Dual Alignment: Partial Negative and Soft-label Alignment for Text-to-Image Person Retrieval

Xulin Song[a,*], Xing Jin[b], Jin Qi[a], Jun Liu[c]

[a]*School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, 210003, Jiangsu, China*
[b]*College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing, 210037, Jiangsu, China*
[c]*School of Computing and Communications, Lancaster University, Lancaster, U.K.*

Abstract

Text-to-image person retrieval is a task to retrieve the right matched images based on a given textual description of the interested person. The main challenge lies in the inherent modal difference between texts and images. Most existing works narrow the modality gap by aligning the feature representations of text and image in a latent embedding space. However, these methods usually leverage the hard label and mine insufficient or incorrect hard negatives to achieve cross-modal alignment, generating incorrect hard negative pairs so as to suboptimal performance. To tackle the above problems, we propose a dual alignment framework, Partial negative and Soft-label Alignment (PASA), which includes the partial negative alignment (PA) strategy and the Soft-label Alignment (SA) strategy. Specifically, PA pushes far away the hard negatives in the triplet loss by considering a certain amount of negatives within each mini-batch as hard negatives, preventing the distraction to the positive text-image pairs. Based on PA, SA further achieves the alignment between the similarity distribution on these hard negatives by the manner of soft-label, as well as the alignment between inter-modal and intra-modal. Extensive experiments on three public datasets, CUHK-PEDES, ICFG-PEDES and RSTPReid, demonstrate that our proposed PASA method can consistently improve the performance of text-to-image person retrieval, and achieve

*Corresponding author
Email addresses: xulinsong@njupt.edu.cn (Xulin Song), xingjin@njfu.edu.cn (Xing Jin), qijin@njupt.edu.cn (Jin Qi), j.liu81@lancaster.ac.uk (Jun Liu)

new state-of-the-art results on the above three datasets.

Keywords: Hard Negative Mining, Soft-Label Alignment, Text-to-Image Retrieval, Person Retrieval

1. Introduction

Text-to-Image Person Retrieval (TIPR) aims to retrieve the image of the interested person from a large-scale image gallery, given a query for a textual description [1, 2, 3]. It provides a complementary solution when the target person's image is not available, showing great potential in many practical applications, *e.g.*, lost persons, tracking suspects for public security [4, 5, 6, 7].

TIPR is still challenging since there are significant intra-identity variations and an inherent heterogeneity gap between the vision modality and the language modality. Most existing methods are devoted to learning discriminative feature representations and then aligning cross-modal features in a joint embedding space [8]. Generally, a textual encoder and a visual encoder are used to encode the texts and images. The greatest challenge is how to align the cross-modal data pairs. To achieve cross-modal alignment, it can be divided into global-level and local-level alignment. Global-level methods utilize vision/language backbones to extract feature representations, then design better cross-modal alignment strategies to achieve matching between texts and images in a joint embedding space [8, 9, 10, 11]. To further capture fine-grained information, local-level methods try to align textual entities with local body regions [2, 12, 13, 14]. Recently, benefiting from the power of pre-trained models, some works [15, 16, 17] utilize BERT [18], ViT [19], and CLIP [20] *et al.* to align global visual and textual features, or to mine more fine-grained local correspondence.

Among them, some existing methods adopt a contrastive learning scheme, namely treating samples corresponding to the same identity from different modalities as positives, while samples corresponding to different identities are considered as negatives. One way is to minimize similarity (similarity distribution) between positives and maximize similarity (similarity distribution) between negatives, which is known as InfoNCE loss [21] and SDM loss [22]. Another way is to push away the sample with its hardest negative by the Triplet Ranking Loss (TRL) [23], or with all negatives by the Triplet Alignment Loss (TAL) [17]. Although these strategies benefit the

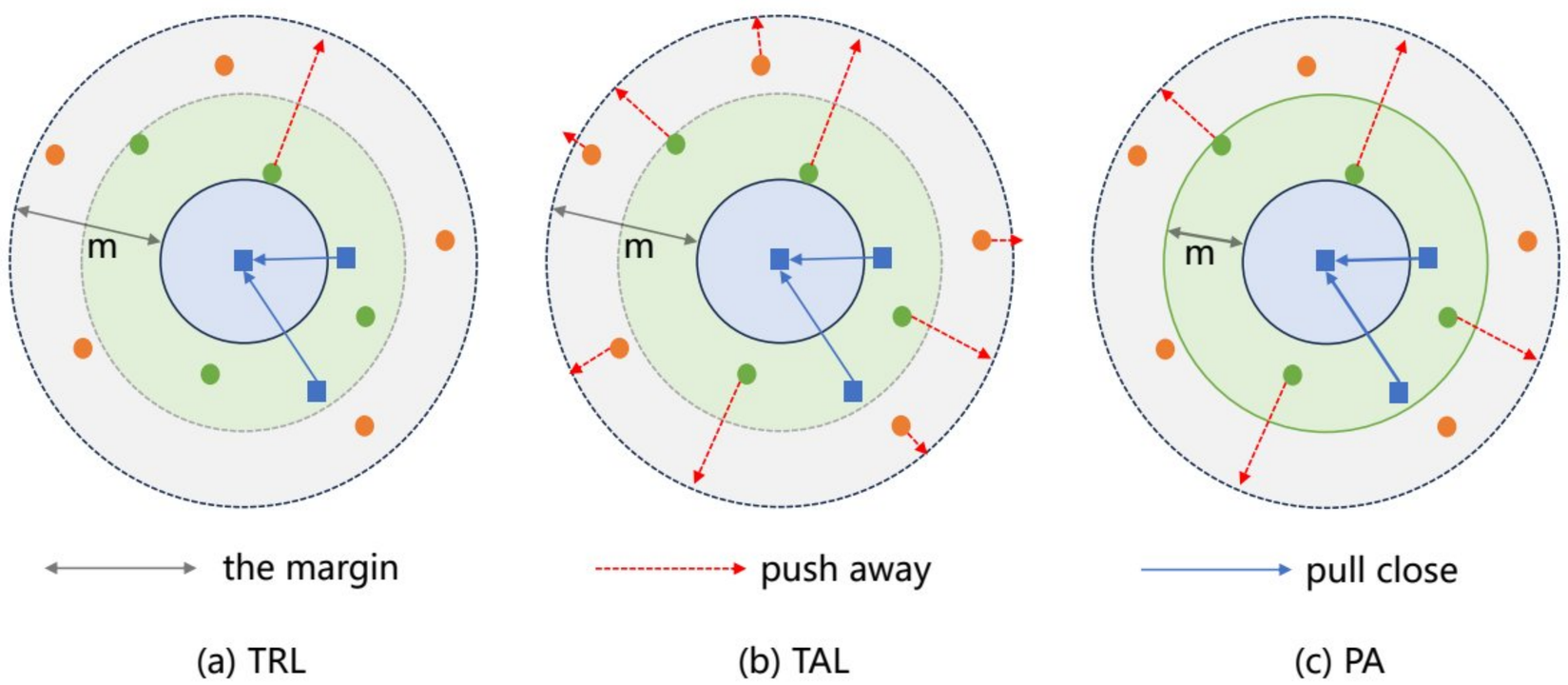

Figure 1: The comparisons between TRL, TAL and the proposed PA. The blue squares indicate the positives, while the green and orange cycles are negatives. (a) The TRL pushes away the hardest negative; (b) The TAL pushes away all the negatives; (c) The proposed PA selects a proportion of negatives as hard negatives, *i.e.*, negatives in the green regions, while the negatives in light gray regions are considered as easy negatives. (Best viewed in color.)

performance of TIPR, they have a great limitation: incorrect hard negative mining during the cross-modal alignment. As shown in Fig. 1, the TRL loss solely pushes away the hardest negative sample, while the TAL loss considers all negatives as hard negatives. However, only one hardest negative might lead to the underfitting learning problem, in which the relationships between data cannot be sufficiently learned. Meanwhile, all negatives would make the model pay much attention to easy data, since there exists a certain amount of negatives that are easy to distinguish. The above negative mining strategy, which optimizes with too many or too few negatives, will cause suboptimal performance.

To tackle this incorrect hard negative mining problem, we propose the dual alignment framework: Partial negative and Soft-label Alignment (PASA), for text-to-image person retrieval. Concretely, the PASA comprises two key modules: Partial-negative Alignment (PA) and Soft-label Alignment (SA). Specifically, Triplet PA mines and selects only a small proportion of hard negatives to push away within each mini-batch. Therefore, the model can sufficiently learn the relationships with these hard text-image pairs, avoiding excessive focusing on these easily distinguishable negative text-image pairs. Based on PA, the SA strategy further aligns the similarity distribution among

the selected hard negative text-image pairs. Motivated by the dual branch of RDE [17], PASA aligns the similarity distributions both between intra-modal and inter-modal by the SA strategy, making the consistent similarity distribution between the intra-modal and inter-modal. As a result, PASA can identify the true matched text-image pairs and better distinguish hard negative pairs simultaneously. We conduct extensive experiments on widely used public Text-to-Image person retrieval datasets, all the results demonstrate that the proposed PASA consistently improves the performance for TIPR task, and also achieves the new state-of-the-art performance. The main contributions can be summarized in the following:

(1) We propose a robust yet effective method, termed PASA, to tackle the incorrect hard negative mining problem for text-to-image person retrieval.

(2) We introduce two strategies: PA and SA. Through PA, a small proportion of hard negatives are selected to construct a triplet loss. Based on PA, SA achieves the consistent alignment distribution both for the intra-modal and inter-modal data pairs.

