# PLOT: Text-based Person Search with Part Slot Attention for Corresponding Part Discovery

Jicheol Park, Dongwon Kim, Boseung Jeong, and Suha Kwak

Pohang University of Science and Technology (POSTECH), South Korea
{jicheol, kdwon, boseung01, suha.kwak}@postech.ac.kr
https://cvlab.postech.ac.kr/research/PLOT

**Abstract.** Text-based person search, employing free-form text queries to identify individuals within a vast image collection, presents a unique challenge in aligning visual and textual representations, particularly at the human part level. Existing methods often struggle with part feature extraction and alignment due to the lack of direct part-level supervision and reliance on heuristic features. We propose a novel framework that leverages a part discovery module based on slot attention to autonomously identify and align distinctive parts across modalities, enhancing interpretability and retrieval accuracy without explicit part-level correspondence supervision. Additionally, text-based dynamic part attention adjusts the importance of each part, further improving retrieval outcomes. Our method is evaluated on three public benchmarks, significantly outperforming existing methods.

**Keywords:** Text-Based Person Search · Multi-Modal Retrieval

## 1 Introduction

Text-based person search is the task of identifying the target person from the vast collection of images with a free-form text query. This task demands extracting identifiable features, such as human parts, from both textual and visual modalities to capture subtle differences between individuals. Hence, establishing correspondence between the extracted human part features across image and text modalities is essential for accurate text-based person search. However, it is not straightforward to extract these part features and establish their correspondences between the two modalities without part-level supervision.

To address this challenge, previous work [2, 22] relies on the heuristic part features obtained by equi-horizontally cropping the entire image; such features are then matched to the free-form text queries for person search. However, the heuristic part features used in this approach are susceptible to deformation caused by occlusion and pose variation. Meanwhile, earlier work [8, 20, 24] proposed a learning-based approach for part feature extraction. Nevertheless, these methods tend to generate redundant part features that lack disentanglement [20], or they demand access to additional part-level supervisions [8, 24].

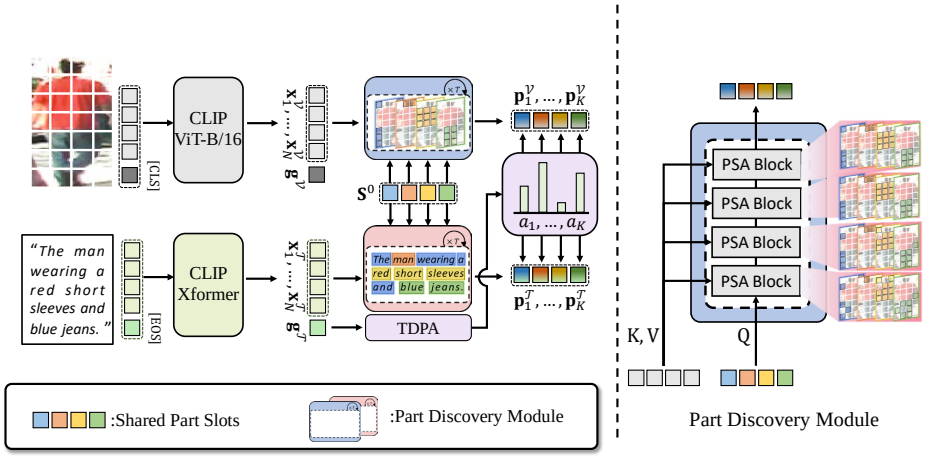


**Fig. 1:** The overall architecture of PLOT.

To tackle the above issues, we introduce a new framework that discovers distinctive parts in both modalities and matches them between the two modalities without any correspondence supervision; its overall architecture is illustrated in Fig. 1. To discover distinctive parts and extract their features from both modalities, we propose PLOT, a **P**art discovery module based on the s**LOT** attention mechanism [15]. Slot attention is an attention mechanism designed for object-centric learning, which segments input data into a set of slots representing individual entities without requiring object-level supervision. In PLOT, we first define a set of learnable embedding vectors, termed *part slots*, that contain primitive information related to human body parts shared between the two modalities. Then, these part slots undergo refinement through several iterative attention processes, where they compete amongst themselves to bind with the input data; ultimately, the part slots are transformed into part features, termed *part embeddings*, that represent distinctive parts in the input data.

To ensure correspondence between part embeddings from different modalities, the part slots are shared between the visual and textual modalities. Part embeddings from the same part slots are then learned to represent the identical part although they are computed from different modalities. This mechanism enables PLOT to match the discovered parts from the two modalities without supervision for the correspondence as well as capturing part-level fine-grained appearance features from the both modalities. Hence, PLOT improves performance of text-based person search through the rich and fine-grained part features, and at the same time, it guarantees interpretable retrieval by providing part-level correspondences between query text and retrieved images.

In addition, PLOT introduces a new similarity aggregation method for part embeddings called text-based dynamic part attention (TDPA) pooling, which dynamically adjusts the weights of part embeddings based on the text query. Given a text query as input, TDPA predicts the importance weight of each

slot for retrieval and applies the predicted importance weights to aggregate the similarities of part embeddings between the two modalities. TDPA allows the entire retrieval system to perform optimized retrieval for each query, leading to improved performance.

Our method was evaluated and compared with prior work on three public benchmarks [4, 11, 32], where it clearly outperformed all existing methods thanks to the rich representation based on part embeddings. The main contribution of our work is four-fold:

- We introduce PLOT, a new framework for text-based person search that discovers distinguished human body parts, extracts their embeddings, and establishes their correspondences between the two modalities with no human intervention.
- The part embeddings provided by PLOT enables an interpretable text-based person search thanks to the part-level correspondences it provides.
- We introduce a novel similarity aggregation method that adaptively determines importance of each discovered part based on each text query and consequently enables retrieval optimized per query.
- Our model with PLOT achieved the best on all the public benchmarks for text-based person search.

## 2 Related Work

### 2.1 Text-Based Person Search

In recent years, the task of text-based person search has gained significant attention in the computer vision community. Li *et al.*, [11] proposed a gated neural attention-based recurrent neural network (GNA-RNN) for learning the affinity between text descriptions and images, along with providing a benchmark dataset CUHK-PEDES for model evaluation. Zhang *et al.*, [30] proposed cross-modal projection matching and classification (CMPM+CMPC) loss, for learning deep discriminative image-text global embeddings. However, these methods primarily focus on global representations of input data and are thus not capable of capturing distinctive human part details, leading to limited performance in the text-based person search.

