RaSa: Relation and Sensitivity Aware Representation Learning for Text-based Person Search

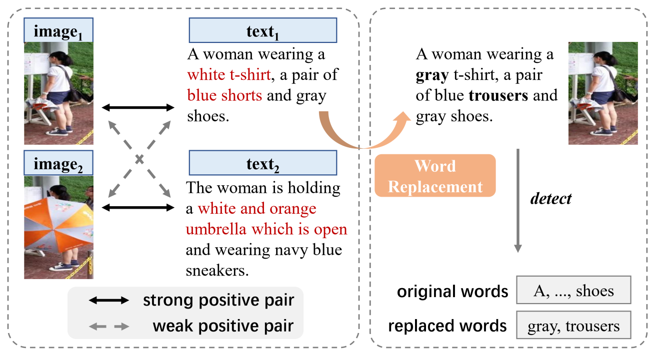
Yang Bai[1] , **Min Cao**[1∗] , **Daming Gao**[1] , **Ziqiang Cao**[1] , **Chen Chen**[2] ,
Zhenfeng Fan[3] , **Liqiang Nie**[4] and **Min Zhang**[1,4]
[1]School of Computer Science and Technology, Soochow University
[2]Institute of Automation, Chinese Academy of Sciences
[3]Institute of Computing Technology, Chinese Academy of Sciences
[4]Harbin Institute of Technology, Shenzhen
ybaibyougert@stu.suda.edu.cn, mcao@suda.edu.cn

Abstract

Text-based person search aims to retrieve the specified person images given a textual description. The key to tackling such a challenging task is to learn powerful multi-modal representations. Towards this, we propose a Relation and Sensitivity aware representation learning method (RaSa), including two novel tasks: Relation-Aware learning (RA) and Sensitivity-Aware learning (SA). For one thing, existing methods cluster representations of all positive pairs without distinction and overlook the noise problem caused by the weak positive pairs where the text and the paired image have noise correspondences, thus leading to overfitting learning. RA offsets the overfitting risk by introducing a novel positive relation detection task (*i.e.*, learning to distinguish strong and weak positive pairs). For another thing, learning invariant representation under data augmentation (*i.e.*, being insensitive to some transformations) is a general practice for improving representation's robustness in existing methods. Beyond that, we encourage the representation to perceive the sensitive transformation by SA (*i.e.*, learning to detect the replaced words), thus promoting the representation's robustness. Experiments demonstrate that RaSa outperforms existing state-of-the-art methods by **6.94%**, **4.45%** and **15.35%** in terms of Rank@1 on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets, respectively. Code is available at: https://github.com/Flame-Chasers/RaSa.

1 Introduction

Text-based person search [Li *et al.*, 2017; Wang *et al.*, 2021] aims at retrieving the person images in a large-scale person image pool given a query of textual description about that person. This task is related to person re-identification [Ji *et al.*, 2021; Wang *et al.*, 2022a] and text-image retrieval [Cao *et al.*, 2022; Li *et al.*, 2021], which have been very active research topics in recent years. It, however, exhibits unique characteristics and challenges. Compared to person re-identification with image queries, text-based person search with more accessible open-form text queries provides a more user-friendly searching procedure while embracing greater challenges due to the cross-modal search. In addition, compared to general image-text retrieval, text-based person search focuses on cross-modal retrieval specific for the person with more fine-grained details, tending to larger intra-class variance as well as smaller inter-class variance, which toughly bottlenecks the retrieval performance.



(a) Two types of positive relations **(b) Replaced token detection**

Figure 1: Illustration of (a) two types of positive relations for relation-aware learning, where the noise interference in the weak positive pairs is highlighted in red, (b) replaced token detection for sensitivity-aware learning, in which word replacement is used as the sensitive transformation and the replaced words are marked in bold.

Targeting learning powerful feature representation and achieving cross-modal alignment for text-based person search, researchers have developed a batch of technologies over the past few years [Wu *et al.*, 2021; Shao *et al.*, 2022]. It has been proved that the model armed with reasonable tasks tends to learn better representation. In this paper, we propose a representation learning method, namely RaSa, with two novel tasks: relation-aware learning and sensitivity-aware learning for text-based person search.

Relation-aware Learning. In existing methods [Han *et al.*, 2021; Li *et al.*, 2022], the *de facto* optimization objective is

∗Corresponding author

to bring image and text representations of the same identity (*i.e.*, positive pairs) together and repel representations of different identities (*i.e.*, negative pairs) away. However, it tends to encounter the following issue. Normally, a textual description is generated by annotating a particular single image in the text-based person search dataset. The text strongly matches the annotated image without a doubt, whereas it is not always well-aligned to other positive images of the same person at the semantic level due to intra-class variation in the image. As shown in Figure 1 (a), the images and texts depict the same person, leading to a positive relation for each image-text pair. However, there exist two different types of positive relations. $text_1$ (*resp.* $text_2$) is the exact description of $image_1$ (*resp.* $image_2$), where they are completely matched and form a strong positive pair. Nevertheless, $image_1$ and $text_2$ (*resp.* $image_2$ and $text_1$) constitute a weak positive pair with the noise interference. For instance, "white t-shirt" and "blue shorts" in $text_1$ correspond to non-existent objects in $image_2$ due to the occlusion. Existing methods endow the strong and weak positive pairs with equal weight in learning representations, regardless of the noise problem from the weak pairs, eventually leading to overfitting learning.

