frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 19 May 2023
DOI 10.3389/fphy.2023.1192412

Check for updates

OPEN ACCESS

EDITED BY
Guanqiu Qi,
Buffalo State College, United States

REVIEWED BY
Hao Tang,
Nanjing University of Science and Technology, China
Kangjian He,
Yunnan University, China
Puhong Duan,
Hunan University, China

*CORRESPONDENCE
Fan Li,
✉ 292789049@qq.com

RECEIVED 23 March 2023
ACCEPTED 18 April 2023
PUBLISHED 19 May 2023

CITATION
Zhou H, Li F, Tian X and Huang Y (2023), Feature semantic alignment and information supplement for Text-based person search.
Front. Phys. 11:1192412.
doi: 10.3389/fphy.2023.1192412

Feature semantic alignment and information supplement for Text-based person search

Hang Zhou[1], Fan Li *, Xuening Tian [2] and Yuling Huang[3]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, [2]LTH Engineering College at Campus Helsingborg, Lund University, Lund, Sweden, [3]School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

The goal of person text-image matching is to retrieve images of specific pedestrians using natural language. Although a lot of research results have been achieved in persona text-image matching, existing methods still face two challenges. First,due to the ambiguous semantic information in the features, aligning the textual features with their corresponding image features is always tricky. Second, the absence of semantic information in each local feature of pedestrians poses a significant challenge to the network in extracting robust features that match both modalities. To address these issues, we propose a model for explicit semantic feature extraction and effective information supplement. On the one hand, by attaching the textual and image features with consistent and clear semantic information, the course-grained alignment between the textual and corresponding image features is achieved. On the other hand, an information supplement network is proposed, which captures the relationships between local features of each modality and supplements them to obtain more complete local features with semantic information. In the end, the local features are then concatenated to a comprehensive global feature, which capable of precise alignment of the textual and described image features. We did extensive experiments on CUHK-PEDES dataset and RSTPReid dataset, the experimental results show that our method has better performance. Additionally, the ablation experiment also proved the effectiveness of each module designed in this paper.

KEYWORDS

cross-modal retrieval, neural network, Text-based person search, deep learning, Text-based image retrieval

1 Introduction

Person text-image matching method has been proposed in order to deal with special cases. For example, if a child is missing in an amusement park, parents can quickly find the area where the child is located in the surveillance equipment by describing the child's appearance. This technique uses the textual description of the pedestrian's appearance provided to retrieve the target pedestrian image. Compared with person re-identificationLi et al. [1]; Zhang et al. [2]; Wang et al. [3]; Zhang et al. [4]; Li et al. [5], it is not limited to the need of pedestrian images as a query condition, so it compensates for the disadvantages of using pedestrian re-identification techniques in the presence of surveillance blind spotsZhu et al. [6]; Li et al. [7]; Lingli et al. [8]; Li et al. [9]. Therefore the technique has practical value.

In this task, ensuring the consistency of text semantic information and image semantic information is one of the factors that affect the retrieval performance. In recent years, many effective methods of feature semantic consistency have been proposed. These methods can be roughly divided into methods based on local relationship correspondence Ding et al. [10]; Zhang et al. [11]; Liu et al. [12]; Zheng et al. [13]; Chen et al. [14], etc.), methods based on external knowledge Jing et al. [15]; Wang et al. [16]; Aggarwal et al. [17]; Wang et al. [18], methods based on similarity measurement Niu et al. [19]; Gao et al. [20], and methods based on multi-head attention mechanism Wang et al. [21]; Li et al. [22]. The method based on local relation correspondence often achieves local alignment of text features and image features through specific functional relations. In the method based on external knowledge, human body semantics Wang et al. [16], pedestrian posture Jing et al. [15] and pedestrian attributes Li et al. [23]; Wang et al. [18] are often used as auxiliary information for text and visual feature alignment. The method based on similarity measure measures the similarity between noun phrases and local patches in the image, and determines the relationship between them according to the predicted weight. Compared with the attention mechanism used in the method based on local relationship correspondence, the method based on multi-head attention mechanism usually assigns different semantics to each head to align the heads with the same semantics.