(3) We conduct extensive experiments on three widely used public text-to-image person retrieval benchmarks. All of the experimental results demonstrate the superiority of the proposed PASA, which can consistently improve the performance of TIPR and achieve new state-of-the-art results.

2. Related Works

2.1. Text-to-Image Person Retrieval

Text-to-image person retrieval is a relatively novel and challenging cross-modal retrieval task. Li *et al.* [1] first introduced this TIPR task by constructing the benchmark dataset CUHK-PEDES. Subsequently, more challenging benchmark datasets are introduced, ICFG-PEDES [24] and RSTPReid [25]. Existing methods can be divided into pre-training separately and pre-training with a CLIP model. The former methods treat TIPR as a complete cross-modal matching task. It first utilizes the different networks as backbones to extract text and image features, respectively [8, 11, 15, 26, 27, 28, 29, 30]. Then, cross-modal matching losses are designed to align these feature representations in the joint feature embedding space. Li *et al.* [26] utilize a CNN-LSTM network as the feature extractor and Cross-Modal Cross-Entropy (CMCE) loss is then used to align cross-modal features. The Cross-Modal Knowledge Adaptation (CMKA) [30] further leverages BERT [31] and ResNet [32] to extract different modality features. Additionally, the

Visual-Textual Attribute Alignment Model (ViTAA) [29] introduces auxiliary attribute segmentation to align the features of the human body parts and the textual attributes. Wu *et al.* [8] propose to use color reasoning information to learn the alignments between text phrases and image regions. Recently, with the success of vision-language pre-training models, the latter approaches achieve cross-modal alignment only by the CLIP model [17, 22, 33, 34]. Han *et al.* [33] first introduce the CLIP to the task of TIPR, and propose the momentum contrastive learning loss to transfer the information from generic text-image pairs to the person text-image pairs. The IRRA framework [22] leverages the CLIP to map the input text-image pairs into a joint embedding space and achieve global-level alignment based on local relation learning. RDE [17] is further proposed to utilize a triplet alignment loss to achieve global alignment with pre-trained CLIP. Although these CLIP-based models have all benefited the person retrieval performance, there exist incorrect hard negatives during the cross-modal global feature alignment which hinders the retrieval performance. In this paper, we propose to mine a certain proportion of hard negatives that are very different from TRL [23] with only the hardest negative, as well as the TAL [17] which regards all negatives as hard negatives.

2.2. Hard Negative Mining

The hard negative mining is to find these negatives that are similar to the positive while are difficult to distinguish from the positive. kalantidis *et al.* [35] proposed to mix the anchor and its negative to synthesize a new instance as the hard negative. AdaS [36] states that the negatives should be neither too hard nor too easy to distinguish from the anchor, so the adaptive sampling strategy is introduced. Xia *et al.* proposed the ProGCL [37] framework to measure the hardness of negatives. HCHSM [38] mines the hard instances through a mutual information agreement gap between negative pairs and positive pairs. In the cross-modal retrieval scenarios, TRL [23] and TAL [17] treat the dissimilar instance and all the other instances (except for positives) within a mini-batch as hard negatives. This paper proposes the partial negative mining strategy that is mostly similar to TRL and TAL. The difference is that the proposed method can be considered as the upper bound of TRL and the lower bound of TAL, which is more effective in mining hard negatives.

2.3. Alignment with Soft-label

The hard label is known as matching the paired text-image inputs in the cross-modal text-to-image retrieval scenario, which is often denoted as the pattern of one-hot label. However, it has been proven that hard labels depend excessively on the incorrect prediction results of the current model. The soft-label which is represented as the probability distribution of logits from the model, is proposed to achieve mutual guidance in the dual-branch models, such as the teacher-student model in the knowledge distillation-based methods. Li et al. [39] introduce the additional knowledge from an extra language model to achieve the cross-modal soft-label alignment. The CUSA [40] model utilizes the soft label as a supervision signal to achieve both the inter-modal and intra-modal alignment. However, existing soft-label-based alignment methods are all designed in the teacher-student scheme, our proposed PASA framework achieves the alignment with soft-label in an end-to-end CLIP model. In addition, soft-label in existing methods is generated from all instances within a mini-batch, while PASA proposes to only utilize the hard negatives to generate the soft-label which is more effective.

3. Methods

In this section, we elaborate on each component of the proposed PASA framework, which is illustrated in Fig. 2. Specifically, we first introduce the problem formulations of TIPR task in Section 3.1, as well as the feature representations in Section 3.2. Then, we present the detailed description and discussion of the PA strategy and the SA strategy in Section 3.3 and Section 3.4, respectively. Finally, we show the overall optimization of our proposed PASA in Section 3.5.

3.1. Problem Formulation

Given the text query, the aim of TIPR is to retrieve the matched images that belong to the same identity from the gallery set. Formally, the training set of a TIPR dataset can be denoted as $\{(T, V), Y\}$, where $T = \{t_i\}_{i=1}^{N}$ and $V = \{v_i\}_{i=1}^{N}$ represent text sentences and images, $\{(t_i, v_i), y_i\}$ denotes a paired text and image that belong to the same identity y_i, and $y_i \in \{1, \cdots, N_P\}$. Thus, N is the number of text-image pairs and N_P is the number of identities in the training set. As illustrated in Fig. 2, the proposed PASA utilizes the pre-trained CLIP as the text encoder and the

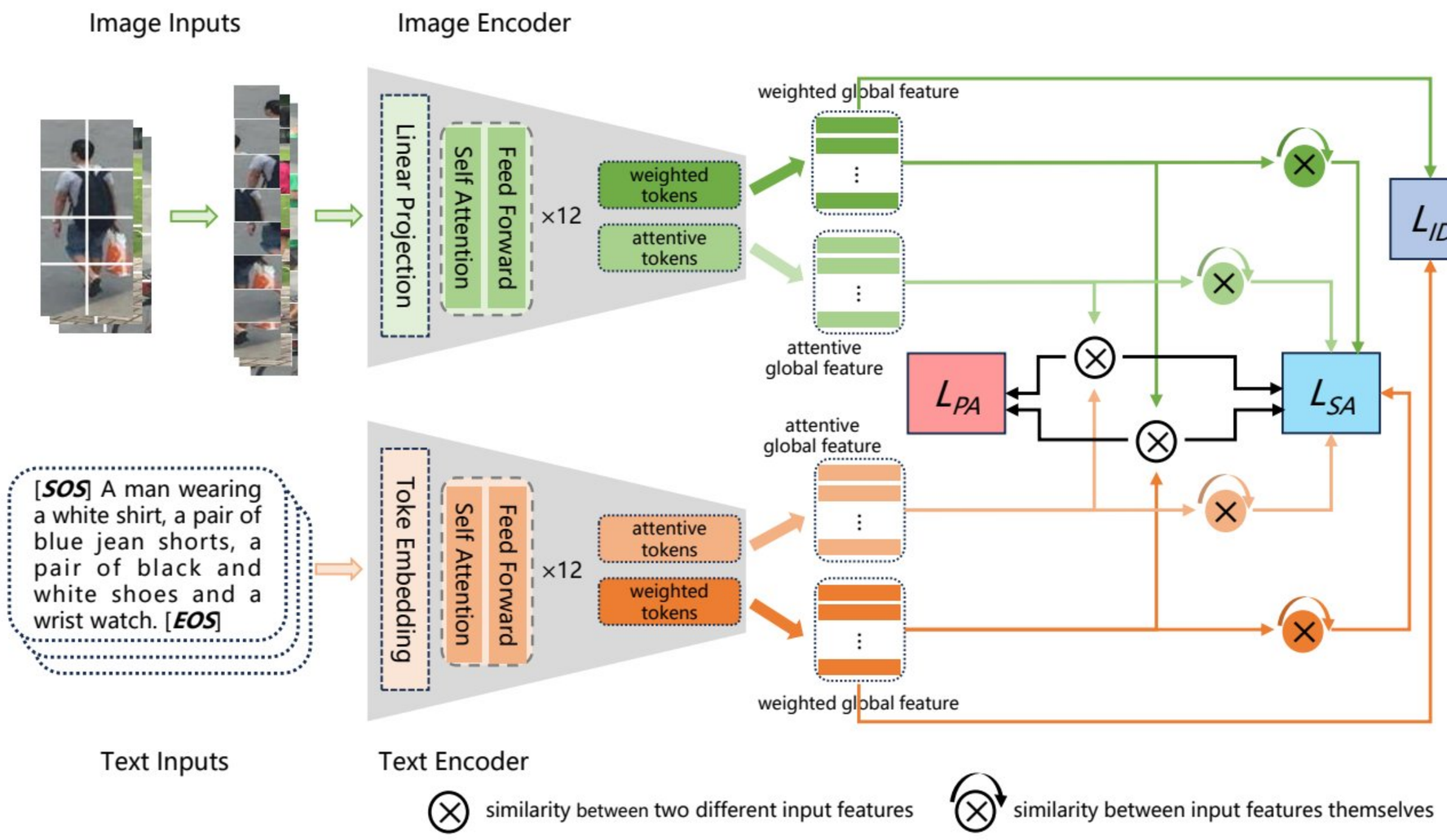

Figure 2: The framework of our proposed PASA. The input texts/images are first fed into the text/image encoder of CLIP to extract the modal-related Weighted Global Feature (WGF) and Attentive Global Feature (AGF). Then the similarity can be obtained between inter-modal and intra-modal, as well as the Partial-negative Alignment (PA) loss and the Soft-label Alignment (SA) loss. (Best viewed in color)

image encoder to map each t_i and v_i into the joint embedding space. Following the previous works IRRA [22] and RDE [17], 12-layer transformer blocks are leveraged as modality-specific encoders to extract token representations. The token representations are further fused into the weighted global feature (WGF), and the attentive global feature (AGF) for text and image, respectively. To mitigate the impact of incorrect hard negatives, we then present two strategies: Partial-negative Alignment (PA) and Soft-label Alignment (SA).