In order to mitigate the impacts of the noise interference from weak positive pairs, we propose a Relation-Aware learning (RA) task, which is composed of a probabilistic Image-Text Matching (p-ITM) task and a Positive Relation Detection (PRD) task. p-ITM is a variant of the commonly-used ITM, aiming to distinguish negative and positive pairs with a probabilistic strong or weak positive inputting, while PRD is designed to explicitly makes a distinction between the strong and weak positive pairs. Therein, p-ITM emphasizes the consistency between strong and weak positive pairs, whereas PRD highlights their difference and can be regarded as the regularization of p-ITM. The model armed with RA can not only learn valuable information from weak positive pairs by p-ITM but also alleviate noise interference from them by PRD, eventually reaching a trade-off.

Sensitivity-aware Learning. Learning invariant representations under a set of manually chosen transformations (also called *insensitive* transformations in this context) is a general practice for improving the robustness of representation in the existing methods [Chen and He, 2021]. We recognize it but there is more. Inspired by the recent success of equivariant contrastive learning [Dangovski *et al.*, 2022], we explore the *sensitive* transformation that would hurt performance when applied to learn transformation-invariant representations. Rather than keeping invariance under insensitive transformation, we encourage the learned representations to have the ability to be aware of the sensitive transformation.

Towards this end, we propose a Sensitivity-Aware learning (SA) task. We adopt the word replacement as the sensitive transformation and develop a Momentum-based Replaced Token Detection (m-RTD) pretext task to detect whether a token comes from the original textual description or the replacement, as shown in Figure 1 (b). The closer the replaced word is to the original one (*i.e.*, more confusing word), the more difficult this detection task is. When the model is trained to well solve such a detection task, it is expected to have the ability to learn better representation. With these in mind, we use Masked Language Modeling (MLM) to perform the word replacement, which utilizes the image and the text contextual tokens to predict the masked tokens. Furthermore, considering that the momentum model, a slow-moving average of the online model, can learn more stable representations than the current online model [Grill *et al.*, 2020] to generate more confusing words, we employ MLM from the momentum model to carry out the word replacement. Overall, MLM and m-RTD together form a Sensitivity-Aware learning (SA), which offers powerful surrogate supervision for representation learning.

Our contributions can be summarized as follows:

- We differentiate between strong and weak positive image-text pairs in learning representation and propose a relation-aware learning task.
- We pioneer the idea of learning representation under the sensitive transformation to the text-based person search and develop a sensitivity-aware learning task.
- Extensive experiments demonstrate RaSa outperforms existing state-of-the-art methods by 6.94%, 4.45% and 15.35% in terms of Rank@1 metric on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets, respectively.

2 Related Work

2.1 Text-based Person Search

Li *et al.* [2017] first introduce the text-based person search task and publish a challenging dataset CUHK-PEDES. Following this, a series of methods are proposed to solve this task. Part of methods [Zheng *et al.*, 2020a; Wang *et al.*, 2021] focus on designing a reasonable cross-modal alignment strategy, while others [Zhang and Lu, 2018; Shao *et al.*, 2022] concentrate on learning powerful feature representation. For cross-modal alignment, it begins with global alignment [Zheng *et al.*, 2020b] or local correspondences (*e.g.*, patch-word or region-phrase correspondences) [Chen *et al.*, 2022; Niu *et al.*, 2020], and evolves into self-adaptively learning semantic alignment across different granularity [Li *et al.*, 2022; Gao *et al.*, 2021]. Beyond that, some works [Wang *et al.*, 2020; Zhu *et al.*, 2021] utilize external technologies (*e.g.*, human segmentation, pose estimation or attributes prediction) to assist with the cross-modal alignment. For representation learning, Wu *et al.* [2021] propose two color-related tasks based on the observation that color plays a key role in text-based person search. Zeng *et al.* [2021] develop three auxiliary reasoning tasks with gender classification, appearance similarity and image-to-text generation. Ding *et al.* [2021] firstly notice the noise interference from weak positive pairs and propose to keep the difference between strong and weak positive pairs by manually assigning different margins in the triplet loss. More recently, some works [Han *et al.*, 2021; Shu *et al.*, 2022; Yan *et al.*, 2022] resort to vision-language pretraining models to learn better representations. In this paper, we design two novel tasks: RA and SA. RA detects the type of the positive pair to weaken noise from weak positive pairs, differently from the method [Ding *et al.*, 2021] with the sophisticated trick. SA focuses on representation learning by

