_{j=1}^{|\mathcal{V}|} = MLP(h_i^m)$. The IRR objective can be formulated as:

$$\mathcal{L}_{irr} = -\frac{1}{|\mathcal{M}||\mathcal{V}|}\sum_{i\in\mathcal{M}}\sum_{j\in|\mathcal{V}|} y_j^i \log\frac{\exp(m_j^i)}{\sum_{k=1}^{|\mathcal{V}|}\exp(m_k^i)}, \quad (3)$$

where $\mathcal{M}$ denotes the set of masked text tokens and $|\mathcal{V}|$ is the size of vocabulary $\mathcal{V}$. m^i is predicted token probability distribution and y^i is a one-hot vocabulary distribution where the ground-truth token has a probability of 1.



(a) Co-attention (b) Merged attention (c) Ours

Figure 4. **Illustration of our multimodal interaction encoder and two other popular interaction modules.** (a) Co-attention, textual and visual features are fed into separate transformer blocks with self-attn and cross-attn independently to enable cross-modal interaction. (b) Merged attention, textual and visual features are concatenated together and then fed into a single transformer block. (c) Our multimodal interaction encoder, textual and visual features are first fused by a cross-attn layer and then fed into a single transformer block.

3.3. Similarity Distribution Matching

We introduce a novel cross modal matching loss termed as Similarity Distribution Matching (SDM), which incorporates the cosine similarity distributions of the $N \times N$ image-text pairs embeddings into KL divergence to associate the representations across different modalities.

Given a mini-batch of N image-text pairs, for each image global representation f_i^v, we construct a set of image-text representation pairs as $\{(f_i^v, f_j^t), y_{i,j}\}_{j=1}^{N}$, where $y_{i,j}$ is a true matching label, $y_{i,j} = 1$ means that (f_i^v, f_j^t) is a matched pair from the same identity, while $y_{i,j} = 0$ indicates the unmatched pair. Let $sim(\mathbf{u},\mathbf{v}) = \mathbf{u}^\top\mathbf{v}/\|\mathbf{u}\|\|\mathbf{v}\|$ denotes the dot product between $\mathcal{L}_2$ normalized $\mathbf{u}$ and $\mathbf{v}$ (*i.e.* cosine similarity). Then the probability of matching pairs can be simply calculated with the following softmax function:

$$p_{i,j} = \frac{\exp(sim(f_i^v, f_j^t)/\tau)}{\sum_{k=1}^{N}\exp(sim(f_i^v, f_k^t)/\tau)}, \quad (4)$$

where τ is a temperature hyperparameter which controls the probability distribution peaks. The matching probability $p_{i,j}$ can be viewed as the proportion of the cosine similarity score between f_i^v and f_j^t to the sum of cosine similarity score between f_i^v and $\{f_j^t\}_{j=1}^{N}$ in a mini-batch. Then the SDM loss from image to text in a mini-batch is computed by :

$$\mathcal{L}_{i2t} = KL(\mathbf{p_i}\|\mathbf{q_i}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{i,j}\log(\frac{p_{i,j}}{q_{i,j}+\epsilon}), \quad (5)$$

where ϵ is a small number to avoid numerical problems, and $q_{i,j} = y_{i,j}/\sum_{k=1}^{N} y_{i,k}$ is the true matching probability.

Symmetrically, the SDM loss from text to image $\mathcal{L}_{t2i}$ can be formulated by exchanging f^v and f^t in Eq.(4) (5), and the bi-directional SDM loss is calculated by:

$$\mathcal{L}_{sdm} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}. \quad (6)$$