Although many effective methods have been proposed, there are still some problems that have not been effectively solved. On the one hand, features alignment based on local relationships always faces the problem of vague semantic information in local features. This is because different images of the same person may include spatial discrepancy and different sentences describing the same person may have differences in the order of expression and logic. On the other hand, some special semantic information in the text, such as “Coat,” corresponds to multiple body parts, such as the upper and lower body, in the image. Therefore, those various semantics are not entirely independent, and there should be correlations between multiple different semantics. Previous methods for extracting local features have not put these correlations in their consideration, resulting in the loss of some semantic information in the local features and presents a challenge in building up local correspondences between text and images. Although attention-based methods Zheng et al. [13]; Liu et al. [12]; Gao et al. [20] can effectively alleviate this problem, they require high computational cost and loss of efficiency.

This paper proposes a method for explicit semantic feature extraction and an information supplement network to address the challenge of aligning textual and image features of pedestrians. The relationships between key information in the sentence have been fully considered. On the one hand, we starts with a class token obtained from the transformer, and predicts the features that correspond to the local regions of the pedestrian image from the global features with relationship embedding, thus obtaining local features that are roughly aligned and have clear semantics. On the other hand, the information supplement network is proposed to adaptively probe the relationships between local features, and such relationships are used to fuse out semantically well-informed local features. In the end, the local features with improved semantic information are precisely aligned and concatenated in a certain order on the channel to form globally aligned features with comprehensive semantics.

Our research contributions are as follows

- In this paper, we propose a explicit semantic feature extraction method. We use local features of pedestrian image with clear semantics to guide text feature extraction with fuzzy semantics, and achieve a rough alignment between local features of text and local features of images.
- To address the challenge of semantics loss in local features, this paper proposes an effective information supplement network to complement the missing information.

2 Related works

2.1 Text-based image retrieval

Text-based image retrieval Liu et al. [24] is a technique that uses natural language to retrieve specific images. It differs from the single-modal task Tang et al. [25,26]; Zha et al. [27]; Li et al. [28] in that it requires overcoming greater modal differences. Depending on the testing process, we can divide these methods into modal interaction methods Gao et al. [20] and modal non-interaction methods Chen et al. [14]; Liu et al. [24]. The modal interaction methods often requires each text feature with all image features to derive results through a complex cross-modal attention mechanism, which undoubtedly increases the time cost and it’s hard to deploy in a real world scenario. The modal non-interaction methods can extract image features or text features separately and does not require two modality features for cross-learning, saving time overhead. Therefore, it can be used for large-scale text image retrieval tasks. However, these methods do not take into account the impact of semantic clarity and missing feature information on cross-modality matching.

2.2 Person Text-image matching

Person text-image matching faces different problems than Text-based Image Retrieval. In a Text-based Image Retrieval task, an image usually contains multiple objects, and the model design often needs to consider the association between objects. In the person text-image matching task, there is usually only one pedestrian object, so the model design needs to consider extracting fine-grained features. Therefore, by comparison, person text-image matching is more challenging. Li et al.Li et al. [29] first proposed the person text-image matching task and successfully completed the task using recurrent neural network with gated neural attention mechanism. Meanwhile, a large-scale person description dataset named the CUHK person description dataset was constructed. Because there are too many defects in the first proposed method. Subsequently, Li et al. Li et al. [30] proposed an identity aware two-stage network. The network extracts robustness features through two steps.