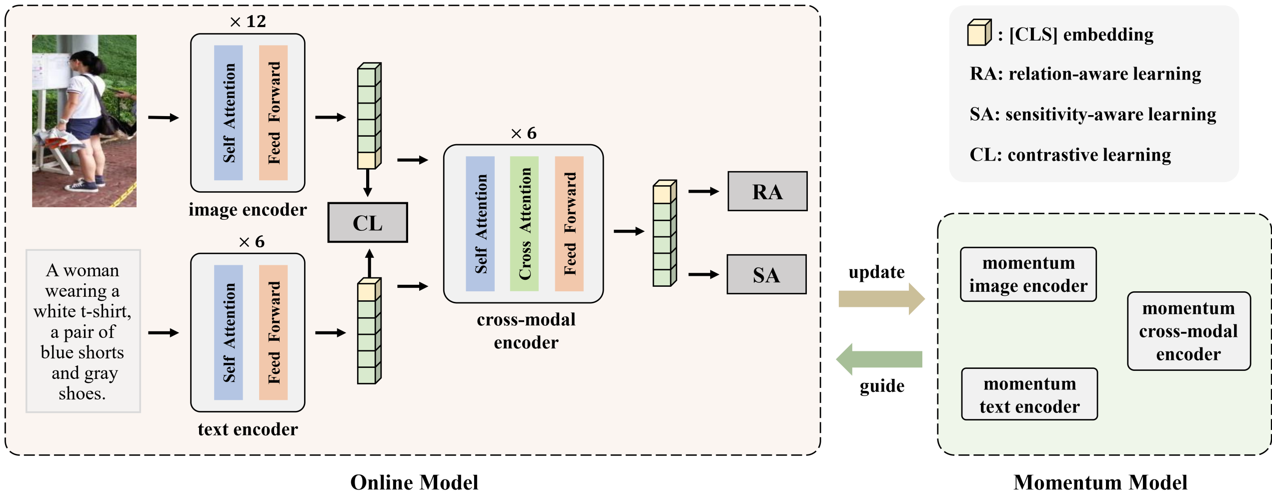


Figure 2: Model architecture of RaSa. It consists of an image encoder, a text encoder and a cross-modal encoder. An intra- and cross-modal CL task is attached after the unimodal encoders for unimodal representation learning. RA and SA tasks are tied after the cross-modal encoders for multi-modal representation learning. The momentum model (a slow-moving of the online model) is used to guide the online model to learn better representations.

detecting sensitive transformation, which is under-explored in the previous methods.

2.2 Equivariant Contrastive Learning

Different from contrastive learning [He *et al.*, 2020] that aims to learn transformation-insensitive representations, equivariant contrastive learning [Dangovski *et al.*, 2022] is recently proposed by additionally encouraging the learned representations to have the ability to be aware of sensitive transformations. Mathematically, the notions of insensitivity and sensitivity can be inductively summarized as: $f(T(x)) = T'(f(x))$ where T denotes a group of transformations of an input instance x, and f is an encoder to compute the representation of x. When T' is the identity transformation, it can be said that f is trained to be insensitive to T; otherwise, f is sensitive to T. Equivariant contrastive learning has shown its successful application in the fields of computer vision (CV) [Dangovski *et al.*, 2022] and natural language processing (NLP) [Chuang *et al.*, 2022], which inspires us to explore sensitive transformations for learning high-quality representations in the cross-modal retrieval task. In this paper, we develop a sensitivity-aware learning with MLM-based word replacement as the sensitive transformation to encourage the model to perceive the replaced words, thus obtaining more informative and discriminative representations.

3 Method

In this section, we take ALBEF [Li *et al.*, 2021] as the backbone[1] and elaborate on the proposed method RaSa by introducing the modal architecture in Section 3.1 and the optimization objectives involving the proposed RA and SA tasks in Section 3.2.

[1]More experiments on other backbones are shown in Appendix.

3.1 Model Architecture

As illustrated in Figure 2, the proposed RaSa consists of two unimodal encoders and a cross-modal encoder. We adopt 12-layer and 6-layer transformer blocks for the image and text encoders, respectively. The cross-modal encoder comprises 6-layer transformer blocks, where a cross-attention module is added after the self-attention module in each block. Considering that the textual description usually covers a part of the information in the corresponding image, we employ a text-guided asymmetric cross-attention module in the cross-modal encoder, *i.e.*, using the textual representation as query and the visual one as key and value. Simultaneously, we maintain a momentum version of the online model via Exponential Moving Average (EMA). Specifically, EMA is formulated as $\hat{\theta} = m\hat{\theta} + (1-m)\theta$, where $\hat{\theta}$ and θ are the parameters of the momentum and online models, respectively, and $m \in [0, 1]$ is a momentum coefficient. The momentum model presents a delayed and more stable version of the online model and is used to guide the online model to learn better representations.

Given an image-text pair (I, T), we first feed the image I into the image encoder to obtain a sequence of visual representations $\{v_{cls}, v_1, \cdots, v_M\}$ with v_{cls} being the global visual representation and v_i $(i = 1, \cdots, M)$ being the patch representation. Similarly, we obtain a sequence of textual representations $\{t_{cls}, t_1, \cdots, t_N\}$ by feeding the text T into the text encoder, where t_{cls} is the global textual representation and t_i $(i = 1, \cdots, N)$ is the token representation. The visual and textual representations are then fed to the cross-modal encoder to obtain a sequence of multi-modal representations $\{f_{cls}, f_1, \cdots, f_N\}$, where f_{cls} denotes the joint representation of I and T, and f_i $(i = 1, \cdots, N)$ can be regarded as the joint representation of the image I and the i-th token in the text T. Simultaneously, the momentum model is employed to obtain a sequence of momentum representations.

3.2 Optimization Objectives

Relation-aware Learning

The vanilla widely-used ITM predicts whether an inputted image-text pair is positive or negative, defined as:

$$L_{itm} = \mathbb{E}_{p(I,T)}\mathcal{H}(y^{itm}, \phi^{itm}(I,T)), \tag{1}$$

where $\mathcal{H}$ represents a cross-entropy function, y^{itm} is a 2-dimension one-hot vector representing the ground-truth label (*i.e.*, $[0,1]^\top$ for the positive pair, and $[1,0]^\top$ for the negative pair), and $\phi^{itm}(I,T)$ is the predicted matching probability of the pair that is computed by feeding f_{cls} into a binary classifier, a fully-connected layer followed by a softmax function.