Method	Type	Ref	Image Enc.	Text Enc.	Rank-1	Rank-5	Rank-10	mAP	mINP
CMPM/C [53]	L	ECCV18	RN50	LSTM	49.37	-	79.27	-	-
TIMAM [36]	G	ICCV19	RN101	BERT	54.51	77.56	79.27	-	-
ViTAA [46]	L	ECCV20	RN50	LSTM	54.92	75.18	82.90	51.60	-
NAFS [12]	L	arXiv21	RN50	BERT	59.36	79.13	86.00	54.07	-
DSSL [55]	L	MM21	RN50	BERT	59.98	80.41	87.56	-	-
SSAN [7]	L	arXiv21	RN50	LSTM	61.37	80.15	86.73	-	-
LapsCore [49]	L	ICCV21	RN50	BERT	63.40	-	87.80	-	-
ISANet [51]	L	arXiv22	RN50	LSTM	63.92	82.15	87.69	-	-
LBUL [48]	L	MM22	RN50	BERT	64.04	82.66	87.22	-	-
Han et al. [13]	G	BMVC21	CLIP-RN101	CLIP-Xformer	64.08	81.73	88.19	60.08	-
SAF [28]	L	ICASSP22	ViT-Base	BERT	64.13	82.62	88.40	-	-
TIPCB [5]	L	Neuro22	RN50	BERT	64.26	83.19	89.10		-
CAIBC [47]	L	MM22	RN50	BERT	64.43	82.87	88.37	-	-
AXM-Net [10]	L	MM22	RN50	BERT	64.44	80.52	86.77	58.73	-
LGUR [38]	L	MM22	DeiT-Small	BERT	65.25	83.12	89.00	-	-
IVT [39]	G	ECCVW22	ViT-Base	BERT	65.59	83.11	89.21	-	-
CFine [50]	L	arXiv22	CLIP-ViT	BERT	69.57	85.93	91.15	-	-
Baseline (CLIP-RN50)	G	-	CLIP-RN50	CLIP-Xformer	57.26	78.57	85.58	50.88	34.44
Baseline (CLIP-RN101)	G	-	CLIP-RN101	CLIP-Xformer	60.27	80.88	87.88	53.93	37.54
Baseline (CLIP-ViT-B/16)	G	-	CLIP-ViT	CLIP-Xformer	68.19	86.47	91.47	61.12	44.86
IRRA (Ours)	G	-	CLIP-ViT	CLIP-Xformer	**73.38**	**89.93**	**93.71**	**66.13**	**50.24**

Table 1. Performance comparisons with state-of-the-art methods on CUHK-PEDES dataset. Results are ordered based on the Rank-1 accuracy. "G" and "L" in "Type" column stand for global-matching/local-matching method.

Optimization. As mentioned previously, the main objective of IRRA is to improve the learning of global image-text representations in joint embedding space. To achieve this goal, the commonly utilized ID loss [54] is also adopted along with SDM loss and IRR loss to optimize IRRA. The ID loss is a softmax loss which classifies an image or text into distinct groups based on their identities. It explicitly considers the intra-modal distance and ensures that feature representations of the same image/text group are closely clustered together in the joint embedding space.

IRRA is trained in an end-to-end manner and the overall optimization objective for training is defined as:

$$\mathcal{L} = \mathcal{L}_{irr} + \mathcal{L}_{sdm} + \mathcal{L}_{id}. \tag{7}$$

4. Experiments

We extensively evaluate our method on three challenging text-to-image person retrieval datasets.

CUHK-PEDES [30] is the first dataset dedicated to text-to-image person retrieval, which contains 40,206 images and 80,412 textual descriptions for 13,003 identities. Following the official data split, the training set consists of 11,003 identities, 34,054 images and 68,108 textual descriptions. The validation set and test set contain 3,078 and 3,074 images, 6158 and 6156 textual descriptions, respectively, and both of them have 1,000 identities.

ICFG-PEDES [7] contains a total of 54,522 images for 4,102 identities. Each image has only one corresponding textual description. The dataset is divided into a training set and a test set, the former comprises 34,674 image-text pairs of 3,102 identities, while the latter contains 19,848 image-text pairs for the remaining 1,000 identities.

RSTPReid [55] contains 20505 images of 4,101 identities from 15 cameras. Each identity has 5 corresponding images taken by different cameras and each image is annotated with 2 textual descriptions. Following the official data split, the training, validation and test set contain 3701, 200 and 200 identities respectively.