To address the above problem, a line of studies focuses on extracting fine-grained representations. One of the prominent examples of exploiting fine-grained information is to cut human images horizontally and use them as human parts [2, 4, 5, 16, 22, 23]. Chen *et al.*, [2] extracts image part embeddings through equi-horizontal cropping of the entire image and aligns additional network to transform the textual global into corresponding parts. However, these heuristic part features inevitably include no-informative information, such as background elements. To avoid the above limitations, Suo *et al.*, [22] proposed the simple and robust correlation filtering method to extract foreground features on the heuristic part features. Yet, the heuristically divided part features fundamentally fall short of capturing the complex human parts, underscoring a critical limitation in their

expressive capability. To move beyond such heuristic part structures, Shao *et al.*, [2] proposes a learning-based approach for part feature extraction. However, this method struggles to extract distinctive part embeddings due to the extraction of redundant part information, which lacks sufficient disentanglement. To extract exquisite human parts, several studies tried to utilize useful information (*e.g.*, human attributes and human keypoints) via external tools [1, 8, 24]. Wang *et al.*, [24] introduced an auxiliary attribute segmentation to align the visual part features with the textual attributes parsed from text description. Jing *et al.*, [8] proposed a new multi-granularity attention network to learn the part feature alignment between visual and textual with human pose estimation. However, these approaches have inevitable limitations of high computational cost and dependence on the performance of external tools for local feature extraction.

Most recent work, Jiang *et al.*, [7] utilizes pre-trained CLIP [18] model, capitalizing on the rich knowledge of models trained on extensive data for text-to-image matching, to excel in text-based person search. Yet, this methods primarily focus on global features, not specifically designed for extracting human part features. Not only do we leverage pre-trained knowledge, but we transcend heuristic human parts methods to extract sophisticated human parts without external tools or part-level supervision.

### 2.2 Slot Attention

Slot attention [15] is a recently proposed attention mechanism for object-centric learning, a problem focusing on discovering constituent visual entities within an image. The unique property of slot attention is that it can represent input images as a set of slots, where the slots are representations corresponding to individual visual entities, without any object-level supervision during training. Within slot attention, slots iteratively compete for aggregating input data, ensuring distinct representations focusing on individual visual entities. By incorporating slot attention, the proposed framework facilitates the unsupervised identification of a structure underlying image and text queries, enabling the model to discern and represent individual human parts without explicit supervision. This capability is particularly valuable in person search datasets, where recognizing subtle differences and understanding the correspondence of human parts across different modalities are crucial.

## 3 Proposed Method

The following subsection offers details of global and part embeddings extraction for each modality with its backbone (Sec 3.1). Subsequently, we present our novel framework, PLOT, which includes the part discovery module (Sec. 3.2) and similarity aggregation between part embeddings (Sec. 3.3), concluding with a discussion on the learning objective designed to optimize our proposed framework (Sec. 3.4) and inference of our framework (Sec. 3.5).

### 3.1 Global and Part Embeddings

In our framework, a single input data is described by two different types of representations: a global embedding and multiple part embeddings. The global embedding is used to represent the input data holistically, while each part embedding describes the appearance of distinctive human parts (*e.g.*, arm, leg, torso, etc.). We below provide details of global and part embeddings computation for visual and textual modalities. Following the previous work [7], a pre-trained CLIP [18] is used as the backbone networks for the visual and textual modalities.
**Visual Modality:** We utilize the vision transformer (ViT) from the CLIP-B/16 [18] architecture as a visual backbone network. Initially, an input image of a person is split into $N$ distinct, non-overlapping patches, which are subsequently transformed into patch tokens through linear projection. The patch tokens and an extra `[cls]` token are then fed into the visual backbone network. The token sequence is processed throughout multiple self-attention blocks, and the `[cls]` token of the last block is used as a global embedding $\mathbf{g}^{\mathcal{V}} \in \mathbb{R}^{D}$. To obtain part embeddings $\mathbf{P}^{\mathcal{V}} \in \mathbb{R}^{K \times D}$, remaining patch tokens of last block $\mathbf{x}^{\mathcal{V}} \in \mathbb{R}^{N \times D}$ is passed to the part discovery module. The part discovery module aggregates the patch features describing coherent human parts into the same part embedding. We provide more detailed information about the part discovery module in Sec. 3.2.
**Textual Modality:** For the textual backbone network, we utilize a transformer architecture from the CLIP-Xformer [18] text encoder. This encoder operates on text input transformed into byte pair encoding (BPE) sequences. Initially, the text query undergoes tokenization via BPE, followed by the enclosing with `[SOS]` and `[EOS]` tokens. The resulting sequence of tokens is then inputted into the textual backbone network. Here, the `[EOS]` token from the final block serves as the global embedding, denoted as $\mathbf{g}^{\mathcal{T}} \in \mathbb{R}^{D}$. Analogous to the approach for the visual modality, we process the remainder of the text tokens, $\mathbf{x}^{\mathcal{T}} \in \mathbb{R}^{L \times D}$, into part embeddings, $\mathbf{P}^{\mathcal{T}} \in \mathbb{R}^{K \times D}$, utilizing the part discovery module.

### 3.2 Part Discovery Module

For extracting part embeddings in each modality, part discovery module aggregates patch tokens $\mathbf{x}^{\mathcal{V}}$ and text tokens $\mathbf{x}^{\mathcal{T}}$ into the visual part embeddings $\mathbf{P}^{\mathcal{V}}$ and textual part embeddings $\mathbf{P}^{\mathcal{T}}$, respectively. It is worth noting that the part discovery module for each modality has identical model architecture and functions equivalently. Therefore, we will only explain the part discovery module on the visual modality for brevity.