In recent years researchers have proposed a variety of methods, which can be broadly classified into the following three categories: similarity relation metrics based methods Niu et al. [19]; Gao et al. [20]; Li et al. [30], external knowledge assistance based methods Wang et al. [16]; Jing et al. [15]; Aggarwal et al. [17]; Wang et al. [18], and multi-granularity relational correspondence based feature alignment methods Ding et al. [10]; Zhang et al. [11]; Liu et al. [12]; Zheng et al. [13]; Chen et al. [14]; Wang et al. [21]. Similarity relation metrics based methods use the similarity between text features and image features as the relationship between them to obtain robust features. Then, during testing, it also requires each text

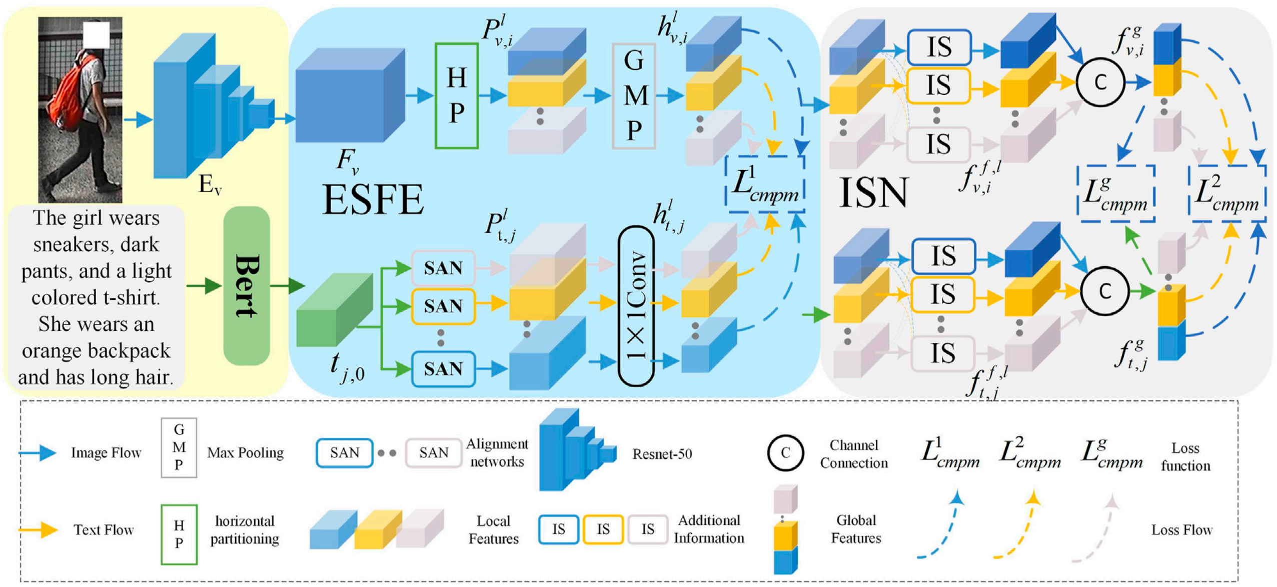


FIGURE 1
Overall framework of the proposed method.

to do the same operation with all images, which greatly reduces the testing efficiency. External knowledge assistance based methods need to construct external knowledge in advance to assist the model in extracting features. However, the model performance is highly dependent on the external knowledge, and the performance also depends on how well the external knowledge is constructed. Multi-granularity relational correspondence based feature alignment methods usually align the features of each granularity directly, which reduces the model performance without explicit semantic. This reduces the performance of the model without explicit semantics. In contrast to the above methods. To accomplish semantic alignment between textual and pedestrian image features, this study proposes obtaining distinct local features by projecting the global features of text onto the local feature space of the corresponding pedestrian image through a nonlinear mapping mechanism. Subsequently, the information supplementation network complements local feature information to achieve refined alignment of local features with comprehensive information. Utilizing these aligned local features, global features with coherent semantic information are then constructed.

3 Proposed method

3.1 Overview

The technical framework of this paper consists of two main parts: Explicit Semantic Feature Extraction (ESFE) and Information Supplementation Network (ISN), as shown in Figure 1. ESFE guides the image features with clear semantics to align with the vague semantic text features, thereby achieving semantic alignment and laying a solid foundation for downstream tasks. ISN is responsible for establishing the relationships between various local features and fusing them based on these relationships to eliminate the incompleteness of the features and obtain more robust features.