Evaluation Metrics. We adopt the popular Rank-*k* metrics (*k*=1,5,10) as the primary evaluation metrics. Rank-*k* reports the probability of finding at least one matching person image within the top-k candidate list when given a textual description as a query. In addition, for a comprehensive evaluation, we also adopt the mean Average Precision (mAP) and mean Inverse Negative Penalty(mINP) [52] as another retrieval criterion. The higher Rank-*k*, mAP and mINP indicates better performance.

Implementation Details. IRRA consists of a pre-trained image encoder, *i.e.*, CLIP-ViT-B/16, a pre-trained text encoder, *i.e.*, CLIP text Transformer, and a random-initialized multimodal interaction encoder. For each layer of the multimodal interaction encoder, the hidden size and number of heads are set to 512 and 8. During training, random horizontally flipping, random crop with padding, and random erasing are employed for image data augmentation. All input images are resized to 384×128. The maximum length of the textual token sequence L is set to 77. Our model is trained with Adam optimizer [24] for 60 epochs with a learning rate initialized to 1×10^{-5} and cosine learning rate decay. At the beginning, we spend 5 warm-up epochs linearly increasing the learning rate from 1×10^{-6} to 1×10^{-5}. For random-initialized modules, we set the initial learning rate to 5×10^{-5}. The temperature parameter τ in SDM

loss is set to 0.02. This work is supported by Huawei MindSpore [19]. We perform our experiments on a single RTX3090 24GB GPU.

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

In this section, we present comparison results with state-of-the-art methods on three public benchmark datasets. Note that the Baseline models in Tab. 1 2 and 3 denotes different CLIP models fine-tuned with the original CLIP loss (InfoNCE [34]).

Performance Comparisons on CUHK-PEDES We first evaluate the proposed method on the most common benchmark, CUHK-PEDES. As shown in Tab. 1, IRRA outperforms all state-of-the-art methods, achieving 73.38% Rank-1 accuracy and 66.13% mAP respectively. It is worth noting that our directly fine-tuned CLIP Baseline has already achieved the recent state-of-the-art method CFine [50], with Rank-1 accuracy and mAP reaching 68.19% and 86.47% respectively. In Tab. 1, we annotate the feature extraction backbones ("Image Enc." and "Text Enc." column) employed by each method, and it is evident that there is a growing demand of powerful feature extraction backbone for text-to-image person retrieval, with transformer-based backbone becoming progressively dominant.

Performance Comparisons on ICFG-PEDES The experimental results on the ICFG-PEDES dataset are reported in Tab. 2. The Baseline can achieve comparable results to recent state-of-the-art methods, with 56.74%, 75.72% and 82.26% on Rank-1, Rank-5 and Rank-10, respectively. Moreover, our proposed IRRA achieves 63.46%, 80.24% and 85.82% on these metrics, which exceed the recent state-of-the-art local-matching method Cfine [50] by a large margin, *i.e.*, +2.63%, +3.69% and +3.4%. It is worth noting that the mINP [52] metric on ICFG-PEDES is relatively low, which indicates the inferior capability of IRRA to find the hardest matching samples.

Performance Comparisons on RSTPReid We also report our experimental results on the newly released RSTPReid dataset in Tab. 3. Our proposed IRRA dramatically surpass the recent global-matching method IVT [39] by +13.5%, +11.3% and +9.4% on Rank-1, Rank-5 and Rank-10, respectively. Compared with the recent local-matching method Cfine [50], IRRA also achieves considerable performance gains, with the rise of +9.65%, +8.8% and +6.6% on Rank-1, Rank-5 and Rank-10, respectively.

In summary, our IRRA consistently achieves the best performance for all metrics on all three benchmark datasets. This demonstrates the generalization and robustness of our proposed method.

4.2. Ablation Study

In this subsection, we analyze the effectiveness of each component in the IRRA framework. Here, we adopt the CLIP-ViT-B/16 model fine-tuned with InfoNCE loss as the Baseline to facilitate the ablation study.