3.2. Feature Representation.

For the text input $t_i \in T$, it is firstly tokenized by the lower-cased byte pair encoding (BPE) [41] which has 49512 vocab words. The text is equipped with [SOS] token before the text and [EOS] token after the text to remind both the starting and the ending of this text sequence. Then the tokenized sequence $t_i = \{t_i^s, t_i^1, \cdots, t_i^{N_t}, t_i^e\} \in \mathbb{R}^{\left(N^t+2\right)\times d}$ is fed into the text encoder

of CLIP to extract token feature representations $F_{ti} = \{f_{ti}^s, f_{ti}^1, \cdots, f_{ti}^{N_t}, f_{ti}^e\}$, where N_t is the used token number. Similarly, given the image input $v_i \in V$ and $v_i \in \mathbb{R}^{H \times W \times C}$, it is first split into $N_v^p = H \times W/P^2$ fixed-sized non-overlapping patches, where H, W, C and P represent the height, the width, the color channels and the size of each split patch, respectively. After performing the positional embedding, the image can be tokenized as $v_i = \{v_i^{cls}, v_i^1, \cdots, v_i^{N_v}\}$, v_i^{cls} and N_v is the extra added $[CLS]$ token and the number of patches. The token feature representation for the image by the vision encoder of CLIP can be denoted as $F_{vi} = \{f_{vi}^{cls}, f_{vi}^1, \cdots, f_{vi}^{N_v}\}$.

Additionally, the weights for each modality from the self-attention map of the last Transformer blocks in CLIP reflect correlations between the $[EOS]$ token and other tokens for text, as well as the $[CLS]$ token and other tokens for image. Therefore, f_{ti}^e is the weighted combination of $\{f_{ti}^s, f_{ti}^1, \cdots, f_{ti}^{N_t}\}$, and f_{vi}^{cls} is also the weighted combination of $\{f_{vi}^1, \cdots, f_{vi}^{N_v}\}$. Following previous works [17, 34], the dual branch is utilized to extract global features for each modality, namely the weighted global feature (WGF), and the attentive global feature (AGF). Firstly, we directly treat the $[EOS]$ token as the weighted global feature $F_{ti}^w = f_{ti}^e$ for the text, as well as $F_{vi}^w = f_{vi}^{cls}$ for the image. Then the attentive global feature can be obtained by

$$F_{ti}^a = MaxPool(MLP(F_{ti}^k) + FC(F_{ti})) \ , \tag{1}$$

$$F_{vi}^a = MaxPool(MLP(F_{vi}^k) + FC(F_{vi})) \ , \tag{2}$$

where $MaxPool(\cdot)$, $MLP(\cdot)$ and $FC(\cdot)$ denote the max-pooling layer, multi-layer perceptron layer, and fully connected layer as [17]. F_{ti}^k and F_{vi}^k are the selected top k tokens that have the higher correlation weights in the self-attention map of the last Transformer blocks in CLIP as [17]. The similarity between the weighted global features $S^w(t_i, v_i) = S(F_{ti}^w, F_{vi}^w)$, as well as the attentive global features $S^a(t_i, v_i) = S(F_{ti}^a, F_{vi}^a)$, can be computed to achieve the cross-modal matching and alignment. In this paper, the cosine metric is used as similarity.

3.3. Partial-negative Alignment

In the cross-modal learning scenario, the Triplet loss mines the hard negatives and is widely used. However, the early Triplet Ranking Loss (TRL) only employs the hardest negative, while the Triplet Alignment Loss (TAL) relaxes to optimize all negatives. Although the TAL loss has achieved promising performance in the TIPR task, we think optimization of all the negatives

is not necessary since there is a certain number of negatives that are easy to distinguish. These easy negatives are also optimized to push away in the TAL which overemphasizes the easy negatives and may lead to overfitting.

3.3.1. Partial-negative Alignment Loss

In this paper, we regarded instances as hard negatives on the condition that similarities with the anchor are larger than or around similarities between the anchor's positive. To achieve this, we propose the PA strategy based on the TRL and TAL. Given a mini-batch text-image input pairs $\{T^B, V^B\} = \{\{t_i\}_{i=1}^B, \{v_i\}_{i=1}^B\}$ where B denotes the size of each mini-batch and (t_i, v_i) is a paired input from the same identity, the proposed Triplet PA loss can be represented as

$$\begin{aligned} L_{PA} = \frac{1}{B}\sum_{i=1}^{B}([m - S^p(t_i) + \tau log(S_{pa}^{hn}(t_i)/\tau)]_+ \\ + \ [m - S^p(v_i) + \tau log((S_{pa}^{hn}(v_i)/\tau)]_+) \ , \end{aligned} \tag{3}$$

where m denotes the positive margin coefficient, τ is the widely used temperature parameter to control the peaks of the probability distribution. The function $[\cdot]_+$ returns the value when it is greater than zero, otherwise returning zero. $exp\,(\cdot)$ represents the exponential function. Specifically, $S^p(t_i) = \sum_{j=1}^B \alpha_{ij} S(t_i, v_j)$ is the weighted average similarity between the text t_i and its positive images from the same identity, since there may appear multiple text-image pairs from the same identity in a mini-batch due to random sampling. Hence, the average weight can be expressed as $\alpha_{ij} = \frac{exp(S(t_i,v_j)/\tau)}{\sum_{k=1}^B exp(S(t_i,v_k)\tau)}$, t_i and v_j have the same identity, and $S\left(t_i, v_j\right) \in \{S^w\left(t_i, v_j\right), S^a\left(t_i, v_j\right)\}$. What's more, we select the hard negatives for t_i as follows

$$S_{pa}^{hn}(t_i) = \sum_{j=1}^{B} l_{ij}^{PA} exp(S(t_i, v_j)/\tau), \ \begin{cases} l_{ij}^{PA} = 1, & l_{ij}^{PA} \in topR(N_{hn}) \ , \\ l_{ij}^{PA} = 0, & otherwise \ , \end{cases} \tag{4}$$

$$N_{hn} = Sorted(q_{i1}S(t_i, v_1), q_{i2}S(t_i, v_2), \ldots, q_{iB}S(t_i, v_B)) \ , \tag{5}$$

where N_{hn} represents the sorted similarity from high to low for the text t_i with all images in a mini-batch, and $q_{ij} = 1$ when t_i and v_j are obtained from the different identity, otherwise $q_{ij} = 0$. Therefore, the selected hard negatives can be obtained by ratio $topR \in (0, 1)$ in Eq. (4) within mini-batch. Similarly, we can select the hard negatives $S_{pa}^{hn}(v_i)$ for v_i following the above solution.

Theoretically, we can prove that the proposed PA loss is the tight upper bound of the TRL, and it is also the lower bound of TAL, **i.e.**,

Lemma 1. PA is a tight upper bound of TRL than TAL, namely, PA is the upper bound with TRL, as well as the lower bound with TAL, **i.e.**, $L_{TRL} < L_{PA} < L_{TAL}$.

As $topR \to 0$, PA approaches TRL and focuses more on the hardest negative with a lower bound, while $topR \to 1$, PA approaches TAL and it relaxes the optimization to all negatives with an upper bound.

To prove Lemma 1, we first take the text-to-image direction as an example. To facilitate a more straightforward description, the Triplet Ranking Loss (TRL) [23] L_{TRL} and the Triplet Alignment Loss [17] L_{TAL} can be expressed as

$$\begin{aligned} L_{TRL} = \frac{1}{B}\sum_{i=1}^{B}([m - S^p(t_i) + \tau log(S^{hn}_{trl}(t_i)/\tau)]_+ \\ + \ [m - S^p(v_i) + \tau log((S^{hn}_{trl}(v_i)/\tau)]_+) \ , \end{aligned} \tag{6}$$

$$S^{hn}_{trl}(t_i) = \sum_{j=1}^{B} l^{TRL}_{ij} exp(S(t_i, v_j)/\tau), \quad \begin{cases} l^{TRL}_{ij} = 1, & l^{TRL}_{ij} \in N^{TRL}_{hn} \ , \\ l^{TRL}_{ij} = 0, & otherwise \ , \end{cases} \tag{7}$$

$$\begin{aligned} L_{TAL} = \frac{1}{B}\sum_{i=1}^{B}([m - S^p(t_i) + \tau log(S^{hn}_{tal}(t_i)/\tau)]_+ \\ + \ [m - S^p(v_i) + \tau log((S^{hn}_{tal}(v_i)/\tau)]_+) \ , \end{aligned} \tag{8}$$

$$S^{hn}_{tal}(t_i) = \sum_{j=1}^{B} l^{TAL}_{ij} exp(S(t_i, v_j)/\tau), \quad \begin{cases} l^{TAL}_{ij} = 1, & l^{TAL}_{ij} \in N^{TAL}_{hn} \ , \\ l^{TAL}_{ij} = 0, & otherwise \ . \end{cases} \tag{9}$$

According to Eq. (3), Eq. (6) and Eq. (8), we conclude that the only difference among TRL, TAL and PA losses is the number of the selected hard negatives. Since the number of hard negative images for text t_i is N_{hn}, TRL selects the hardest negative image, TAL selects all the hard negative images, the proposed PA loss selects partial hard negative images, namely

$$\begin{cases} N^{TRL}_{hn} = 1 \ , \\ N^{TAL}_{hn} = N_{hn} \ , \\ N^{PA}_{hn} = topR(N_{hn}) \ , \end{cases} \tag{10}$$

their relationship is $N_{hn}^{TRL} < topR(N_{hn}^{PA}) < N_{hn}^{TAL}$. Therefore, we have $S_{trl}^{hn}(t_i) < S_{pa}^{hn}(t_i) < S_{trl}^{hn}(t_i)$.