Part discovery module consists of initial part slots $\mathbf{S}^0 \in \mathbb{R}^{K \times D}$ and $T$ multiple iteration of the *part slot attention block* (PSA block). To extract part embeddings, we first initialize a set of learnable embeddings *part slots* $\mathbf{S}^0 \in \mathbb{R}^{K \times D}$, where $K$ indicates the number of part slots. Then, through a series of $T$ iterations of our PSA block, the initial $\mathbf{S}^0$ evolves into refined $\mathbf{S}^T$, where each slot captures distinct parts within input data. The refined part slots $\mathbf{S}^T$ are used as visual part embeddings $\mathbf{P}^{\mathcal{V}}$. The part discovery module can be formulated as follows:

$$\mathbf{P}^{\mathcal{V}} := \mathbf{S}^T, \text{where} \quad \mathbf{S}^t = \texttt{PSA\_Block}^{\mathcal{V}}(\mathbf{x}^{\mathcal{V}}; \mathbf{S}^{t-1}). \tag{1}$$

The PSA block first transforms the inputs $\mathbf{S}^{t-1}$ and $\mathbf{x}^{\mathcal{V}}$ with layer normalization and linear projection layers $q(\cdot), k(\cdot)$ and $v(\cdot)$ to obtain embeddings of $D_h$ dimension. Then the attention map $A \in \mathbb{R}^{N \times K}$ between $\mathbf{S}^{t-1}$ and $\mathbf{x}^{\mathcal{V}}$ is computed by

$$A_{n,k} = \frac{e^{M_{n,k}}}{\sum_{i=1}^{K} e^{M_{n,i}}}, \text{ where } M = \frac{k(\mathbf{x}^{\mathcal{V}}) q(\mathbf{S}^{t-1})^{\top}}{\sqrt{D_h}}. \tag{2}$$

The attention map is obtained through normalization across part slots; this normalization encourages competition among themselves to bind distinct sets of patch tokens to each slot. We update part slots with weighted mean of patch tokens obtained by attention map $A_{n,k}$ and then feed it to a gated recurrent unit (GRU) using $\mathbf{S}^{t-1}$ as hidden state as follows:

$$\bar{A}_{n,k} = \frac{A_{n,k}}{\sum_{i=1}^{N} A_{i,k}}, \ \bar{\mathbf{S}}^{t} = \texttt{GRU}\big(\mathbf{S}^{t-1}, \bar{A}^{\top} v(\mathbf{x}^{\mathcal{V}})\big). \tag{3}$$

Then, we obtain $t$-th part slot by feeding $\bar{\mathbf{S}}^{t-1}$ into a multi-layer perceptron (MLP) with layer normalization, ReLU activation, and residual connection:

$$\mathbf{S}^{t} = \texttt{PSA\_Block}^{\mathcal{V}}(\mathbf{x}^{\mathcal{V}}; \mathbf{S}^{t-1}) = \texttt{MLP}(\bar{\mathbf{S}}^{t-1}) + \bar{\mathbf{S}}^{t-1}. \tag{4}$$

Finally, we can obtain the visual part embeddings $\mathbf{P}^{\mathcal{V}} = \{\mathbf{p}_k^{\mathcal{V}}\}_{k=1}^{K}$ which is the output of $T$-th iteration of PSA block: $\mathbf{P}^{\mathcal{V}} = \mathbf{S}_{\mathcal{V}}^{T} \in \mathbb{R}^{K \times D}$.

**Part Correspondence through Slot Sharing:** Additionally, we share the learnable part slots $\mathbf{S}^0$ between two part discovery modules to establish correspondences between part embeddings extracted from each modality, considering that part embeddings extracted from the same part slot are corresponding part across modalities, thereby contributing to a clearer comparison between modalities.

### 3.3 Measuring Similarity between Embeddings

In the context of training and applying our retrieval model, selecting an appropriate similarity function between embeddings is crucial. The challenge lies in dealing with two distinct types of embeddings: global embeddings and part embeddings. For global embeddings, cosine similarity offers a straightforward and effective means of measuring similarity.

However, the situation becomes more complex when considering part embeddings. A direct method for addressing this complexity involves calculating the average cosine similarity across all pairs of part embeddings. This approach, while straightforward, has its drawbacks, primarily because it treats all part-wise similarities as equally significant. In reality, the relevance of specific part embeddings to the actual similarity between data instances can significantly vary, influenced by the context of a text query. For instance, if a text query focuses exclusively on particular features of a human figure, the similarity contributions from other unrelated part embeddings should be less relevant. This challenge highlights the necessity for an approach that can dynamically assess and prioritize the relevance of part embeddings based on the context provided by the query.

To address this challenge, we introduce text-based dynamic part attention (TDPA) to aggregate the similarities between part embeddings. Firstly, textual global embedding $\mathbf{g}^{\mathcal{T}}$ is transformed to the TDPA $\mathbf{a} \in \mathbb{R}^K$, using MLP and a softmax function. Using TDPA, the aggregated similarity between part embeddings are computed as follows:

$$\mathbf{a} = \sigma\big(\texttt{MLP}(\mathbf{g}^{\mathcal{T}})\big) \in \mathbb{R}^K, \tag{5}$$

$$c_{\text{agg}}(\mathbf{P}^{\mathcal{V}}, \mathbf{P}^{\mathcal{T}}; \mathbf{g}^{\mathcal{T}}) := \sum_{k=1}^{K} \mathrm{a}_k \cdot c(\mathbf{p}_k^{\mathcal{V}}, \mathbf{p}_k^{\mathcal{T}}), \tag{6}$$

where $c(\cdot,\cdot)$ denotes cosine similarity between two embeddings, $\mathrm{a}_k$ is the $k$-th value of $\mathbf{a}$, and $\sigma(\cdot)$ is a softmax function. If given textual global embedding $\mathbf{g}^{\mathcal{T}}$, TDPA is computed by $\texttt{MLP}$ which is learned in an end-to-end manner by minimizing our partNCE loss, as will be introduced in Sec 3.4. This dynamic attention enables us to adaptively assign importance to each part embedding, which reduces the impact of non-informative part embedding and promotes a more informative similarity measurement that reflects the significant semantic similarity of each text query.

### 3.4 Learning Objective

Our model is trained through the establishment of cross-modal alignments, considering both global and part levels. The alignment at the global-level is accomplished by aligning global embeddings $\mathbf{g}^{\mathcal{V}}$ and $\mathbf{g}^{\mathcal{T}}$ that capture the comprehensive information of each modality. For part-level alignment, we leverage part embeddings $\mathbf{P}^{\mathcal{V}}$ and $\mathbf{P}^{\mathcal{T}}$ that have locally exclusive features within each modality due to the slot attention. Furthermore, by dynamically adjusting weights of specific part slots based on the text global embedding $\mathbf{g}^{\mathcal{T}}$, we facilitate the learning of more informative alignments.