Method	Type	Rank-1	Rank-5	Rank-10	mAP	mINP
Dual Path [54]	G	38.99	59.44	68.41	-	-
CMPM/C [53]	L	43.51	65.44	74.26	-	-
ViTAA [46]	L	50.98	68.79	75.78	-	-
SSAN [7]	L	54.23	72.63	79.53	-	-
IVT [39]	G	56.04	73.60	80.22	-	-
ISANet [51]	L	57.73	75.42	81.72	-	-
CFine [50]	L	60.83	76.55	82.42	-	-
Baseline (CLIP-RN50)	G	41.46	63.68	73.04	21.00	2.46
Baseline (CLIP-RN101)	G	44.09	66.27	74.75	22.59	2.84
Baseline (CLIP-ViT-B/16)	G	56.74	75.72	82.26	31.84	5.03
IRRA (Ours)	G	**63.46**	**80.25**	**85.82**	**38.06**	**7.93**

Table 2. Performance comparisons with state-of-the-art methods on ICFG-PEDES dataset.

Method	Type	Rank-1	Rank-5	Rank-10	mAP	mINP
DSSL [55]	G	39.05	62.60	73.95	-	-
SSAN [7]	L	43.50	67.80	77.15	-	-
LBUL [48]	L	45.55	68.20	77.85	-	-
IVT [39]	G	46.70	70.00	78.80	-	-
CFine [50]	L	50.55	72.50	81.60	-	-
Baseline (CLIP-RN50)	G	41.40	68.55	77.95	31.51	12.71
Baseline (CLIP-RN101)	G	43.45	67.75	78.40	29.91	11.18
Baseline (CLIP-ViT-B/16)	G	54.05	80.70	88.00	43.41	22.31
IRRA (Ours)	G	**60.20**	**81.30**	**88.20**	**47.17**	**25.28**

Table 3. Performance comparisons with state-of-the-art methods on RSTPReid dataset.

Ablations on proposed components To fully demonstrate the impact of different components in IRRA, we conduct a comprehensive empirical analysis on three public datasets (*i.e.*, CUHK-PEDES [30], ICFG-PEDES [7] and RSTPReid [55]). The Rank-1, Rank-5, Rank-10 accuracies (%) are reported in Tab. 4.

IRR learns local relations through MLM task which can be easily integrated with other transformer-based methods to facilitate fine-grained cross-modal alignment. The efficacy of IRR is revealed via the experimental results of No.0 *vs*. No.4, No.2 *vs*. No.6 and No.5 *vs*. No.7. Merely adding the IRR to Baseline improves the Rank-1 accuracy by 3.04%, 4.22% and 3.85% on the three datasets, respectively. The above results clearly show that IRR module are beneficial for cross-modal matching.

To demonstrate the effectiveness of our proposed similarity distribution matching (SDM) loss, we compare it with the commonly used cross-modal projection matching (CMPM) loss [53] (No.1 *vs*. No.2) on the three public datasets, the SDM loss promotes the Rank-1 accuracy of the CMPM loss by 11.11%, 6.62%, and 2.2%, respectively. Besides, replace the original InfoNCE loss with the commonly used CMPM loss (No.0 *vs*. No.1) does not improve the performance on text-to-image person retrieval, yet it leads to performance degradation. In contrast, the SDM loss promotes the Rank-1 accuracy of the Baseline by 2.23%, 3.71%, and 3.15% on three datasets, respectively. These results demonstrate that the proposed SDM loss well aligns the features representations between the two modalities. In