Similarly, we can prove $S_{trl}^{hn}(v_i) < S_{pa}^{hn}(v_i) < S_{trl}^{hn}(v_i)$ in the image-to-text direction. Thus, we can get $L_{TRL} < L_{PA} < L_{TAL}$, and the proof for Lemma 1 is completed.

3.3.2. Gradient Analysis of PA Loss

To provide the underlying theoretical analysis, we compute the gradients for TRL, TAL, and PA. For the sake of representation and analysis, only one direction as in RDE [17] is considered since text-to-image and image-to-text are symmetrical. Suppose there is only one paired text and image for each identity within a mini-batch, and taking the text-to-image direction as an example, the gradients generated by TRL, TAL, and PA can be simplified as

$$\frac{\partial L_{TRL}}{\partial t_i} = \bar{v_i} - v_i \ , \frac{\partial L_{TRL}}{\partial v_i} = -t_i \ , \frac{\partial L_{TRL}}{\partial \bar{v_i}} = t_i \ , \tag{11}$$

$$\frac{\partial L_{TAL}}{\partial t_i} = -v_i + \sum_{j=1}^{B} \beta v_j \ , \frac{\partial L_{TAL}}{\partial v_i} = -t_i \ , \frac{\partial L_{TAL}}{\partial v_j} = \beta t_i \ , \tag{12}$$

$$\frac{\partial L_{PA}}{\partial t_i} = -v_i + \sum_{j=1}^{B} \gamma v_j \ , \frac{\partial L_{PA}}{\partial v_i} = -t_i \ , \frac{\partial L_{PA}}{\partial v_j} = \gamma t_i \ , \tag{13}$$

where v_i, v_j, and $\bar{v_i}$ represent the positive image, negative image, and the hardest negative image corresponding to the text t_i, respectively, and $\beta = \frac{l_{ij}^{TAL} exp\left(t_i^T v_j/\tau\right)}{\sum_{k=1}^{B} l_{ij}^{TAL} exp\left(t_i^T v_k/\tau\right)}$, $\gamma = \frac{l_{ij}^{PA} exp\left(t_i^T v_j/\tau\right)}{\sum_{k=1}^{B} l_{ij}^{PA} exp\left(t_i^T v_k/\tau\right)} = \frac{exp\left(t_i^T v_j/\tau\right)}{\sum_{k\neq i}^{B} exp\left(t_i^T v_k/\tau\right)}$. Since the hardest negative is most similar to the positive, $\frac{\partial L_{TRL}}{\partial t_i}$ in Eq. (11) would easily approach 0, resulting in bad local minima early during the training. While TAL adjusts the gradients by taking all negatives into consideration, $\frac{\partial L_{TAL}}{\partial t_i}$ in Eq. (12) would keep away from 0. However, it may lead the model to pay much attention to easy data, since there exists a certain amount of negatives that are easy to distinguish, resulting in a suboptimal performance. Therefore, PA proposed to consider part of relatively harder negatives by utilizing L_{ij}^{PA} which can be obtained by $TopR$, leading to obtaining the optimal point around the selected relatively harder negatives. Thus, PA not only avoids the optimization being dominated by the hardest negative like TRL, but also

provides a more stable training process by considering relatively harder negatives than TAL which takes all negatives into consideration, leading to the performance gain.

3.4. Soft-label Aligning

We have been inspired by the work [40] that these unpaired texts and images that belong to different identities (can also be called negatives) may have a potential semantic association.

SA based on partial negative mining. Existing methods mostly emphasize the alignment of paired text and image that belong to the same identity (also called positive), and neglect the alignment between negatives which may lead to the semantically associated text and image being wrongly pushed away. To tackle this problem, we propose the Soft-label Alignment (SA) strategy, which is designed to emphasize hard negative pairs (with relatively higher similarity) and neglect these easier pairs. Specifically, SA is performed among these hard negatives based on the PA strategy since the easy negatives with lower similarity can be easily distinguished. The similarity between text t_i and image v_j can be calculated after the within-batch normalization (BN) [42] and denoted as $S^w\left(t_i, v_j\right)$ / $S^w\left(v_j, t_i\right)$ between WGF and $S^a\left(t_i, v_j\right)$ / $S^a\left(v_j, t_i\right)$ between AGF, respectively. Then the similarity probability of t_i and v_j based WGF can be expressed as

$$P_{w\_ij}^{t2v} = \frac{l_{ij} exp(S^w(t_i, v_j)/\tau)}{\sum_{k=1}^{B} l_{ik} exp(S^w(t_i, v_k)/\tau)} , \tag{14}$$

where l_{ij} is used to indicate hard negatives and can also be obtained from Eq. (4), which also indicates that SA depends on PA. In a mini-batch, the similarity probability distribution between text t_i and all other images within a mini-batch is denoted as $P_{w\_i}^{t2v} = (P_{w\_i1}^{t2v}, \cdots, P_{w\_iB}^{t2v})$.

Similarly, we can get $P_{w\_j}^{v2t} = (P_{w\_j1}^{v2t}, \cdots, P_{w\_jB}^{v2t})$ as the similarity probability distribution between image v_j and all other texts within a mini-batch, as well as $P_{a\_i}^{t2v} = (P_{a\_i1}^{t2v}, \cdots, P_{a\_iB}^{t2v})$ and $P_{a\_j}^{v2t} = (P_{a\_j1}^{v2t}, \cdots, P_{a\_jB}^{v2t})$ for AGF. By utilizing the indicator l_{ij}, the similarity probability distributions only contain these cross-modal hard negatives. Without a doubt, the hard negatives of text t_i and the hard negatives of image v_j should be consistent when t_i and v_j are the paired text-image input. That is, the similarity probability distribution P_i^{t2v} and P_j^{v2t} should be similar enough, we call this soft-label alignment-based hard negative mining. We leverage the KL divergence to

measure the similarity of two probability distributions. The soft-label alignment loss based on hard negatives is defined as

$$L_{SA}^{hn} = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}(KL(P_{w\_i}^{t2v}||P_{w\_j}^{v2t}) + KL(P_{w\_j}^{v2t}||P_{w\_i}^{t2v})) \\ +\frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}(KL(P_{a\_i}^{t2v}||P_{a\_j}^{v2t}) + KL(P_{a\_j}^{v2t}||P_{a\_i}^{t2v}))\ . \tag{15}$$

SA based on Cross-modality. Both the weighted global feature and the attentive global feature point to the same text or image, so the similarity probability distribution between weighted global features and attentive global features should also be consistent and similar enough. In addition, the inter-modal relationship is more important than the intra-modal relationship since the text-to-image person retrieval is a cross-modal task. The similarity distribution between F_{ti}^{w} and F_{tj}^{w} should be consistent with F_{ti}^{a} and F_{vj}^{a}. The probability of t_i and t_j is computed by

$$P_{w\_ij}^{t2t} = \frac{exp(S^w(t_i,t_j)/\tau)}{\sum_{k=1}^{B} exp(S^w(t_i,t_k)/\tau)}\ . \tag{16}$$

Then the probability distribution of weighted global features for t_i with other texts within a mini-batch is termed as $P_{w\_i}^{t2t} = (P_{w\_i1}^{t2t}, \cdots, P_{w\_iB}^{t2t})$. Similarly, the probability distribution of weighted global features for v_i is $P_{w\_i}^{v2v} = (P_{w\_i1}^{v2v}, \cdots, P_{w\_iB}^{v2v})$, as well as the probability distribution of attentive global features are $P_{a\_i}^{t2t} = (P_{a\_i1}^{t2t}, \cdots, P_{a\_iB}^{t2t})$ and $P_{a\_i}^{v2v} = (P_{a\_i1}^{v2v}, \cdots, P_{a\_iB}^{v2v})$, respectively. Due the redundancy of tokens, we regard $P_{a\_i}^{t2t}$ and $P_{a\_i}^{v2v}$ as target distribution in the K-L divergence, *i.e.*,

$$L_{SA}^{cm} = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}(KL(P_{w\_i}^{t2t}||P_{a\_i}^{t2t}) + KL(P_{w\_i}^{v2v}||P_{a\_i}^{v2v})) \\ +\frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}(KL(P_{w\_i}^{t2t}||P_{a\_i}^{t2i}) + KL(P_{w\_i}^{v2v}||P_{a\_i}^{v2t}))\ . \tag{17}$$

Finally ,the soft-label alignment loss is defined as

$$L_{SA} = L_{SA}^{hn} + L_{SA}^{cm}\ . \tag{18}$$

Algorithm 1 Training procedure of PASA

Require: Labeled N text-image pairs $\{(T, V), Y\}$,
pre-trained cross-modal model $\phi(\cdot; W)$,
the hyperparameters $topR$, τ, m, k.
Ensure: Well-trained cross-modal model $\hat{\phi}(\cdot; W)$

```
Initialize the backbones of the pre-trained corss-modal CLIP;
for each epoch do
    features.
    for each mini-batch do
        Extract features: WGF and AGF for each input text-image pairs;
        Compute similarities between all texts/images within mini-batch;
        Compute similarity probability distributions within mini-batch;
        Optimize the Triplet PA loss by Eq. (3).
        Optimize the SA loss by Eq. (18).
        Update and optimize φ(·; W) according to Eq. (19).
    end for
end for
```

3.5. Overall Loss

The PASA framework is trained in an end-to-end manner and the optimization procedure can be summarized in the Algorithm (1). The overall optimization objective can be expressed as :

$$L = L_{PA} + L_{SA} + L_{ID} \tag{19}$$

where L_{PA} and L_{SA} denote the hard mining loss defined in Eq. (3) and the soft-label alignment loss in Eq. (18). Specifically, L_{ID} is the commonly used ID loss [10] which leverages the cross-entropy loss that classifies each text or image into distinct groups based on their identities. In this paper, it is also utilized as the basic loss along with PA loss and SA loss.