**Global Alignment Loss:** To align global embeddings $\mathbf{g}^{\mathcal{V}}$ and $\mathbf{g}^{\mathcal{T}}$ extracted from each modality backbone, we first define a batch of global embeddings $\mathcal{B}_{\text{global}} = \{(\mathbf{g}_i^{\mathcal{V}}, \mathbf{g}_i^{\mathcal{T}})\}_{i=1}^{B}$, where $B$ is batch size. Then we adopt the InfoNCE loss [17] which is a contrastive learning objective that maximizes the similarity between embeddings of positive pairs while minimizing the similarity between negative pairs in the batch. Consequently, our globalNCE loss with cosine similarity function $c(\cdot,\cdot)$ is formulated as follows:

$$\mathcal{L}_{\text{NCE}} = -\sum_{i=1}^{B}\left(\log\frac{e^{c(\mathbf{g}_i^{\mathcal{V}}, \mathbf{g}_i^{\mathcal{T}})/\tau}}{\sum_{j=1}^{B} e^{c(\mathbf{g}_i^{\mathcal{V}}, \mathbf{g}_j^{\mathcal{T}})/\tau}} + \log\frac{e^{c(\mathbf{g}_i^{\mathcal{V}}, \mathbf{g}_i^{\mathcal{T}})/\tau}}{\sum_{j=1}^{B} e^{c(\mathbf{g}_j^{\mathcal{V}}, \mathbf{g}_i^{\mathcal{T}})/\tau}}\right), \tag{7}$$

where the $\tau$ is temperature term. Additionally, we employ an identity classification loss $\mathcal{L}_{\text{ID}}$ to ensure that the embeddings extracted from the same identity become similar. The $\mathcal{L}_{\text{ID}}$ is denoted by

$$\mathcal{L}_{\text{ID}} = -\sum_{i=1}^{B}\Big(\boldsymbol{y}_i \log\sigma(\mathbf{g}_i^{\mathcal{V}}\mathbf{W}_{\text{ID}}) + \boldsymbol{y}_i \log\sigma(\mathbf{g}_i^{\mathcal{T}}\mathbf{W}_{\text{ID}})\Big), \tag{8}$$

where $\boldsymbol{y}_i \in \mathbb{R}^C$ is the identity ground truth of corresponding global embedding $\mathbf{g}_i$ represented by a one-hot vector and $C$ is the number of identities, $\mathbf{W}_{\text{ID}} \in \mathbb{R}^{D \times C}$ is a classifier shared between the two modalities. For cross-modal alignments, the commonly used CMPM loss [7, 30] is additionally adopted. Finally, we describe global alignment loss as below:

$$\mathcal{L}_{\text{Global}} = \mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{CMPM}}. \tag{9}$$

**Part Alignment Loss:** Similar to the alignment of global embeddings, we first define a batch of part embeddings $\mathcal{B}_{\text{part}} = \{(\mathbf{P}_i^{\mathcal{V}}, \mathbf{P}_i^{\mathcal{T}})\}_{i=1}^{B}$ with batch size $B$, and we adopt the InfoNCE loss to align part embeddings extracted from two modalities. For learning text query-based informative alignment of part embeddings between modalities, we employ text query-based similarity aggregation function $c_{\text{agg}}$ in Eq. 6 to compute InfoNCE loss. We termed this InfoNCE with $c_{\text{agg}}$ as PartNCE loss, and it is formulated by

$$\mathcal{L}_{\text{PartNCE}} = -\sum_{i=1}^{B}\left(\log\frac{e^{c_{\text{agg}}(\mathbf{P}_i^{\mathcal{V}},\mathbf{P}_i^{\mathcal{T}};\, \mathbf{g}_i^{\mathcal{T}})/\tau}}{\sum_{j=1}^{B} e^{c_{\text{agg}}(\mathbf{P}_i^{\mathcal{V}},\mathbf{P}_j^{\mathcal{T}};\, \mathbf{g}_j^{\mathcal{T}})/\tau}} + \log\frac{e^{c_{\text{agg}}(\mathbf{P}_i^{\mathcal{V}},\mathbf{P}_i^{\mathcal{T}};\, \mathbf{g}_i^{\mathcal{T}})/\tau}}{\sum_{j=1}^{B} e^{c_{\text{agg}}(\mathbf{P}_j^{\mathcal{V}},\mathbf{P}_i^{\mathcal{T}};\, \mathbf{g}_i^{\mathcal{T}})/\tau}}\right), \tag{10}$$

Similar to global alignment loss, we adopt the identity loss that shares classifier weights between the two modalities for part alignment; however, the difference is that the part embeddings are concatenated along the embedding dimension. The identity loss for part alignment is formulated by

$$\mathcal{L}_{\text{PartID}} = -\sum_{i=1}^{B}\left(\boldsymbol{y}_i \log\sigma([\mathbf{P}_i^{\mathcal{V}}]\mathbf{W}_{\text{PartID}}) + \boldsymbol{y}_i \log\sigma([\mathbf{P}_i^{\mathcal{T}}]\mathbf{W}_{\text{PartID}})\right), \tag{11}$$

where $[\mathbf{P}] \in \mathbb{R}^{KD}$ is the concatenation of part embeddings $\mathbf{P}$ along the embedding dimension, and $\mathbf{W}_{\text{PartID}} \in \mathbb{R}^{KD \times C}$ is a classifier shared between both modalities. Finally, the part alignment loss is computed by

$$\mathcal{L}_{\text{Part}} = \mathcal{L}_{\text{PartNCE}} + \mathcal{L}_{\text{PartID}}. \tag{12}$$

**Cross-Modal Masked Language Modeling Loss:** Following the practices of employing Transformer-based backbones in existing methods [7, 21], we adopt an auxiliary loss, the cross-modal masked language modeling (CMLM) loss, to facilitate the learning of interactions between modalities. Similar to BERT [3], given a text description, we randomly select text tokens with a 15% probability and replace them with the learnable [MASK] token. The masked text description is then processed through the text backbone to obtain the masked textual tokens. After obtaining the masked textual tokens, we concatenate them with the visual tokens extracted via the image backbone.