However, it is unreasonable to directly adopt the vanilla ITM in text-based person search. On the one hand, there exists noise interference from weak positive pairs, which would hamper the representation learning. On the other hand, the weak positive pairs contain certain valuable alignment information that can facilitate representation learning. As a result, to reach a balance, we retain a proportion of weak positive pairs in ITM by introducing the probabilistic inputting. Specifically, we input the weak positive pair with a small probability of p^w and the strong positive pair with a probability of $1-p^w$. To distinguish with the vanilla ITM, we denote the proposed probabilistic ITM as p-ITM.

Furthermore, we continue to alleviate the noise effect of the weak pairs. We propose a Positive Relation Detection (PRD) pretext task to detect the type of the positive pair (*i.e.*, strong or weak), which is formulated as:

$$L_{prd} = \mathbb{E}_{p(I,T^p)}\mathcal{H}(y^{prd}, \phi^{prd}(I,T^p)), \tag{2}$$

where (I,T^p) denotes a positive pair, y^{prd} is the ground truth label (*i.e.*, $[1,0]^\top$ for the strong positive pair and $[0,1]^\top$ for the weak pair), and $\phi^{prd}(I,T^p)$ is the predicted probability of the pair which is computed by appending a binary classifier to the joint representation f_{cls} of the pair.

Taken together, we define the Relation-Aware learning (RA) task as:

$$L_{ra} = L_{p\text{-}itm} + \lambda_1 L_{prd}, \tag{3}$$

where the weight λ_1 is a hyper-parameter.

During the process of the optimization, p-ITM focuses on the consistency between strong and weak positive pairs, while PRD highlights their difference. In essence, PRD plays a role of a regularized compensation for p-ITM. As a whole, RA achieves a trade-off between the benefits of the weak pair and the risk of its side effects.

Sensitivity-aware Learning

Learning invariant representations under the *insensitive* transformation of data is a common way to enhance the robustness of the learned representations. We go beyond it and propose to learn representations that are aware of the *sensitive* transformation. Specifically, we adopt the MLM-based word replacement as the sensitive transformation and propose a Momentum-based Replaced Token Detection (m-RTD) pretext task to detect (*i.e.*, being aware of) the replacement.

Given a strong positive pair (I,T^s), MLM loss is formulated as:

$$L_{mlm} = \mathbb{E}_{p(I,T^{msk})}\mathcal{H}(y^{mlm}, \phi^{mlm}(I,T^{msk})), \tag{4}$$

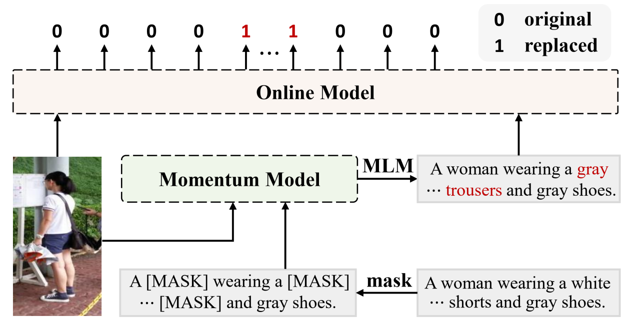


Figure 3: Illustration of m-RTD. It aims to detect whether a token is from the original textual description or the replacement with the aid of the information of the contextual tokens and the paired image. The text with word replacement is obtained by the result of the Masked Language Modeling (MLM) from the momentum model.

where T^{msk} is a masked text in which each token in the input text T^s is randomly masked with a probability of p^m, y^{mlm} is a one-hot vector denoting the ground truth of the masked token and $\phi^{mlm}(I,T^{msk})$ is the predicted probability for the masked token based on the information of the contextual text T^{msk} and the paired image I.

We use the result of MLM from the momentum model as the word replacement, denoted as T^{rep}. The momentum model is a slow-moving of the online model and can learn more stable representations. Therefore, the momentum model is expected to generate more confusing tokens. As m-RTD detects such challenging tokens well, the model is motivated to learn more informative representations to distinguish the tiny differences. Remarkably, besides serving as a generator for the word replacement, MLM also plays a role of token-level optimization, promoting fine-grained representation learning.

Next, m-RTD performs a detection of the MLM-based token replacement. Specifically, the pair (I,T^{rep}) is inputted to the model to obtain a sequence of multi-modal representations $\{f_{cls}, f_1, ..., f_N\}$, and a binary classifier works on $\{f_1, ..., f_N\}$ to predict whether the i-th token is replaced or not. m-RTD minimizes a cross-entropy loss:

$$L_{m\text{-}rtd} = \mathbb{E}_{p(I,T^{rep})}\mathcal{H}(y^{m\text{-}rtd}, \phi^{m\text{-}rtd}(I,T^{rep})), \tag{5}$$

where $y^{m\text{-}rtd}$ is a one-hot vector denoting the ground truth of the replaced token and $\phi^{m\text{-}rtd}(I,T^{rep})$ is the predicted replacement probability. We illustrate the pipeline of m-RTD in Figure 3 for clarity.

Overall, Sensitivity-Aware learning (SA) loss is defined as:

$$L_{sa} = L_{mlm} + \lambda_2 L_{m\text{-}rtd}, \tag{6}$$

where the weight λ_2 is a hyper-parameter.