4. Experiments

4.1. Experimental Details

We conducted extensive experiments on three public text-to-image person retrieval datasets: CUHK-PEDES [1], ICFG-PEDES [24] and RSTPReid [25]. A brief summary of the above datasets, together with the used evaluation protocols and the experimental details, is shown in the following sections.

4.1.1. Datasets and Metrics

CUHK-PEDES is the first dataset proposed for the TIPR task. It aggregates 13002 person identities, 40206 person images and each image has two textual descriptions, *i.e.* 80412 texts. As in the official data split, the training set includes the 11003 identities text-image pairs with 68108 textual descriptions corresponding to 34054 images. Both the validation set and the test set consist of 1000 identities, 6158 textual descriptions with 3078 images for validation, and the rest 6156 texts with 3078 images for testing.

ICFG-PEDES has 4102 identities with 54522 text-image pairs, each textual description corresponding to one image. Following most TIPR methods [22, 43, 17], it is only divided into the training set and the test set. There are 34674 text-image pairs for 3102 identities in the former set, as well as the remaining 19848 text-image pairs for 1000 identities in the latter set.

RSTPReid is the most challenging TIPR dataset which comprises 20505 images for 4101 identities from 15 cameras. Specifically, there are 5 images taken from different cameras for each identity, each image has 2 textual descriptions. Following the official set, the 3701 identities text-image pairs construct the training set, as well as 200 identities for the validation set, and the rest 200 identities for the test set.

Metrics. To evaluate the performance of our proposed PASA, we mainly adopt the widely-used Recall-K metric which is shortened to RK with K=1, 5, 10. Since the Rank-K metric measures whether the top k images include the first matched image corresponding to the query text. It is sensitive to the first matched image's position, and is more suitable for the condition that there is only one true-matched image corresponding to the query text in the gallery set. Therefore, the mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP) are also employed as auxiliary metrics to give a comprehensive evaluation.

4.1.2. Implementation Details

The proposed PASA model is trained and tested on a single NVIDIA RTX4090 24G GPU. To be fair, the pre-trained CLIP, the CLIP text Transformer and the CLIP-ViTB/16, is adopted as the text/image encoder as IRRA [22] and RDE [17]. During training, we adopt different data augmentation strategies to increase the diversity of text and image training data. For the input texts, the random masking/placement/removing of each word token is used as data augmentation, while the data augmentation for each input image consists of the random horizontal flipping, random crop with

Table 1: Performance comparison of the proposed PASA with these SOTA methods on the dataset CUHK-PEDES. The "Ref." column shows the source of the methods. The "T/I Enc." column is the text/image encoder of each referenced method. RN50 is short for the ResNet50 neural network. * For a fair comparison, we only state the results of the global matching of CADA [44].

Methods	Ref.	T/I Enc.	R1	R5	R10	MAP	mINP
GNA-RNN[1]	CVPR17	VGG16	19.05	-	53.64	-	-
CMPM/C[9]	ECCV18	LSTM/RN50	49.37	-	79.21	-	-
Dual-Path[45]	ACMMM20	RN50/RN50	44.40	66.26	76.07	-	-
ViTAA[29]	ECCV20	LSTM/Resnet50	54.92	75.18	82.90	51.60	-
DSSL[25]	TMM21	BERT/RN50	59.98	80.41	87.56	-	-
SSAN[24]	Arxiv21	LSTM/RN50	61.37	80.15	86.73	-	-
Lapscore[8]	ICCV21	BERT/RN50	63.4	-	87.8	-	-
AXM-Net[46]	AAAI22	BERT/RN50	64.44	80.52	86.77	58.73	-
PBSL[47]	ACMMM23	BERT/RN50	65.32	83.81	89.26	-	-
BEAT[48]	ACMMM23	BERT/RN101	65.61	83.45	89.54	-	-
LCR^{2}S[49]	ACMMM23	TextCNN/RN50	67.36	84.19	89.62	59.24	-
CFine[34]	TIP23	BERT/ViT	69.57	85.93	91.15	-	-
IRRA[22]	CVPR23	CLIP	73.38	89.93	93.71	66.13	50.24
DCEL[50]	ACMMM23	CLIP	75.02	90.89	94.52	-	-
Rasa[43]	IJCAI23	ALBEF	76.51	90.29	94.25	**69.38**	-
APTM[51]	ACMMM23	BERT/Swim-B	76.53	90.04	94.15	66.91	-
VGSG[52]	TIP24	CLIP/RN50	71.38	86.75	91.86	67.91	-
CADA*[44]	TMM24	BERT/ViT	73.48	89.57	94.10	65.82	-
TPBS[53]	AAAI24	CLIP	73.54	88.19	92.35	65.38	-
RDE[17]	CVPR24	CLIP	75.94	90.14	94.12	67.56	51.44
PASA	Ours	CLIP	**76.58**	**90.81**	**94.67**	67.93	**51.56**

padding and random erasing. Specifically, the maximum length of each input textual token sequence is 77, and the size of each input image is 384×128. We adopt the AdamW optimizer for a total of 60 epochs with a batch size of 128. The initial learning rate is set to $1e-5$ together with the cosine learning rate decay. In addition, the ratio of the hard negatives $topR = 0.1$ for the datasets CUHK-PEDES, and $topR = 0.2$ for ICFG-PEDES and RSTPReid datasets. Finally, the temperature τ is set to 0.02, the margin $m = 0.05$ in Eq. (3) for loss L_{PA}, the attentive token selection ratio $k = 0.5$ for the

Table 2: Performance comparison of the proposed PASA with these SOTA methods on the dataset ICFG-PEDES.

Methods	Ref	R1	R5	R10	MAP	mINP
CMPM+CMPC[9]	ECCV18	43.51	65.44	74.26	-	-
Dual-Path[45]	ACMTMM20	38.99	59.44	68.41	-	-
ViTAA[29]	ECCV20	50.98	68.79	75.78	-	-
SSAN[24]	Arxiv21	54.23	72.63	79.53	-	-
CFine[34]	TIP23	60.83	76.55	82.42	-	-
IRRA[22]	CVPR23	63.46	80.25	85.82	38.06	7.93
Rasa[43]	IJCAI23	65.28	80.04	85.12	**41.29**	-
APTM[51]	ACMMM23	68.51	82.99	87.56	41.22	-
TPBS-CLIP[53]	AAAI24	65.05	80.34	85.47	39.83	-
RDE[17]	CVPR24	67.68	82.47	87.36	40.06	7.87
PASA	Ours	**67.89**	**82.52**	**87.43**	41.11	**8.26**

attentive global feature.

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

In this section, we display the comparison results between the proposed PASA method with the existing SOTA approaches on three public datasets.

Comparison results on CUHK-PEDES. We first evaluate the performance on the most widely used dataset, *i.e.*, CUHK-PEDES, and the comparison results are shown in Table 1. From Table 1, we can obviously conclude that PASA outperforms the displayed state-of-the-art methods, achieving the new SOTA performance on all the proposed metrics except for R10, where our PASA is still the second best. Specifically, similar to the proposed PASA method, the IRRA [22] and the RDE [17] methods are all global-based cross-modal alignment methods, and RDE is also the method for negative mining. PASA can surpass IRRA [22] by 3.2% on R1 and 1.8% on mAP accuracy, respectively. Meanwhile, PASA consistently outperforms the recent SOTA method RDE by 0.64% on R1, 0.27% on mAP, demonstrating the advantages and effectiveness of our proposed method.

Comparison results on ICFG-PEDES and RSTPReid. We report the experimental results on the challenging dataset ICFG-PEDES in Table 2, as well as the RSTPReid dataset in Table 3. On the ICFG-PEDES dataset, the proposed PASA can achieve the performance with R1=**67.89%**,

Table 3: Performance comparison of the proposed PASA with these SOTA methods on the dataset RSTPReid.

Methods	Ref	R1	R5	R10	MAP	mINP
DSSL[25]	TMM21	39.05	62.60	73.95	-	-
SSAN[24]	Arxiv21	43.50	67.80	77.15	-	-
LCR^{2}S[49]	ACMMM23	54.95	76.65	84.70	40.92	-
CFine[34]	TIP23	50.55	72.50	81.60	-	-
IRRA[22]	CVPR23	60.20	81.30	88.20	47.17	25.28
Rasa[43]	IJCAI23	66.90	86.50	91.35	**52.31**	-
TPBS-CLIP[53]	AAAI24	61.95	83.55	88.75	48.26	-
RDE[17]	CVPR24	65.35	83.95	89.90	50.88	28.08
PASA	Ours	**66.1**	**85.3**	**91.55**	51.31	**29.17**

Table 4: Ablation studies of the proposed PASA on the dataset CUHK-PEDES.