3.2 Feature extraction

In the extraction of person image features, ResNet50 is used as the backbone and is denoted as E_v. As shown in Figure 1, the output image features is denoted as $F_v \in^{H\times W\times C}$, and split F_v horizontally into N patches. Where H, W and C denote the length, width, and number of channels in the feature map, respectively. The feature maps of the l-th patch of the i-th Pedestrian image in a batch are represented as $X^l_{v,i} \in R^{H/N\times W\times C}$, where N denotes the total number of patches. We performed maximum pooling on each patch to obtain the feature vector $h^l_{v,i}$.

In the extraction of text features, the pre-trained Bidirectional Encoder Representation from Transformers (BERT) model Kingma and Ba (2014) is used as the backbone. The output text features are $Y_j = (t_{j,0}, t_{j,1}...t_{j,M}) \in R^{(M+1)\times D}$, where $t_{j,1}, \ldots, t_{j,M}$ represents the features of M words. and $t_{j,0}$ represents the global features of Y_j.

3.3 Explicit semantic feature extraction

Since the text describing the same pedestrian may have multiple sentences and inconsistent features after encoding, this unclear semantics leading to ineffective alignment. To address this problem, this paper proposes the Explicit Semantic Feature Extraction (ESFE) module. In general, this module bases on the fact that each divided region of the pedestrian image has clearer semantic information, which we can use to guide the learning of the text features. By aligning the semantic information between text and image features, the proposed module endows text features with clear

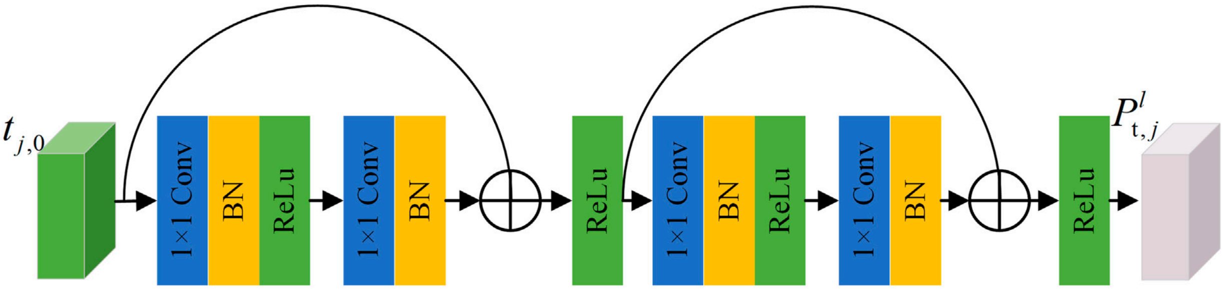


FIGURE 2
Semantic alignment network.

semantic information. Specifically, the Semantic Alignment Network (SAN) is employed in the ESFE module to map classification token to features with local semantics that are consistent with pedestrian image features.

The structure of the semantic alignment network is shown in Figure 2, in order to generate the same amount as the N local features of the image. The Explicit Semantic Feature Extraction (ESFE) module also contains N of Semantic Alignment Network (SAN). Each SAN has its own set of parameters and is used to map the global feature of text to different semantic spaces that correspond to different local features of the pedestrian. Assume that j-th text feature of the pedestrian encoded by the Bert network is $t_{j,0}$, the resulting feature we get from SAN is represent as $P^l_{\mathrm{t},j}$.

It is necessary to ensure that the feature channels of both modalities are the same, and we use 1 × 1convolution to expand the number of channels of $P^l_{\mathrm{t},j}$ to obtain $h^l_{t,j}$. We also use Cross-Modal Projection Matching (CMPM) loss Zhang and Lu [31]. $h^l_{v,i}$ and $h^l_{t,j}$ are matching probabilities which can be calculated by the following equation:

$$\boldsymbol{S}^l_{v2t,i,j} = \frac{\exp\left(\left(h^l_{v,i}\right)^T \bar{h}^l_{t,j}\right)}{\sum_{j=1}^{n} \exp\left(\left(h^l_{v,i}\right)^T \bar{h}^l_{t,j}\right)}, \tag{1}$$