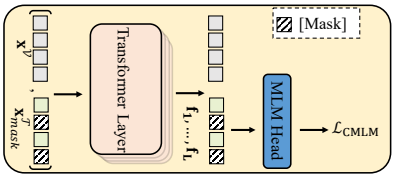


**Fig. 2:** Illustration of CMLM.

The concatenated tokens are then fed into a transformer to acquire cross-modal fused tokens. Among the fused tokens, those corresponding to the indices of textual tokens are denoted as $\mathbf{F}$ composed as $\{\mathbf{f}_1, \cdots \mathbf{f}_L\}$, where a fused token $\mathbf{f} \in \mathbb{R}^{1\times D}$ and $L$ indicates the max length of input text descriptions. Ultimately, these tokens are fed into a CMLM classifier $\mathbf{W}_{\text{CMLM}}$ to predict the probability of vocabulary IDs. The overall procedure of CMLM is illustrated in Fig. 2 and through this procedure we compute the $\mathcal{L}_{\text{CMLM}}$ loss as follows:

$$\mathcal{L}_{\text{CMLM}} = -\frac{1}{L}\sum_{l=1}^{L} \boldsymbol{y}_l \log\left(\sigma(\mathbf{f}_l \mathbf{W}_{\text{CMLM}})\right), \tag{13}$$

where $\mathbf{W}_{\text{CMLM}} \in \mathbb{R}^{D\times V}$ and $\boldsymbol{y}_l \in \mathbb{R}^V$ is the vocabulary ground truth of $l$-th text tokens, represented by a one-hot vector and $V$ is the size of vocabulary. Finally, our overall objective function for training is denoted by

$$\mathcal{L} = \mathcal{L}_{\text{Global}} + \mathcal{L}_{\text{Part}} + \mathcal{L}_{\text{CMLM}}. \tag{14}$$

### 3.5 Inference

During testing, the global and part embeddings of each modality input are fully exploited to calculate the similarity between the image-text pair. In particular, the similarities between visual part embeddings and their corresponding textual part embeddings are linearly combined with the attention weights $\mathbf{a}$ to aggregate them. The image-text pair similarity is defined as the sum of the similarity between the global embeddings of the image-text pair and the similarity between the part embeddings of it, which can be computed by $c(\mathbf{g}^{\mathcal{V}}, \mathbf{g}^{\mathcal{T}}) + c_{\text{agg}}(\mathbf{P}^{\mathcal{V}}, \mathbf{P}^{\mathcal{T}}; \mathbf{g}^{\mathcal{T}})$. Finally, given the text query, the images in the gallery are ranked according to similarity scores between the images and the text for inference.

## 4 Experiments

In this section, we provide a detailed account of our experimental setup (Sec. 4.1), evaluate our method, and compare it with state of the arts on three benchmark datasets for text-based person search (Sec. 4.2). Furthermore, we qualitatively present retrieval results and analyze the effectiveness of the part discovery module and TDPA with visualization results (Sec. 4.3). We also conduct ablation studies on the losses employed in model training, the methodologies of part discovery, and the strategies for part similarity aggregation (Sec. 4.4).

### 4.1 Experimental Setup

**Datasets:** On three benchmark datasets, CUHK-PEDES [11], ICFG-PEDES [4], and RSTPReid [32], we evaluate and compare the performance of our method against previous methods. In CUHK-PEDES collected from five existing person re-identification datasets [6,12,13,28,31], it contains 40,206 images corresponding

to 13,003 individual IDs, with each image being approximately matched with two annotated text descriptions. We follow the data split of [11] with 34,054 images from 11,003 person IDs and 68,126 text descriptions for training, 3,078 images from 1,000 IDs and 6,158 text descriptions for validation, and 3,074 images from 1,000 IDs and 6,156 text descriptions for testing. The remaining two datasets are collected from MSMT17 [26]. ICFG-PEDES consists of 54,522 image-text pairs from 4,102 individual IDs, which are split into 34,674 and 19,848 for training and testing, respectively. RSTPReid contains 20,505 images of 4,101 individual IDs, with each ID having 5 images and each image associated with the corresponding two annotated text descriptions. We follow the data split of [32] with 18,505 images from 3,701 IDs and 37,010 text descriptions for training, 1,000 images from 200 IDs and 2,000 text descriptions for validation, and 1,000 images from 200 IDs and 2,000 text descriptions for testing, respectively.

**Evaluation Protocol:** We employ the standard metric of rank at K (R@K=1,5, 10) for all retrieval experiments. Specifically, given a query text, images are sorted based on their similarity to the query text. The search is considered correct if at least one relevant image appears in the top K positions of the ranking.

**Network Architecture:** We adopt the pre-trained CLIP models from OpenAI [18] for both image and text encoders, where the size of the image encoder is ViT-B/16. The input images are resized to 384×128. Random horizontal flipping, random cropping, and random erasing are applied for the data augmentation in training time. The maximum text length is set to 77.

**Network Optimization:** Our model is trained using the Adam optimizer for 60 epochs with a batch size of 128 for all experiments. For the CLIP encoders, the initial learning rate is set to 5e−6, using a cosine schedule with the warm-up strategy at the first five epochs; we use a high learning rate for the remaining parameters by scaling 20 times.

**Hyperparameters:** We set the number of part slots to 8. The number of iterations in the part slot attention block is set to 5. The temperature parameter $\tau$ is set to 0.015.

### 4.2 Quantitative Results

We compare our method with previous text-based person search methods on CUHK-PEDES [11], ICFG-PEDES [4], and RSTPReid [32]. The performance comparison and the backbones for each modality employed by each method are shown in Table 1. Specifically, our method achieves an outstanding R@1 metric of 75.28%, 65.76%, and 61.80% on the CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets, respectively, outperforming previous methods utilizing different backbones from ours. Moreover, our method improves the previous state of the art, IRRA [7], on R@1 by a large margin as 1.9%p, 1.7%p, and 1.6%p, respectively. Since IRRA only focuses on aligning the global embeddings of each modality, it is hard to capture the fine-grained differences. In contrast, our method not only takes account of the global embeddings but also aims to discover and align discriminative part embeddings from each modality; it allows the model to effectively find the target person.