	L_{ID}	L_{SA}^{cm}	L_{SA}^{hn}	L_{PA}	R1	R5	R10	mAP	mINP
backbone	✓				65.33	84.05	90.33	59.15	43.1
	✓			✓	75.83	90.46	94.53	67.66	51.33
	✓	✓		✓	76.15	90.19	94.13	67.36	51.11
	✓		✓	✓	76.32	90.33	94.22	67.46	51.23
	✓	✓			0.19	0.80	1.38	0.43	0.21
	✓		✓		7.02	14.81	20.24	6.63	2.87
	✓	✓	✓		73.47	87.98	92.64	65.12	48.76
PASA	✓	✓	✓	✓	**76.58**	**90.81**	**94.67**	**67.93**	**51.56**

R5=**82.52%**, R10=**87.43%**, mAP=41.11% and mINP=**8.26%**, it achieves a new SOTA performance. Our method surpass RDE with rise of 0.21% on R1 accuracy and 0.39% on mINP, respectively. Specifically, PASA obtains the performance gain on mAP by a large margin of 1.05%. As shown in Table (3), PASA further achieves considerable performance gains by 0.75%, 0.43% and 1.09% on R1 accuracy, mAp and mINP, respectively.

In summary, our PASA consistently achieves SOTA performance on all three datasets, which indicates the effectiveness of the proposed PASA method.

4.3. Ablation Studies

To investigate the effectiveness, we conduct ablation studies on the CUHK-PEDES dataset to compare the contributions of the Partial-hard-negative-Alignment strategy and the Soft-label Alignment strategy in PASA method. As shown in Table 4, we take the L_{ID} loss as the backbone framework. Subsequently, different losses are added to the backbone to test the contribution of the PA and SA strategy. It is worth noting that the SA strategy includes L_{SA}^{cm} and L_{SA}^{hn}. All the experimental results could give the following observations: (1) Only the L_{SA}^{cm} or the L_{SA}^{hn} loss seems not to work, while any one of the L_{SA}^{cm} or the L_{SA}^{hn} loss combined with the PA loss can make an improvement on the performance. This also indicates that SA loss depends on PA loss. (2) By combining PA and SA, PASA can achieve the best performance on all metrics, demonstrating the effectiveness and complementarity of both components, which also further certifies the superiority of the proposed Triplet PA loss.

4.4. Hyperparameter sensitivity analysis

To examine the sensitivity of our proposed PASA to these hyperparameters, we conducted experiments on the CUHK-PEDES dataset. For **the ratio of hard negatives** $topR$, we set it to vary within a certain range (0,

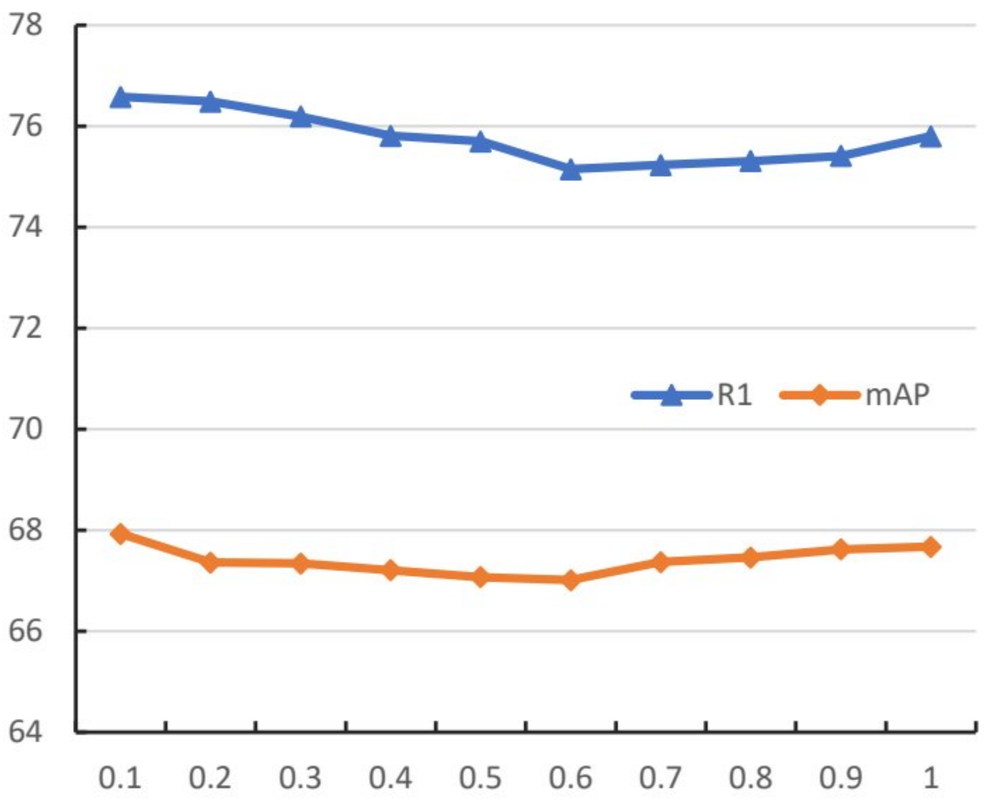

Figure 3: The impact on the performance of hyperparameter topR (ratio of hard negatives) on the dataset CUHK-PEDES. (Best viewed in color)

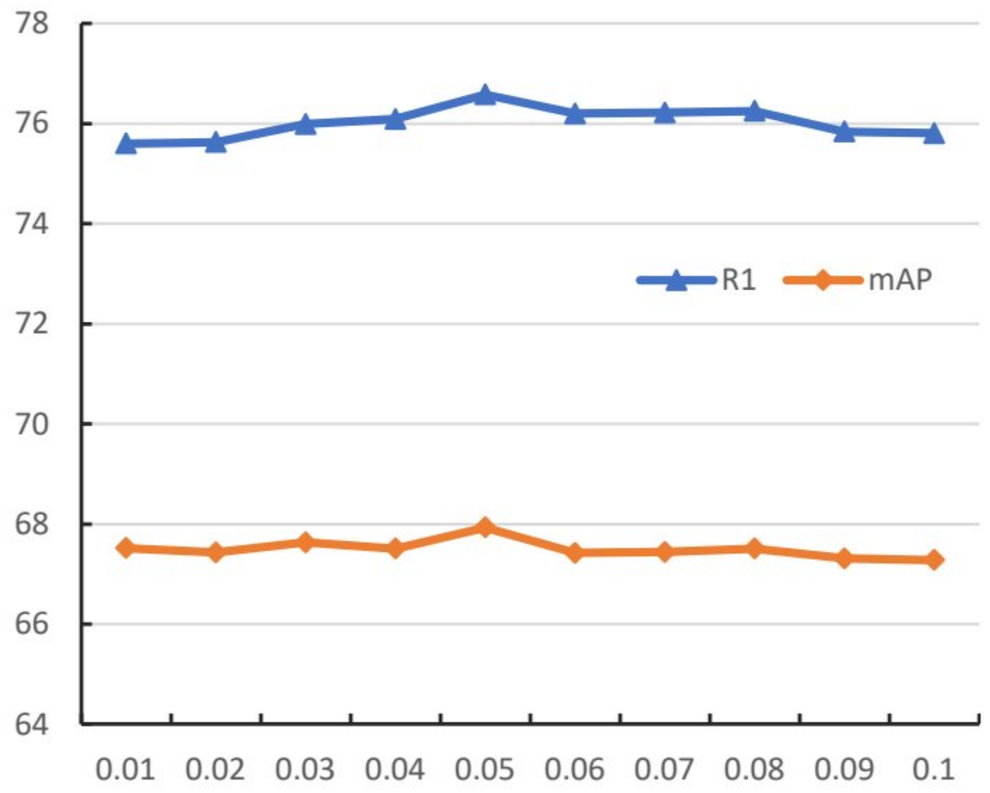

Figure 4: The impact on the performance of hyperparameter m (margin) on the dataset CUHK-PEDES. (Best viewed in color)

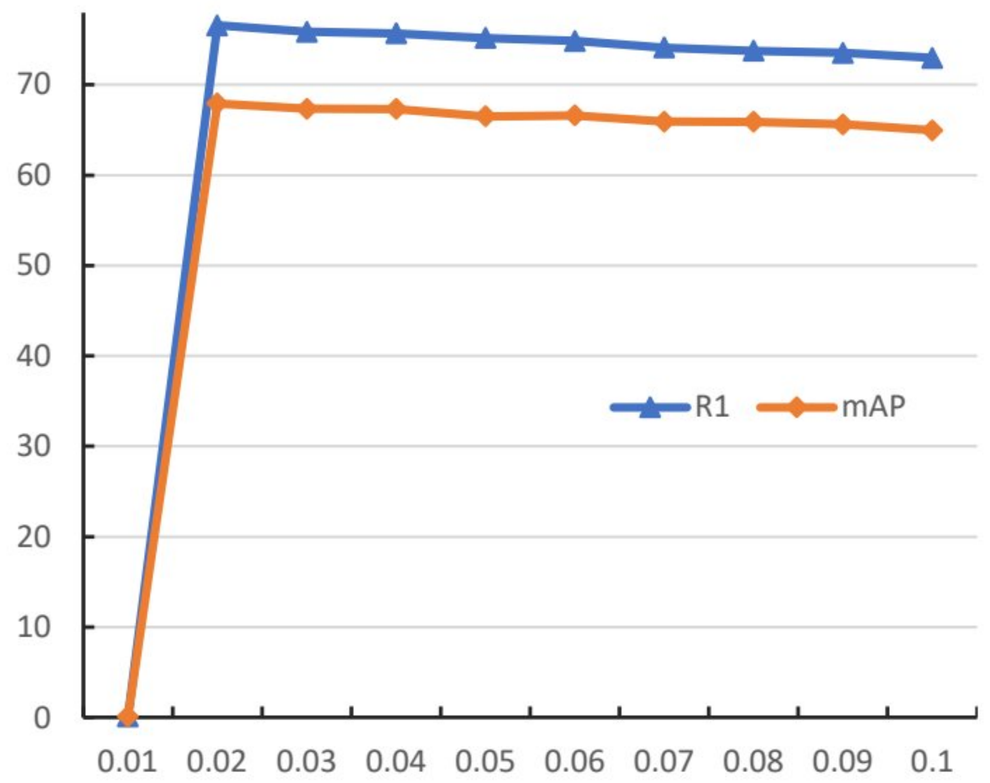

Figure 5: The impact on the performance of hyperparameter τ (temperature) on the dataset CUHK-PEDES. (Best viewed in color)