where $\bar{h}^l_{t,i} = \frac{h^l_{t,i}}{\|h^l_{t,i}\|_2}$, and the matching loss from image to text in a mini-batch is computed by:

$$\boldsymbol{L}_{v2t}\left(E_v\right) = \frac{1}{n}\sum_{l=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} S^l_{v2t,i,j}\log\left(\frac{S^l_{v2t,i,j}}{z^l_{i,j}+\varepsilon}\right), \tag{2}$$

where $\varepsilon = 10^{-8}$, n is the batchsize,$z^l_{i,j} = y^l_{i,j} / \sum_{j=1}^{N} y^l_{i,j}$, and $y^l_{i,j} = 1$ indicates that both belong to the same ID. The loss in the v2t direction adds the loss in the t2v direction to obtain the CMPM loss. The formula is shown as follows:

$$\boldsymbol{L}^1_{cmpm}\left(E_v\right) = L_{v2t}\left(E_v\right) + L_{t2v}\left(E_v\right), \tag{3}$$

3.4 Information supplementation network learning

To address the issue of information incompleteness in the individual local features, which hinders a comprehensive

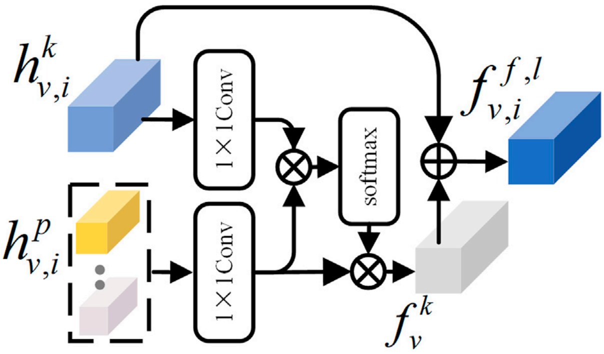


FIGURE 3
Information supplementation network.

representation of the features, we propose the Information Supplementation Network (ISN) to enrich the semantic information of the local features and thus enhance the feature representation. For the image modality, the local features $h^l_{v,i}$ with missing semantic information are supplemented using ISN (as shown in Figure 3) to obtain locally complete features$f^l_{v,i}$ which will later be concatenated in a specific channel order to form robust global features $f^g_{v,i}$. We illustrate this process using the k-th visual local feature as an example. First, we compute the similarity between $h^k_{v,i}$ and $h^p_{v,i}\,(p \neq k)$ embedded in a common space:

$$\boldsymbol{S}_{\boldsymbol{k},\boldsymbol{p}} = \frac{W_k\left(h^k_{v,i}\right)^T W_p\left(h^p_{v,i}\right)}{\|W_k\left(h^k_{v,i}\right)\|\|W_p\left(h^p_{v,i}\right)\|}, \tag{4}$$

where, W_k, $W_p$are two parameter matrices that can be updated during training. Then, the association strength between the k-th image local feature and the other local features can be expressed as follows:

$$\boldsymbol{\alpha}_{\boldsymbol{k},\boldsymbol{p}} = \frac{\exp\left(S_{k,p}\right)}{\sum_{p=1,p\neq k}^{N}\exp\left(S_{k,p}\right)}, \tag{5}$$

After extracting the missing information of local feature $h^k_{v,i}$ among N-1 local features using $\alpha_{k,p}$, we can obtain the missing information of the k-th visual local feature f^k_v, with the following equation.

$$\boldsymbol{f}_v^k = W_a\left(\sum\nolimits_{p=1,p\neq k}^{N} \alpha_{k,p} W_P\left(h_{v,i}^k\right)\right), \tag{6}$$

Finally, the semantic information is refined by fusing the missing information with the original information, and the formula can be expressed as follows.