In conclusion, RA works on the global representation f_{cls} and mainly focuses on the correlation between the image and text, which can be regarded as a coarse-grained optimization. As a complement, SA acts on the token representations $\{f_1, ..., f_N\}$ and pays more attention to the interaction between the image and textual tokens, exhibiting a fine-grained optimization. The two complementary tasks effectively facilitate representation learning.

Contrastive Learning

The proposed RA and SA are directly applied on the multi-modal representations from the cross-modal encoder. Furthermore, we introduce an intermediate Contrastive Learning task (CL) on the representations from the unimodal encoders, so as to make the subsequent cross-modal fusion easier to perform multi-modal representation learning.

Given an image-text pair (I, T), we feed it into the unimodal encoders and obtain the global visual and textual representations v_{cls} and t_{cls}. Then a linear layer is applied to project them to lower-dimensional representations v'_{cls} and t'_{cls}. Meanwhile, we obtain the output of momentum unimodal encoders, denoted as $\hat{v}'_{cls}$ and $\hat{t}'_{cls}$. We maintain an image queue $\hat{Q}_v$ and a text queue $\hat{Q}_t$ to store the recent R projected representations $\hat{v}'_{cls}$ and $\hat{t}'_{cls}$, similarly to MoCo [He *et al.*, 2020]. The introduction of the queues implicitly enlarges the batch size, and a larger batch will provide more negative samples, thereby facilitating representation learning.

In CL, the general form of InfoNCE loss is formulated as:

$$L_{nce}(x, x_+, Q) = -\mathbb{E}_{p(x,x_+)}[\log \frac{\exp(s(x, x_+)/\tau)}{\sum\limits_{x_i \in Q} \exp(s(x, x_i)/\tau)}], \quad (7)$$

where τ is a learnable temperature parameter, Q denotes a maintained queue, and $s(x, x_+) = x^{\mathrm{T}} x_+ / \|x\| \|x_+\|$ measures the cosine similarity between x and x_+.

Beyond the widely-used cross-modal image-text contrastive learning (ITC) [Li *et al.*, 2021; Radford *et al.*, 2021], denoted as:

$$L_{itc} = [L_{nce}(v'_{cls}, \hat{t}'_{cls}, \hat{Q}_t) + L_{nce}(t'_{cls}, \hat{v}'_{cls}, \hat{Q}_v)] \;/\; 2, \quad (8)$$

we additionally explore the intra-modal contrastive learning (IMC). The representations of the same person are supposed to stay closer than those of different persons within each modality. IMC loss is formulated as:

$$L_{imc} = [L_{nce}(v'_{cls}, \hat{v}'_{cls}, \hat{Q}_v) + L_{nce}(t'_{cls}, \hat{t}'_{cls}, \hat{Q}_t)] \;/\; 2. \quad (9)$$

Taken together, we define the overall loss for CL as:

$$L_{cl} = (L_{itc} + L_{imc}) \;/\; 2. \quad (10)$$

Joint Learning

Overall, we formulate the joint optimization objective as:

$$L = L_{ra} + L_{sa} + \lambda_3 L_{cl}, \quad (11)$$

where λ_3 is a hyper-parameter.

During inference, given a query text and a large-scale image pool, we use the predicted matching probability from p-ITM to rank all images. Considering the inefficiency of the cross-modal encoder with quadratic interaction operation, we refer to ALBEF [Li *et al.*, 2021] and exclude a large number of irrelevant image candidates prior to the cross-modal encoder, thereby speeding up the inference. Specifically, we first calculate each pair's similarity $s(t_{cls}, v_{cls})$ via the unimodal encoders, and then select the first 128 images with the highest similarities to send them to the cross-modal encoder and compute the p-ITM matching probabilities for ranking.

	Method	R@1	R@5	R@10	mAP
w/o VLP	GNA-RNN [Li *et al.*, 2017]	19.05	-	53.64	-
	Dual Path [Zheng *et al.*, 2020b]	44.40	66.26	75.07	-
	CMPM/C [Zhang and Lu, 2018]	49.37	71.69	79.27	-
	ViTAA [Wang *et al.*, 2020]	55.97	75.84	83.52	-
	DSSL [Zhu *et al.*, 2021]	59.98	80.41	87.56	-
	MGEL [Wang *et al.*, 2021]	60.27	80.01	86.74	-
	ACSA [Ji *et al.*, 2022]	63.56	81.40	87.70	-
	SAF [Li *et al.*, 2022]	64.13	82.62	88.40	58.61
	TIPCB [Chen *et al.*, 2022]	64.26	83.19	89.10	-
	CAIBC [Wang *et al.*, 2022b]	64.43	82.87	88.37	-
	$\mathrm{C_2A_2}$ [Niu *et al.*, 2022]	64.82	83.54	89.77	-
	LGUR [Shao *et al.*, 2022]	65.25	83.12	89.00	-
w/ VLP	PSLD [Han *et al.*, 2021]	64.08	81.73	88.19	60.08
	IVT [Shu *et al.*, 2022]	65.59	83.11	89.21	-
	CFine [Yan *et al.*, 2022]	69.57	85.93	91.15	-
	ALBEF(backbone) [Li *et al.*, 2021]	60.28	79.52	86.34	56.67
	RaSa (Ours)	**76.51**	**90.29**	**94.25**	**69.38**

Table 1: Comparison with other methods on CUHK-PEDES. VLP denotes vision-language pretraining. For a fair comparison, all reported results come from the methods without re-ranking.