No.	Methods	Components			CUHK-PEDES			ICFG-PEDES			RSTPReid		
		SDM	$\mathcal{L}_{id}$	IRR	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
0	Baseline				68.19	86.47	91.47	56.74	75.72	82.26	54.05	80.70	88.00
1	+$\mathcal{L}_{cmpm}$ [53]				59.31	79.66	86.11	53.83	72.20	79.02	55.40	77.70	85.25
2	+SDM	✓			70.42	86.73	92.04	60.45	77.88	83.86	57.20	79.90	88.10
3	+$\mathcal{L}_{id}$		✓		65.33	84.05	90.33	53.38	72.70	79.70	54.15	76.65	85.00
4	+IRR			✓	71.23	88.89	93.24	60.96	79.02	84.90	57.90	80.85	88.50
5	+SDM+$\mathcal{L}_{id}$	✓	✓		70.52	87.59	92.12	61.03	78.26	83.89	58.65	80.70	87.05
6	+SDM +IRR	✓		✓	72.81	89.31	93.39	63.27	80.10	85.77	59.25	79.70	88.00
7	IRRA	✓	✓	✓	**73.38**	**89.83**	**93.71**	**63.46**	**80.25**	**85.82**	**60.20**	**81.30**	**88.20**

Table 4. Ablation study on each component of IRRA on CUHK-PEDES, ICFG-PEDES and RSTPReid.

Method	Param(M)	Time(ms)	Rank-1	Rank-5	Rank-10
Co-attn	33.62	24.30	73.28	89.04	93.44
Merged attn	**12.61**	19.20	73.21	89.18	93.70
Ours	13.66	**6.42**	**73.38**	**89.83**	**93.71**

Table 5. Comparisons between different Multimodal Interaction Module of IRRA on CUHK-PEDES.

addition, the experimental results of No.2 *vs*. No.5 and No.6 *vs*. No.7 demonstrate the effectiveness of the ID loss.

Analysis of the Multimodal Interaction Encoder To demonstrate the advantages of our proposed Multimodal Interaction Module, we compare it with two other popular multimodal interaction modules in Tab. 5. The Multimodal Interaction Module in IRR is a computationally efficient operation to fuse multimodal features, building the connection between the two modalities. We extensively compare it with *Co-attn* and *Merged attn* under our proposed IRRA setting, and observe slight but consistent performance gain on all Rank-*k* metrics. Notably, our major advantage is the computational efficiency.

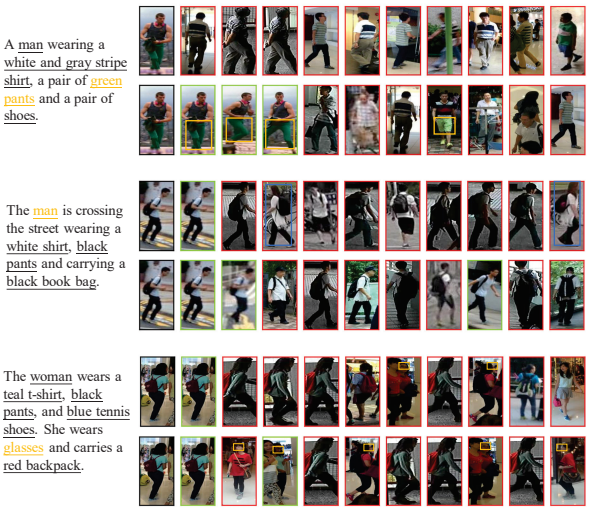


Figure 5. Comparison of top-10 retrieved results on CUHK-PEDES between Baseline (the first row) and IRRA (the second row) for each text query. The image corresponding to query text., matched and mismatched images are marked with black, green and red rectangles, respectively.

4.3. Qualitative Results

Fig. 5 compares the top-10 retrieval results from the Baseline and our proposed IRRA. As the figure shows, IRRA achieves much more accurate retrieval results and obtains accurate retrieval results when Baseline fails to retrieve them. This is mainly due to the Implicit Relation Reasoning (IRR) modules we designed, which fully exploit fine-grained discriminative clues to distinguish different pedestrians. This is illustrated in the orange highlighted text and image regions box in Fig. 5. Moreover, We found that our model only learns the semantic information of the word-level but unable to understand the semantics of the phrase-level in the description text, which leads to the distortion of semantic information. This is because we only masked random single tokens in MLM, and did not perform phrase-level masking. We plan to address this issue in the future.