1) together with an interval 0.1. As shown in Fig. 3, the accuracy of Rank-1 and mAP initially decreases and then increases, the best performance can be obtained when $topR$ is set to 0.1. That is, the hard negatives only exist between instances within a relatively smaller ratio of the batch size, the remaining instances can be regarded as negatives that are easy to distinguish.

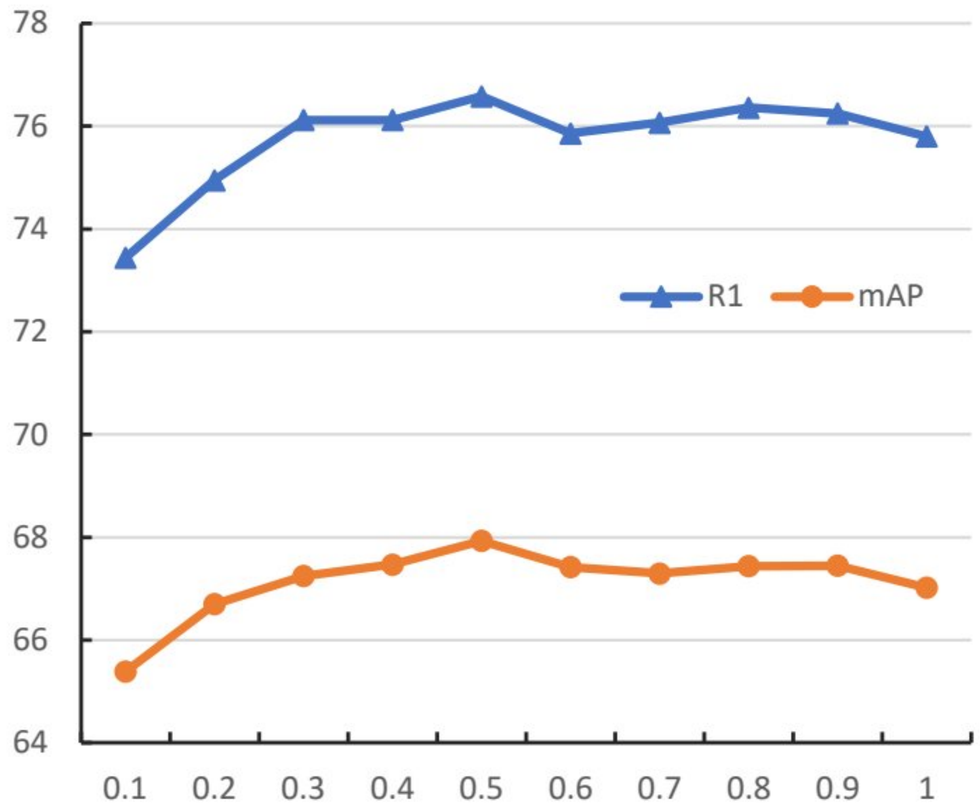

Figure 6: The impact on the performance of hyperparameter k (attention token selection ratio) on the dataset CUHK-PEDES. (Best viewed in color)

The margin in loss L_{PA} is a hyperparameter to decide the difference in similarity between positives and negatives. From Fig. 4, we can see that the best performance can be obtained with $m = 0.05$. The increase or decrease in the value of m leads to suboptimal performance on Rank-1 and mAP. Specifically, the results of Rank-1 and mAP on **temperature** τ are shown in Fig. 5. We can obviously find that a too small value gives the failed training, *i.e.* $\tau = 0.01$, while the increasing value of τ decreases both the Rank-1 and mAp accuracy. During the extensive experiments, the best performance can be achieved when $\tau = 0.02$. To sum up, we choose $topR = 0.1$, $m = 0.05$ and $\tau = 0.02$ in all experiments on the CUHK-PEDES dataset. The impact of the attention token selection ratio k on the performance is presented in Fig. 6. Through the result changes on Rank1 and mAP, it can be obviously seen that the Rank1 and MAP accuracy first rise and then drop as the value of k increases, and the best performance can be obtained with k=0.5. It means that a small ratio of local tokens may lose important information, while a large ratio of local tokens leads to overfitting and decreases the performance to a certain extent.

4.5. Generalization Analysis

To test the generalization ability of the proposed PASA, we conducted experiments under noise. In view of fairness, we leverage the noisy training

Table 5: Performance comparison under different noise rates. CUHK represents the dataset CUHK-PEDES, ICFG and RSTP are short for the datasets ICFG-PEDES and RSTPReid, respectively.

Noise		0%			20%			50%		
Methods		PASA	RDE	IRRA	PASA	RDE	IRRA	PASA	RDE	IRRA
CUHK	R1	**76.58**	75.94	73.38	**75.06**	74.46	69.44	**72.24**	71.33	62.41
	R5	**90.81**	90.14	89.93	**89.85**	89.42	87.09	**87.78**	87.41	82.23
	R10	**94.67**	94.12	93.71	**94.02**	93.63	92.04	**92.35**	91.81	88.40
	mAP	**67.93**	67.56	66.13	**66.7**	66.13	62.16	**64.47**	63.50	55.52
	mINP	**51.56**	51.44	50.24	**50.24**	49.66	45.70	**48.39**	47.36	38.48
ICFG	R1	**67.89**	67.68	63.46	**66.68**	66.54	60.76	**64.26**	63.76	52.53
	R5	**82.52**	82.47	80.25	**82.35**	81.70	78.26	**79.98**	79.53	71.99
	R10	**87.43**	87.36	85.82	**87.16**	86.70	84.01	**85.17**	84.91	79.41
	mAP	**41.11**	40.06	38.06	**39.77**	39.08	35.87	**38.63**	37.38	29.05
	mINP	**8.26**	7.87	7.93	**7.98**	7.55	6.80	**7.74**	6.80	4.43
RSTP	R1	**66.1**	65.35	60.20	**65.1**	64.45	58.75	**63.45**	62.85	56.65
	R5	**85.3**	83.96	81.30	**84.45**	83.50	81.90	**84.35**	83.20	78.40
	R10	**91.55**	89.90	88.20	**90.55**	90.00	88.25	**90.35**	89.15	86.55
	mAP	**51.31**	50.88	47.17	**50.5**	49.78	46.38	**48.76**	47.67	42.41
	mINP	**29.17**	28.08	25.28	**28.19**	27.43	24.78	**26.23**	23.97	21.05

data as generated in RDE [17]. That is, the 20% and 50% synthetic noises are used to simulate these insufficiently aligned text-image pairs in a real-world scenario. The experimental results under different noise rates are shown in Table 5. From Table 5, we can see that the proposed PASA has strong generalization ability, since its performance can outperform the recent similar methods RDE[17] and IRRA [22] on all metrics under different noise settings, demonstrating the strong anti-interference ability and generalization ability of the proposed PASA.

4.6. Visualization of Retrieval Results

To visually compare the retrieval performance, we exhibit the top-10 retrieval examples of our proposed PASA method and the recent methods RDE [17] and IRRA [22]. As shown in Fig. 7, we show two texts as the query, and its top-10 retrieved images from our PASA are more accurate. Specifically, the matched images with the query texts from PASA rank in front of the top-10 ranking results. For example, for query 1, the top-3 results

Figure 7: The examples of top-10 retrieval results from two query texts on the CUHK-PEDES dataset from our PASA, RDE [17] and IRRA [22]. For each query text, the first row is the top-10 retrieved images for our PASA, while the second row and the third row are the results from RDE and IRRA. The positive matched images and negative mismatched images are marked by green and red rectangles, respectively. (Best viewed in color)

of the proposed PASA are all the matched positive images, while there are no matched positive images for RDE, and the only matched positive image ranks second for IRRA. From the mismatch negative images, we can also conclude that PASA has more matched parts with the query text than RDE and IRRA, indicating the reliability and robustness of our proposed method.

5. Conclusion

In this paper, we have proposed a dual alignment method for text-to-image person retrieval, *i.e.* the PASA method, which includes Partial negative Alignment and Soft-label Alignment. Our method first utilizes the CLIP model to extract the weighted global feature (WGF) and the attentive global feature (AGF) for each input text-image pair. Then the dual alignment mechanism is performed on WGF and AGF to calculate the similarities between inter-modal and intra-modal. By using Partial Alignment, we then achieve the alignment for these hard negatives within each mini-batch. With Soft-label Alignment, the alignment of similarity distribution for inter-modal and intra-modal can be achieved, especially for the inter-modal hard negatives. Finally, we conducted extensive experiments on widely-used text-to-image person retrieval datasets, *i.e.*, CUHK-PEDES, ICFG-PEDES and RSTPReid. All experimental results demonstrated that our proposed PASA method consistently improves performance in all metrics and achieves the new state-of-the-art results.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62172235.