4 Experiments

We conduct experiments on three text-based person search datasets: CUHK-PEDES [Li *et al.*, 2017], ICFG-PEDES [Ding *et al.*, 2021] and RSTPReid [Zhu *et al.*, 2021]. *The introduction of each dataset and the implementation details of the proposed method are shown in Appendix.*

4.1 Evaluation Protocol

We adopt the widely-used Rank@K (R@K for short, K=$1, 5, 10$) metric to evaluate the performance of the proposed method. Specifically, given a query text, we rank all the test images via the similarity with the text and the search is deemed to be successful if top-K images contain any corresponding identity. R@K is the percentage of successful searches. We also adopt the mean average precision (mAP) as a complementary metric.

4.2 Backbones

Most text-based person search methods [Li *et al.*, 2022; Shao *et al.*, 2022] rely on two feature extractors pre-trained on unaligned images and texts separately, such as ResNet [He *et al.*, 2016] or ViT [Dosovitskiy *et al.*, 2020] for the visual extractor, Bi-LSTM [Hochreiter and Schmidhuber, 1997] or BERT [Devlin *et al.*, 2018] for the textual extractor. Recently, some works [Shu *et al.*, 2022; Yan *et al.*, 2022] have applied vision-language pretraining (VLP) to text-based person search and obtained impressive results. Following this, we adopt VLP models as the backbone.

The proposed RaSa can be plugged into various backbones. To adequately verify the effectiveness, we conduct RaSa on three VLP models: ALBEFF [Li *et al.*, 2021], TCL [Yang *et al.*, 2022] and CLIP [Radford *et al.*, 2021]. We use ALBEF as the backbone by default in the following experiments, which is pre-trained on 14M image-text pairs and adopts ITC and

	Method	R@1	R@5	R@10	mAP
w/o VLP	Dual Path [Zheng *et al.*, 2020b]	38.99	59.44	68.41	-
	CMPM/C [Zhang and Lu, 2018]	43.51	65.44	74.26	-
	ViTAA [Wang *et al.*, 2020]	50.98	68.79	75.78	-
	SSAN [Ding *et al.*, 2021]	54.23	72.63	79.53	-
	SAF [Li *et al.*, 2022]	54.86	72.13	79.13	32.76
	TIPCB [Chen *et al.*, 2022]	54.96	74.72	81.89	-
	SRCF [Suo *et al.*, 2022]	57.18	75.01	81.49	-
	LGUR [Shao *et al.*, 2022]	59.02	75.32	81.56	-
w/ VLP	IVT [Shu *et al.*, 2022]	56.04	73.60	80.22	-
	CFine [Yan *et al.*, 2022]	60.83	76.55	82.42	-
	ALBEF(backbone) [Li *et al.*, 2021]	34.46	52.32	60.40	19.62
	RaSa (Ours)	**65.28**	**80.40**	**85.12**	**41.29**

Table 2: Comparison with other methods on ICFG-PEDES.

	Method	R@1	R@5	R@10	mAP
w/o VLP	DSSL [Zhu *et al.*, 2021]	32.43	55.08	63.19	-
	SSAN [Ding *et al.*, 2021]	43.50	67.80	77.15	-
	SAF [Li *et al.*, 2022]	44.05	67.30	76.25	36.81
	CAIBC [Wang *et al.*, 2022b]	47.35	69.55	79.00	-
	ACSA [Ji *et al.*, 2022]	48.40	71.85	81.45	-
	C_2A_2 [Niu *et al.*, 2022]	51.55	76.75	85.15	-
w/ VLP	IVT [Shu *et al.*, 2022]	46.70	70.00	78.80	-
	CFine [Yan *et al.*, 2022]	50.55	72.50	81.60	-
	ALBEF(backbone) [Li *et al.*, 2021]	50.10	73.70	82.10	41.73
	RaSa (Ours)	**66.90**	**86.50**	**91.35**	**52.31**

Table 3: Comparison with other methods on RSTPReid.

ITM tasks for image-text retrieval. *The details and experiments on TCL and CLIP are shown in Appendix.*

4.3 Comparison with State-of-the-art Methods

We compare the proposed RaSa with the existing text-based person search methods on CUHK-PEDES, ICFG-PEDES and RSTPReid, as shown in Table 1, 2 and 3, respectively. RaSa achieves the highest performance in terms of all metrics, outperforming existing state-of-the-art methods by a large margin. Specifically, compared with the current best-performing method CFine [Yan *et al.*, 2022], RaSa gains a significant R@1 improvement of 6.94%, 4.45% and 15.35% on the three datasets, respectively. The comparison clearly demonstrates the effectiveness of RaSa in text-based person search.

4.4 Ablation Study

We analyze the effectiveness and contribution of each optimization objective in RaSa by conducting a series of ablation experiments on CUHK-PEDES, as shown in Table 4.

Effectiveness of Optimization Objectives

RaSa consists of three optimization objectives. CL provides an explicit alignment before the cross-modal fusion. RA implements the deep fusion by the cross-modal encoder with an alleviation of noise interference. And SA encourages the learned representations to be sensitive to the MLM-based token replacement.