5. Conclusion

In this paper, we introduce a cross modal implicit relation reasoning and aligning framework(IRRA) to learn discriminative global image-text representations. To achieve full cross-modal interaction, we propose an Implicit Relation Reasoning module that exploits MLM to mine fine-grained relations between visual and textual tokens. We further propose a Similarity Distribution Matching loss to effectively enlarge the variance between non-matching pairs and the correlation between matching pairs. These modules collaborate to align images and text into a joint embedding space. Significant performance gains on three popular benchmarks datasets prove the superiority and effectiveness of our proposed IRRA framework. We believe that the CLIP-based approach will be the future trend for text-to-image person retrieval.

Acknowledgement. This work is partially supported by the Key Research and Development Program of Hubei Province (2021BAA187), National Natural Science Foundation of China under Grant (62176188), Zhejiang lab (NO.2022NF0AB01), the Special Fund of Hubei Luojia Laboratory (220100015) and CAAI-Huawei MindSpore Open Fund.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3

[2] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1879–1887. IEEE, 2018. 3

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[4] Yucheng Chen, Rui Huang, Hong Chang, Chuanqi Tan, Tao Xue, and Bingpeng Ma. Cross-modal knowledge adaptation for language-based person search. *IEEE Transactions on Image Processing*, 30:4057–4069, 2021. 3

[5] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022. 2, 3, 6

[6] 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 *European conference on computer vision*, pages 104–120. Springer, 2020. 2, 3

[7] 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. 2, 3, 6, 7

[8] 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. In *International Conference on Learning Representations*, 2020. 3

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 2, 3, 4

[10] Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. Axm-net: Implicit cross-modal feature alignment for person re-identification. 36(4):4477–4485, 2022. 3, 6

[11] Zhiyi Fu, Wangchunshu Zhou, Jingjing Xu, Hao Zhou, and Lei Li. Contextual representation learning beyond masked language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2701–2714, 2022. 4

[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. 6

[13] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*, 2021. 2, 3, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[15] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 1

[16] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021. 4

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. 2

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[19] Huawei. Mindspore, https://www.mindspore.cn/, 2020. 7

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[21] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11189–11196, 2020. 2

[22] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 3

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 6

[25] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 3

[26] Jie Lei, Xinlei Chen, Ning Zhang, Mengjiao Wang, Mohit Bansal, Tamara L Berg, and Licheng Yu. Loopitr: Combining dual and cross encoder architectures for image-text retrieval. *arXiv preprint arXiv:2203.05465*, 2022. 1

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[28] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2724–2728. IEEE, 2022. 6

[29] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1890–1899, 2017. 3

[30] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1970–1979, 2017. 1, 3, 6, 7

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[32] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1

[33] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021. 1

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 7

[35] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[36] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5814–5824, 2019. 3, 6

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 4

[38] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. *arXiv preprint arXiv:2207.07802*, 2022. 2, 3, 6

[39] 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. *arXiv preprint arXiv:2208.08608*, 2022. 2, 6, 7

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[41] 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. 2, 3

[42] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 982–997, 2021. 1

[43] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953. 4

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[45] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7307, 2022. 1

[46] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *European Conference on Computer Vision*, pages 402–420. Springer, 2020. 2, 3, 6, 7

[47] 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. *arXiv preprint arXiv:2209.05773*, 2022. 3, 6

[48] 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 *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1984–1992, 2022. 6, 7

[49] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: Language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633, 2021. 3, 6

[50] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *arXiv preprint arXiv:2210.10276*, 2022. 2, 3, 6, 7

[51] Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. Image-specific information suppression and implicit local alignment for text-based person search. *arXiv preprint arXiv:2208.14365*, 2022. 3, 6, 7

[52] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 6, 7

[53] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–701, 2018. 2, 3, 6, 7, 8

[54] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional

image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020. 2, 6, 7

[55] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217, 2021. 3, 6, 7