**Table 1:** Performance of text-based person search methods on the three datasets. **Bold** and underline denote the best and the second best.

| | Backbone | | CUHK-PEDES | | | ICFG-PEDES | | | RSTPReid | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | Image | Text | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| GNA-RNN [11] | RN50 | LSTM | 19.05 | - | 53.64 | - | - | - | - | - | - |
| CMPM/C [30] | RN50 | LSTM | 49.37 | 71.69 | 79.27 | 43.51 | 65.44 | 74.26 | - | - | - |
| PMA [8] | RN50 | BERT | 53.81 | 73.54 | 81.23 | - | - | - | - | - | - |
| TIMAM [19] | RN101 | BERT | 54.51 | 77.56 | 84.78 | - | - | - | - | - | - |
| SCAN [9] | RN50 | BERT | 55.86 | 75.97 | 83.69 | 50.05 | 69.65 | 77.21 | - | - | - |
| ViTAA [24] | RN50 | LSTM | 55.97 | 75.84 | 83.52 | 50.98 | 68.79 | 75.78 | - | - | - |
| NAFS [5] | RN50 | BERT | 59.94 | 79.86 | 86.70 | - | - | - | - | - | - |
| DSSL [32] | RN50 | BERT | 59.98 | 80.41 | 87.56 | - | - | - | 32.43 | 55.08 | 63.19 |
| MGEL [23] | RN50 | LSTM | 60.27 | 80.01 | 86.74 | - | - | - | - | - | - |
| SSAN [4] | RN50 | LSTM | 61.37 | 80.15 | 86.73 | 54.23 | 72.63 | 79.53 | 43.50 | 67.80 | 77.15 |
| LapsCore [27] | RN50 | BERT | 63.40 | - | 87.80 | - | - | - | - | - | - |
| SRCF [22] | RN50 | BERT | 64.04 | 82.99 | 88.81 | 57.18 | 75.01 | 81.49 | - | - | - |
| LGUR [20] | RN50 | BERT | 64.21 | 81.94 | 87.93 | 57.42 | 74.97 | 81.45 | - | - | - |
| TIPCB [2] | RN50 | BERT | 64.26 | 83.19 | 89.10 | - | - | - | - | - | - |
| CAIBC [25] | RN50 | BERT | 64.43 | 82.87 | 88.37 | - | - | - | 47.35 | 69.55 | 79.00 |
| SAF [10] | ViT-B/16 | BERT | 64.13 | 82.62 | 88.40 | - | - | - | - | - | - |
| IVT [21] | ViT-B/16 | BERT | 65.59 | 83.11 | 89.21 | 56.04 | 73.60 | 80.22 | 46.70 | 70.00 | 78.80 |
| CFine [29] | CLIP-ViT-B/16 | BERT | 69.57 | 85.93 | 91.15 | 60.83 | 75.55 | 82.42 | 50.55 | 72.50 | 81.60 |
| IRRA [7] | CLIP-ViT-B/16 | CLIP-Xformer | 73.38 | 89.93 | 93.71 | 63.46 | 80.24 | 85.82 | 60.20 | 81.30 | 88.20 |
| Ours | CLIP-ViT-B/16 | CLIP-Xformer | **75.28** | **90.42** | **94.12** | **65.76** | **81.39** | **86.73** | **61.80** | **82.85** | **89.45** |

### 4.3 Qualitative Results

**Retrieval Results:** Top-5 retrieval results of our method on the CUHK-PEDES dataset are illustrated in Fig. 3. Above all, it shows the overall satisfactory retrieval results. In particular, we can observe our model retrieves targets well, even with distinctive human parts that are small or located in various positions. For instance, the small distinctive human parts like "ponytails" and "high skirts" in Fig. 3(a), "black shoes" in Fig. 3(b), and "blue plaid shorts" in Fig. 3(c), as well as human parts that could appear in various viewpoints such as "yellow shoulder bags" in Fig. 3(a), "floaty dresses" in Fig. 3(b), and "red backpacks" in Fig. 3(d). The CUHK-PEDES dataset typically contains three target images on average in the search space, thereby most of the false matches in the figure are observed by a lack of additional targets. Despite the false matching due to the limitation of the dataset, the retrieval results are reasonable in that the retrieved false matching contains distinctive human parts described by the query description.

**Visualization of Attention Map $\bar{A}_k$ in PSABlock:** To demonstrate the effectiveness of our part discovery module, we visualize the attention map $\bar{A}_k$ in $T$-th iteration of the PSA block for both visual and textual modalities (in Eq. (2)). The visualization results are illustrated in Fig. 4. It not only demonstrates that the part embeddings extracted by our part discovery module capture distinctive human parts but also shows that the part embeddings extracted from the same part slot attend to the semantically identical human parts regardless of modality. For example, the 1st part slot typically focuses on bottom clothes, the 4th on footwear, the 5th on objects being held, the 7th on top clothes, and the 8th on the person's head. Furthermore, in Fig. 4 compared (a) and (b) of visual modality,

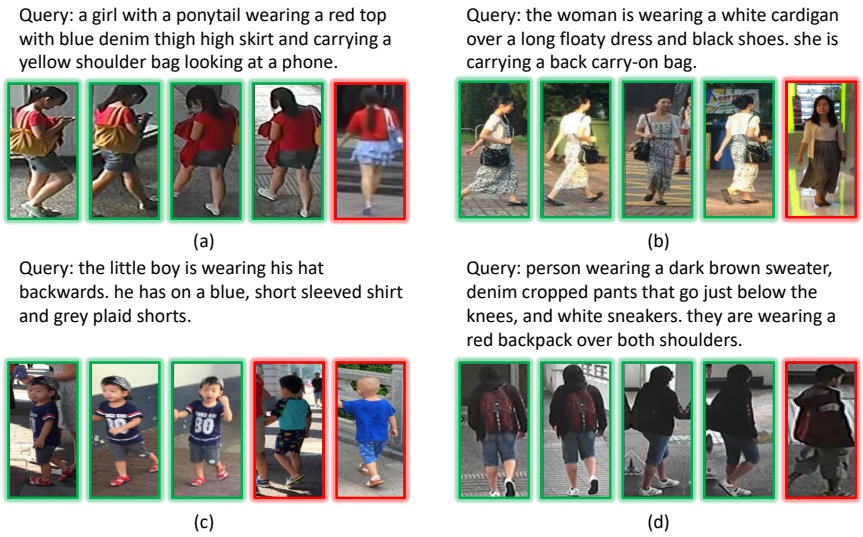


**Fig. 3:** Top-5 retrieval results of our method on the CUHK-PEDES dataset. Images are sorted from left to right according to their ranks below each text query. Green and red boxes indicate true and false matches, respectively.

our part discovery module is capable of capturing the distinctive human parts, while it is robust against pose variations and viewpoint changes.