We can see from Table 4, (1) RaSa with a single CL achieves a modest performance of 61.35% and 59.44% in terms of R@1 and mAP, respectively. On account of the modality gap between the image and text and the fine-grained intra-class variation, CL contributes a coarse alignment with a lack of deep interaction across modalities, which is not enough to handle such a challenging retrieval task. (2) When adding RA(p-ITM + PRD), the performance has a remarkable improvement of 12.85% at R@1 and 8.67% at mAP, effectively demonstrating that deep cross-modal fusion with RA is extraordinarily significant to text-based person search. And (3) with the aid of SA(MLM + m-RTD), RaSa achieves the best performance of 76.51% at R@1 and 69.38% at mAP. SA utilizes the visual information and the contextual token information of the corresponding text to detect whether a token has been replaced or not. In order to handle such a challenging detection task, the learned representations are encouraged to be powerful enough to distinguish the tiny difference between the original token and the replaced one.

Module	Setting	R@1	R@5	R@10	mAP
CL	ITC + IMC	61.35	80.44	86.91	59.44
+RA	ITM	71.29	86.70	91.46	67.82
	s-ITM	73.52	88.71	92.98	66.74
	p-ITM	72.58	87.98	92.51	68.29
	ITM + PRD	73.03	87.75	92.45	68.45
	p-ITM + PRD	74.20	89.02	92.95	68.11
++SA	MLM	74.81	89.85	93.66	68.32
	MLM + f-RTD	75.13	89.93	93.47	69.17
	MLM + o-RTD	75.99	90.21	94.09	69.35
	MLM + m-RTD	76.51	90.29	94.25	69.38

Table 4: Comparison of RaSa with different settings on CUHK-PEDES. ITM learns from all positive pairs without a probabilistic inputting. s-ITM learns from only strong positive pairs and discards all weak positive pairs. p-ITM uses a probabilistic inputting of strong and weak positive pairs. f-RTD adopts DistilBERT [Sanh *et al.*, 2019] as a fixed generator to produce the replaced tokens. o-RTD uses the online model as the generator, while m-RTD is based on the momentum model.

Analysis of RA

RA contains p-ITM and PRD, where the former focuses on the consistency between the strong and weak positive pairs, while the latter highlights their difference, serving as a regularization of p-ITM.

The vanilla ITM learns from all positive pairs without the probabilistic inputting. However, there exists too much noise interference from weak positive pairs. Intuitively, we can discard all weak positives to get rid of the noise. s-ITM only uses the strong positive pairs and gains a boost of 2.23% at R@1 compared to the vanilla ITM. Nevertheless, such a straightforward way ignores the weak supervision from the weak positives which is also beneficial to representation learning. To reach a trade-off between the benefits of the weak supervision and the risk of side effects, p-ITM resorts to the probabilistic inputting and retains a small proportion of the weak positives. Compared with the vanilla ITM and s-ITM, p-ITM achieves an intermediate performance. Not surprisingly at all, the more noise there exists, the more it affects the retrieval

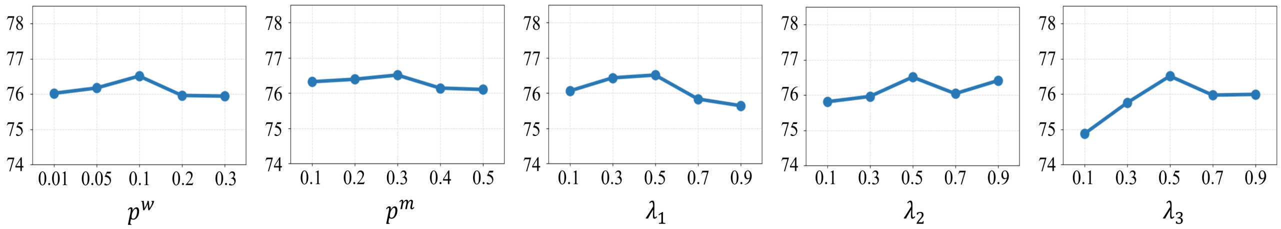


Figure 4: The impact of the hyper-parameters at R@1 on CUHK-PEDES. p^w denotes the probability of inputting weak positive pairs in RA. p^m means the masking ratio of the tokens in a text in SA. λ_1, λ_2 and λ_3 are the loss weights.

performance. In order to alleviate the impact of the noise, we further propose PRD to perform an explicit distinction between the strong and weak positives, which serve as a regularization for p-ITM. Significantly, no matter whether adding PRD to the vanilla ITM or p-ITM, PRD can obtain consistent performance improvement, which powerfully demonstrates its effectiveness.

Analysis of SA

SA includes MLM and m-RTD. MLM not only plays the role of generating the text with word replacement but also performs a token-level optimization. m-RTD detects the replaced tokens by virtue of the visual information and the contextual token information.

Based on CL and RA, adding a single MLM without the replacement detection task brings a slight boost of 0.61% at R@1. Furthermore, we introduce the detection task and use the momentum model as the generator to produce the replaced tokens. In order to adequately investigate the effectiveness of the generator, we compare three different variants. (1) Following DiffCSE [Chuang *et al.*, 2022], we use DistilBERT [Sanh *et al.*, 2019] as a fixed generator for the word replacement, which is denoted as f-RTD. From Table 4, RaSa with f-RTD gains a modest performance of 75.13% at R@1. We argue that the generated tokens from a fixed generator can be easily detected as the training advances and thus provides a limited effect on learning representation. (2) o-RTD adopts the online model as the generator. RaSa with o-RTD achieves a better performance of 75.99% at R@1. Compared with f-RTD, o-RTD resorts to a dynamic generator which is optimized constantly during the whole training process and can produce more confusing tokens with the proceeding of the model's training, effectively increasing the difficulty of replaced tokens detection and facilitating representation learning. And (3) m-RTD adopts the momentum model as the generator and reaches the best performance of 76.51% at R@1. The momentum model is a slow-moving of the online model and can obtain more stable representations. As the training goes ahead, the momentum model iteratively bootstraps MLM to generate more challenging tokens for detection, which encourages the learned representations to be powerful enough to distinguish the tiny difference and substantially improve results.