References

[1] S. Li, T. Xiao, H. Li, B. Zhou, X. Wang, Person Search with Natural Language Description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1970–1979.

[2] C. Wang, Z. Luo, Y. Lin, S. Li, Text-based Person Search via Multi-Granularity Embedding Learning, in: International Joint Conference on Artificial Intelligence (IJCAI), 2021, pp. 1068–1074.

[3] T. Gong, J. Wang, L. Zhang, Cross-modal Semantic Aligning and Neighbor-aware Completing for Robust Text–image Person Retrieval, Information Fusion 112 (2024) 102544.

[4] Z. Wang, R. Hu, Y. Yu, C. Liang, W. Huang, Multi-Level Fusion for Person Re-identification with Incomplete Marks, in: the 23rd ACM international conference on Multimedia, 2015, pp. 1267–1270.

[5] C. E. B. Ham, Learning disentangled representation for robust person re-identification, in: Advances in Neural nformation Processing Systems (NIPS), 2019.

[6] N. Huang, J. Liu, Y. Miao, Q. Zhang, J. Han, Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review, Information Fusion 91 (2023) 396–411.

[7] N. Huang, B. Xing, Q. Zhang, J. Han, J. Huang, Co-segmentation assisted cross-modality person re-identification, Information Fusion 104 (2024) 102194.

[8] Y. Wu, Z. Yan, X. Han, G. Li, C. Zou, S. Cui, LapsCore: Language-Guided Person Search via Color Reasoning, in: the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1624–1633.

[9] Y. Zhang, H. Lu, Deep Cross-modal Projection Learning for Image-text Matching, in: the European conference on computer vision (ECCV), 2018, pp. 707–723.

[10] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y. Shen, Dual-path Convolutional Image-Text Embeddings with Instance Loss, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16 (2020) 1–23.

[11] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, See finer, see more:Implicit modality alignment for text-based person retrieval, in: the European conference on computer vision (ECCV), 2022, pp. 624–641.

[12] Y. Jing, C. Si, J. Wang, W. Wang, T. Tan, Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search, in: the AAAI Conference on Artificial Intelligence, 2020, pp. 11189–11196.

[13] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving Description-Based Person Re-Identification by Multi-Granularity Image-Text Alignments, IEEE Transactions on Image Processing (TIP) 29 (2020) 5542–5556.

[14] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, C. Ding, Learning Granularity-Unified Representations for Text-to-Image Person Reidentification, in: the 30th ACM International Conference on Multimedia, 2022.

[15] S. Yan, H. Tang, L. Zhang, J. Tang, Image-Specific Information Suppression and Implicit Local Alignment for Text-Based Person Search, IEEE Transactions on Neural Networks and Learning Systems (TNNLS) (2023) 1–14.

[16] D. Jiang, M. Ye, Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2787–2797.

[17] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, P. Hu, Noisy-Correspondence Learning for Text-to-Image Person Re-Identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27197–27206.

[18] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv abs/1810.04805 (2018). URL: `http://arxiv.org/abs/1810.04805`.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: The International Conference on Learning Representations (ICLR), 2021.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, Learning Transferable Visual Models From Natural Language Supervision, in: International conference on machine learning (ICML), 2021, pp. 8748–8763.

[21] A. V. D. Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, arXiv:1807.03748 (2018).

[22] D. Jiang, M. Ye, 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.

[23] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, VSE++: Improving Visual-Semantic Embeddings with Hard Negatives, arXiv:1707.05612 (2017).

[24] Z. Ding, C. Ding, Z. Shao, D. Tao, Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification, arXiv preprint arXiv:2107.12666 (2021).

[25] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, G. Hua, DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval, in: the ACM International Conference on Multimedia (ACM MM), 2021.

[26] S. Li, T. Xiao, H. Li, W. Yang, X. Wang, Identity-Aware Textual-Visual Matching with Latent Co-attention, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1890–1899.

[27] Y. Zhang, H. Lu, Deep Cross-Modal Projection Learning for Image-Text Matching, in: the European conference on computer vision (ECCV), 2018.

[28] N. Sarafianos, X. Xu, I. Kakadiaris, Adversarial Representation Learning for Text-to-Image Matching, the IEEE/CVF International Conference on Computer Vision (ICCV) (2019).

[29] Z. Wang, Z. Fang, J. Wang, Y. Yang, ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language, in: European Conference on Computer Vision (ECCV), 2020, pp. 402–420.

[30] Y. Chen, R. Huang, H. Chang, C. Tan, B. Ma, Cross-Modal Knowledge Adaptation for Language-Based Person Search, IEEE Transactions on Image Processing (TIP) PP (2021) 4057–4069.

[31] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:2107.12666 (2018).

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[33] X. Han, S. He, L. Zhang, T. Xiang, Text-Based Person Search with Limited Data, arXiv preprint arXiv:2110.10807 (2021).

[34] S. Yan, N. Dong, L. Zhang, J. Tang, Clip-driven fine-grained text-image person re-identification, IEEE Transactions on Image Processing 32 (2023) 6032–6046.

[35] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus, Hard Negative Mixing for Contrastive Learning, in: Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 21798–21809.

[36] S. Wan, Y. Zhan, S. Chen, S. Pan, J. Yang, D. Tao, C. Gong, Boosting Graph Contrastive Learning via Adaptive Sampling, IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 35 (2024) 15971–15983.

[37] J. Xia, L. Wu, G. Wang, J. Chen, S. Z. Li, ProGCL: Rethinking Hard Negative Mining in Graph Contrastive Learning, in: International Conference on Machine Learning (ICML), 2022, pp. 24332–24346.

[38] Y. Liu, X. Yang, S. Zhou, X. Liu, Z. Wang, K. Liang, W. Tu, L. Li, J. Duan, C. Chen, Hard Sample Aware Network for Contrastive Deep Graph Clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2023.

[39] Z. Li, C. Guo, Z. Feng, J.-N. Hwang, Z. Du, Integrating Language Guidance Into Image-Text Matching for Correcting False Negatives, IEEE Transactions on Multimedia (TMM) 26 (2024) 103–116.

[40] H. Huang, Z. Nie, Z. Wang, Z. Shang, Cross-Modal and Uni-Modal Soft-Label Alignment for Image-Text Retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), volume 38, 2024, pp. 18298–18306.

[41] R. Sennrich, B. Haddow, A. Birch, Neural Machine Translation of Rare Words with Subword Units, arXiv preprint arXiv:1508.07909 (2015).

[42] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: Proceedings of the International Conference on Machine Learning (ICML), volume 37, 2015, pp. 448–456.

[43] Y. Bai, M. Cao, D. Gao, Z. Cao, C. Chen, Z. Fan, L. Nie, M. Zhang, RaSa: Relation and Sensitivity Aware Representation Learning for Text-based Person Search, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI), 2023, pp. 555–563.

[44] D. Lin, Y.-X. Peng, J. Meng, W.-S. Zheng, Cross-Modal Adaptive Dual Association for Text-to-Image Person Retrieval, IEEE Transactions on Multimedia (TMM) 26 (2024) 6609–6620.

[45] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, Y. Shen, Dual-Path Convolutional Image-Text Embedding, ACM Transactions on Multimedia Computing, Communications, and Applications (2020) 1–23.

[46] A. Farooq, M. Awais, J. Kittler, S. S. Khalid, AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), volume 36, 2022, pp. 4477–4485.

[47] F. Shen, X. Shu, X. Du, J. Tang, Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval, in: Proceedings of the 31st ACM International Conference on Multimedia (ACMMM), 2023, pp. 8922–8931.

[48] Y. Ma, X. Sun, J. Ji, G. Jiang, W. Zhuang, R. Ji, Beat: Bi-directional One-to-Many Embedding Alignment for Text-based Person Retrieval, in: Proceedings of the 31st ACM International Conference on Multimedia (ACMMM), 2023, pp. 4157–4168.

[49] S. Yan, N. Dong, J. Liu, L. Zhang, J. Tang, Learning Comprehensive Representations with Richer Self for Text-to-Image Person Re-Identification, in: Proceedings of the 31st ACM International Conference on Multimedia (ACMMM), 2023, pp. 6202–6211.

[50] S. Li, X. Xu, Y. Yang, F. Shen, Y. Mo, Y. Li, H. T. Shen, Dcel: Deep Cross-modal Evidential Learning for Text-based Person Retrieval, in:

Proceedings of the 31st ACM International Conference on Multimedia (ACMMM), 2023, pp. 6292–6300.

[51] S. Yang, Y. Zhou, Y. Wang, Y. Wu, L. Zhu, Z. Zheng, Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark, in: Proceedings of the 2023 ACM on Multimedia Conference (ACMMM), 2023.

[52] S. He, H. Luo, W. Jiang, X. Jiang, H. Ding, VGSG: Vision-Guided Semantic-Group Network for Text-Based Person Search, IEEE Transactions on Image Processing (TIP) 33 (2024) 163–176.

[53] M. Cao, Y. Bai, Z. Zeng, M. Ye, M. Zhang, An Empirical Study of CLIP for Text-Based Person Search, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), volume 38, 2024, pp. 465–473.