**Visualization of TDPA:** We visualize TDPA weights $\mathbf{a}$ in Eq. (6) to demonstrate the effectiveness of similarity aggregation between part embeddings of two modalities with TDPA. The visualization results are presented in the bottom left in Fig. 4(a, b), respectively. Comparing examples (a) and (b), the 5th part slot typically focuses on human parts associated with held objects, like bags, leading to a predicted lower TDPA weight for this part slot in (b) due to the absence of such distinctive information in its query description. However, the presence of the "hat" in the query description of (b) leads to having a high TDPA weight for the 8th part slot that typically focuses on the human head part. In contrast, since this related human head part information is not provided in the query description of (a), the TDPA weight for the 8th part slot in (a) is predicted to be a low value. These observations highlight the capability of TDPA, which adaptively improves part-based retrieval depending on the contents of the text query.

### 4.4 Ablation Studies

In our ablation studies conducted on the CUHK-PEDES datasets, we evaluate the effectiveness of proposed components and their combinations in improving text-based person search performance. In Table 2, we compare several configurations: The baseline method employs only global embeddings $\mathbf{g}$ trained by InfoNCE ($\mathcal{L}_{\text{NCE}}$) loss. When incorporating the cross-modal masked language modeling ($\mathcal{L}_{\text{CMLM}}$), there is a slight improvement across all metrics. The addition

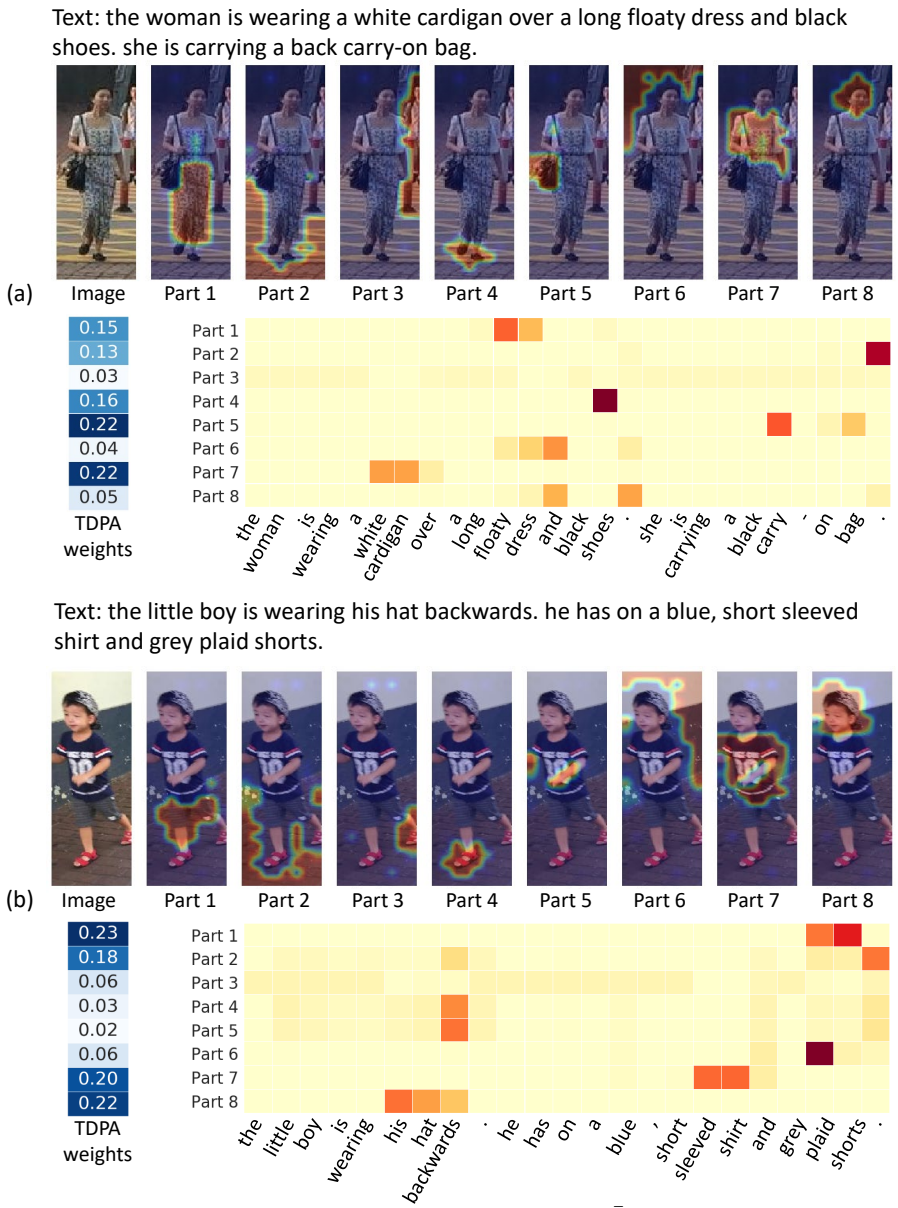


**Fig. 4:** Visualization of each modality's attention map $\bar{A}_k$ in $T$-th iteration of PSA block and TDPA weights **a** on CUHK-PEDES dataset.

of part embeddings $\mathbf{P}$ that without $\mathcal{L}_{\text{PartID}}$ also shows significant enhancement in R@1 and R@5, where we observe an increase of 3.46%p and 2.64%p, respectively. Notably, the full configuration achieves the best performance with $\mathcal{L}_{\text{PartID}}$, underscoring the importance of part embeddings for accurate text-based person search.

**Table 2:** Ablation studies on the CUHK-PEDES datasets.

| Method | Loss | | | | | CUHK-PEDES | | |
|---|---|---|---|---|---|---|---|---|
| | $\mathcal{L}_{\text{NCE}}$ | $\mathcal{L}_{\text{ID}}$ | $\mathcal{L}_{\text{CMLM}}$ | $\mathcal{L}_{\text{PartNCE}}$ | $\mathcal{L}_{\text{PartID}}$ | R@1 | R@5 | R@10 |
| Global Only | ✓ | ✗ | ✗ | ✗ | ✗ | 71.39 (+0.0) | 87.65 (+0.0) | 92.74 (+0.0) |
| | ✓ | ✓ | ✗ | ✗ | ✗ | 71.83 (+0.44) | 88.06 (+0.41) | 92.58 (-0.16) |
| + CMLM | ✓ | ✓ | ✓ | ✗ | ✗ | 72.65 (+1.26) | 88.58 (+0.93) | 92.93 (+0.19) |
| + Part Embeddings | ✓ | ✓ | ✓ | ✓ | ✗ | 74.85 (+3.46) | 90.29 (+2.64) | 94.10 (+1.36) |
| | ✓ | ✓ | ✓ | ✓ | ✓ | **75.28** (+3.89) | **90.42** (+2.77) | **94.12** (+1.38) |

This suggests that our part embeddings significantly contribute to discerning fine-grained details critical for text-based person search.