Hyper-parameters

In Section 3.2, we use the inputting probability p^w to retain a small proportion of weak positive pairs to alleviate the noise, the masking ratio p^m to randomly mask tokens to perform the replaced token detection, and the loss weights λ_1, λ_2, λ_3 to make a trade-off. We show how these hyper-parameters impact the performance of RaSa in Figure 4. (1) The best result is achieved at $p^w = 0.1$. The inputting probability p^w in RA is introduced to seek a balance between the useful information and the noise from weak positives. A larger p^w may introduce too much noise, while a smaller p^w hinders the model from making full use of the useful information. (2) RaSa performs best at $p^m = 0.3$. A larger p^m brings more perturbations to the text, making the detection task too difficult to be carried out. In contrast, when p^m goes smaller, SA will contribute less to representation learning. And (3) for the loss weights λ_1, λ_2 and λ_3, they present an overall trend of first increasing and then decreasing. Empirical results show that RaSa performs best when they are set as 0.5.

4.5 Extended Experiments and Visualization

To go a step further and validate the effectiveness of RaSa, we perform extended experiments on two coarse-grained image-text retrieval datasets (Flickr30K [Plummer *et al.*, 2015] and COCO [Lin *et al.*, 2014]), as well as two fine-grained datasets (CUB [Reed *et al.*, 2016] and Flowers [Reed *et al.*, 2016]). *The experimental results are shown in Appendix.* Besides, we conduct a series of domain generalization experiments following LGUR [Shao *et al.*, 2022] in Appendix to verify the generalization ability of RaSa. These results clearly demonstrate the effectiveness and the generalization ability of RaSa.

For a qualitative analysis, we also present the retrieval visualization in Appendix, vividly showing the excellent retrieval ability of RaSa.

5 Conclusion

In this paper, we propose a Relation and Sensitivity aware representation learning method (RaSa) for text-based person search, which contains two novel tasks, RA and SA, to learn powerful multi-modal representations. Given that the noise from the weak positive pairs tends to result in overfitting learning, the proposed RA utilizes an explicit detection between strong and weak positive pairs to highlight the difference, serving as a regularization of p-ITM that focuses on their consistency. Beyond learning transformation-insensitive representations, SA encourages the sensitivity to MLM-based token replacement. Extensive experiments on multiple benchmarks demonstrate the effectiveness of RaSa.

Acknowledgments

This work is supported by the National Science Foundation of China under Grant NSFC 62002252, and is also partially supported by the National Science Foundation of China under Grant NSFC 62106165.

References

[Cao *et al.*, 2022] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5410–5417. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Survey Track.

[Chen and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

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

[Chuang *et al.*, 2022] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. Diffcse: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, 2022.

[Dangovski *et al.*, 2022] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2022.

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

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

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

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

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[Han *et al.*, 2021] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 337, 2021.

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

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

[Ji *et al.*, 2021] Deyi Ji, Haoran Wang, Hanzhe Hu, Weihao Gan, Wei Wu, and Junjie Yan. Context-aware graph convolution network for target re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1646–1654, 2021.

[Ji *et al.*, 2022] Zhong Ji, Junhua Hu, Deyin Liu, Lin Yuanbo Wu, and Ye Zhao. Asymmetric cross-scale alignment for text-based person search. *IEEE Transactions on Multimedia*, pages 1–11, 2022.

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

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc., 2021.

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

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[Niu *et al.*, 2022] Kai Niu, Linjiang Huang, Yan Huang, Peng Wang, Liang Wang, and Yanning Zhang. Cross-modal co-occurrence attributes alignments for person search by language. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4426–4434, 2022.

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

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

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58, 2016.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Shao *et al.*, 2022] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5566–5574, 2022.

[Shu *et al.*, 2022] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. *arXiv preprint arXiv:2208.08608*, 2022.

[Suo *et al.*, 2022] Wei Suo, Mengyang Sun, Kai Niu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. A simple and robust correlation filtering method for text-based person search. In *European Conference on Computer Vision*, pages 726–742. Springer, 2022.

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

[Wang *et al.*, 2021] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. Text-based person search via multi-granularity embedding learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1068–1074, 2021.

[Wang *et al.*, 2022a] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2540–2549, 2022.

[Wang *et al.*, 2022b] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5314–5322, 2022.

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

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

[Yang *et al.*, 2022] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[Zeng *et al.*, 2021] Pengpeng Zeng, Shuaiqi Jing, Jingkuan Song, Kaixuan Fan, Xiangpeng Li, Liansuo We, and Yuan Guo. Relation-aware aggregation network with auxiliary guidance for text-based person search. *World Wide Web*, pages 1–18, 2021.

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

[Zheng *et al.*, 2020a] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. Hierarchical gumbel attention network for text-based person search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3441–3449, 2020.

[Zheng *et al.*, 2020b] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[Zhu *et al.*, 2021] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217, 2021.