**Ablation Study on Part Discovery Methods:** To validate the effectiveness of our proposed part discovery module, we conduct experiments by replacing our part discovery module with other methods such as TIPCB [2] and PAT [14]. TIPCB is a simple yet effective heuristic part discovery method, which extracts image part embeddings through equi-horizontal cropping of the entire image and transforms the global embedding of the text modality into corresponding parts by using additional learnable MLPs for each image part. PAT performs part discovery with a querying transformer approach by leveraging learnable queries and conventional cross-attention. A key difference of our method from PAT is that ours explicitly encourages the discovered parts to be spatially separated since our part slots compete with each other to aggregate input data. In contrast, parts found by PAT, based on conventional cross-attention, often over-capture salient regions and are likely to miss fine details. To demonstrate this limitation, we show the visualization of part attentions with PAT applied to ours in the supplementary materials. As indicated in Table 3, our part discovery method outperforms these heuristic and conventional cross-attention based part discovery methods across all metrics, with the most significant difference observed in the R@1 metric, which requires a precise discernment of human parts.

**Table 3:** Ablation study of different part discovery methods on the CUHK-PEDES.

| Methods | R@1 | R@5 | R@10 |
|---|---|---|---|
| Ours + TIPCB [2] | 73.23 | 89.10 | 94.04 |
| Ours + PAT [14] | 72.76 | 89.23 | 93.42 |
| Ours | **75.28** | **90.42** | **94.12** |

## 5 Conclusion

We proposed a novel framework that extracts distinctive human parts corresponding across visual and textual modalities through part discovery module without part-level supervision. The introduced TDPA further refines the retrieval process by adjusting the importance of each part slot based on the contents of the text query, leading to more precise and relevant retrieval.

**Limitation:** Since slots learn to occupy the entire image and text, some of them may indicate irrelevant part of input. TDPA addresses this issue to some extent, but a more explicit solution would further enhance our system.

## Acknowledgements

This work was supported by the IITP grants and the NRF grants funded by Ministry of Science and ICT, Korea (RS-2019-II191906; RS-2022-II220926; NRF-2018R1A5A1060031; NRF-2021R1A2C3012728)

## References

1. Aggarwal, S., Radhakrishnan, V.B., Chakraborty, A.: Text-based person search via attribute-aided matching. In: Proc. Winter Conference on Applications of Computer Vision (WACV) (2020)
2. Chen, Y., Zhang, G., Lu, Y., Wang, Z., Zheng, Y.: Tipcb: A simple but effective part-based convolutional baseline for text-based person search. Neurocomputing **494** (2022)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Annual Conference ofthe North American Chapter ofthe Association for Computational Linguistics (2019)
4. Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666 (2021)
5. Gao, C., Cai, G., Jiang, X., Zheng, F., Zhang, J., Gong, Y., Peng, P., Guo, X., Sun, X.: Contextual non-local alignment over full-scale representation for text-based person search. arXiv preprint arXiv:2101.03036 (2021)
6. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS) (2007)
7. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
8. Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., Tan, T.: Pose-guided multi-granularity attention network for text-based person search. In: Proc. AAAI Conference on Artificial Intelligence (AAAI) (2020)
9. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proc. European Conference on Computer Vision (ECCV) (2018)
10. Li, S., Cao, M., Zhang, M.: Learning semantic-aligned feature representation for text-based person search. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2022)
11. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
12. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11. Springer (2013)
13. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
14. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

15. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention (2020)
16. Niu, K., Huang, Y., Ouyang, W., Wang, L.: Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Transactions on Image Processing **29**, 5542–5556 (2020)
17. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. International Conference on Machine Learning (ICML) (2021)
19. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Adversarial representation learning for text-to-image matching. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2019)
20. Shao, Z., Zhang, X., Fang, M., Lin, Z., Wang, J., Ding, C.: Learning granularity-unified representations for text-to-image person re-identification. In: Proc. ACM Multimedia Conference (ACMMM) (2022)
21. Shu, X., Wen, W., Wu, H., Chen, K., Song, Y., Qiao, R., Ren, B., Wang, X.: See finer, see more: Implicit modality alignment for text-based person retrieval. In: Proc. European Conference on Computer Vision Workshop on Real-World Surveillance, (ECCVW) (2022)
22. Suo, W., Sun, M., Niu, K., Gao, Y., Wang, P., Zhang, Y., Wu, Q.: A simple and robust correlation filtering method for text-based person search. In: Proc. European Conference on Computer Vision (ECCV) (2022)
23. Wang, C., Luo, Z., Lin, Y., Li, S.: Text-based person search via multi-granularity embedding learning. In: Proc. International Joint Conferences on Artificial Intelligence (IJCAI) (2021)
24. Wang, Z., Fang, Z., Wang, J., Yang, Y.: Vitaa: Visual-textual attributes alignment in person search by natural language. In: Proc. European Conference on Computer Vision (ECCV) (2020)
25. Wang, Z., Zhu, A., Xue, J., Wan, X., Liu, C., Wang, T., Li, Y.: Caibc: Capturing all-round information beyond color for text-based person retrieval. In: Proc. ACM Multimedia Conference (ACMMM) (2022)
26. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
27. Wu, Y., Yan, Z., Han, X., Li, G., Zou, C., Cui, S.: Lapscore: language-guided person search via color reasoning. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2021)
28. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. arXiv preprint arXiv:1604.01850 (2016)
29. Yan, S., Dong, N., Zhang, L., Tang, J.: Clip-driven fine-grained text-image person re-identification. IEEE Transactions on Image Processing (2023)
30. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Proc. European Conference on Computer Vision (ECCV) (2018)
31. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2015)
32. Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., Hua, G.: Dssl: deep surroundings-person separation learning for text-based person retrieval. In:

Proceedings of the 29th ACM International Conference on Multimedia. pp. 209–217 (2021)