$$\boldsymbol{f}_{v,i}^k = W_f\left(f_v^k + h_{v,i}^k\right), \tag{7}$$

where W_f, W_a are two learnable matrices. Similar to the visual features, we also process the text local features by the same steps as mentioned above to obtain semantically perfect text local features $f_{t,j}^l$. To ensure the consistency of local features, the following loss function is used to optimize the training.

$$\begin{aligned}\boldsymbol{L}_{cmpm}^2\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = L_{v2t}\left(E_v, E_c, W_l, W_p, W_f, W_a\right) \\ + L_{t2v}\left(E_v, E_c, W_l, W_p, W_f, W_a\right),\end{aligned} \tag{8}$$

$$\boldsymbol{L}_{v2t}\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = \frac{1}{n}\sum_{l=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{i,j}^l \log\left(\frac{p_{i,j}^l}{q_{i,j}^l + \varepsilon}\right), \tag{9}$$

$$\boldsymbol{L}_{t2v}\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = \frac{1}{n}\sum_{l=1}^{N}\sum_{j=1}^{n}\sum_{i=1}^{n} p_{i,j}^l \log\left(\frac{p_{i,j}^l}{q_{i,j}^l + \varepsilon}\right), \tag{10}$$

$$\boldsymbol{p}_{i,j}^l = \frac{\exp\left(\left(f_{v,i}^l\right)^T f_{t,j}^l\right)}{\sum_{j=1}^{n}\exp\left(\left(f_{v,i}^l\right)^T f_{t,j}^l\right)}, \tag{11}$$

To make the semantics on the global feature channel consistent as well, we concatenate together different local features on the channel in a specific order to form semantically comprehensive global features $f_{v,i}^g$ and $f_{t,j}^g$, and use the following loss function to optimize the network parameters.

$$\begin{aligned}\boldsymbol{L}_{cmpm}^g\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = L_{v2t}^g\left(E_v, E_c, W_l, W_p, W_f, W_a\right) \\ + L_{t2v}^g\left(E_v, E_c, W_l, W_p, W_f, W_a\right),\end{aligned} \tag{12}$$

where L_{v2t}^g and L_{t2v}^g can be similarly obtained from Eq. 10 Throughout the training process, the final loss function of the model can be expressed as:

$$\begin{aligned}\boldsymbol{L}\left(E_v, E_c, W_l, W_p, W_f, W_a\right) = L_{cmpm}^g\left(E_v, E_c, W_l, W_p, W_f, W_a\right) \\ + \lambda_2 L_{cmpm}^2\left(E_v, E_c, W_l, W_p, W_f, W_a\right) \\ + \lambda_1 L_{cmpm}^1\left(E_v\right),\end{aligned} \tag{13}$$

where, λ_1 and λ_2 are used as parameters to balance the importance of different modules.

4 Experiments

4.1 Datasets and evaluation protocols

To verify the effectiveness of our proposed algorithm, we demonstrate its performance on two challenging datasets CUHK-PEDES Li et al. (2017a) and RSTPReid Zhu et al. [32].

CUHK-PEDES: This dataset is the first publicly available dataset for this task. We adopted the same data partitioning strategy as Chen et al. [33], where the dataset was divided into training, validation, and testing sets. The training set contains 11,003 individuals with a total of 34,054 images and 68,126 textual descriptions. Some sample images and text descriptions are shown in Figure 3, Figure 4. The validation set contains 1,000 individuals with 3,078 images and 6,158 textual descriptions, while the testing set contains 1,000 individuals with 3,074 images and 6,156 textual descriptions.

RSTPReid: This dataset is the latest public dataset. This dataset contains 4101 pedestrians with different identities, each with five different images, resulting in a total of 20,505 person images, with two textual descriptions per image. Following the data partitioning strategy in Zhu et al. [32], we divided this dataset into training, validation, and testing sets, where the training set contains 18,505 images from 3,701 individuals, the validation set contains 1,000 images from 200 individuals, and the testing set contains 1,000 images from 200 individuals. Similar to existing methods, we employ the Cumulative Match Characteristic metric to evaluate the performance of our model.

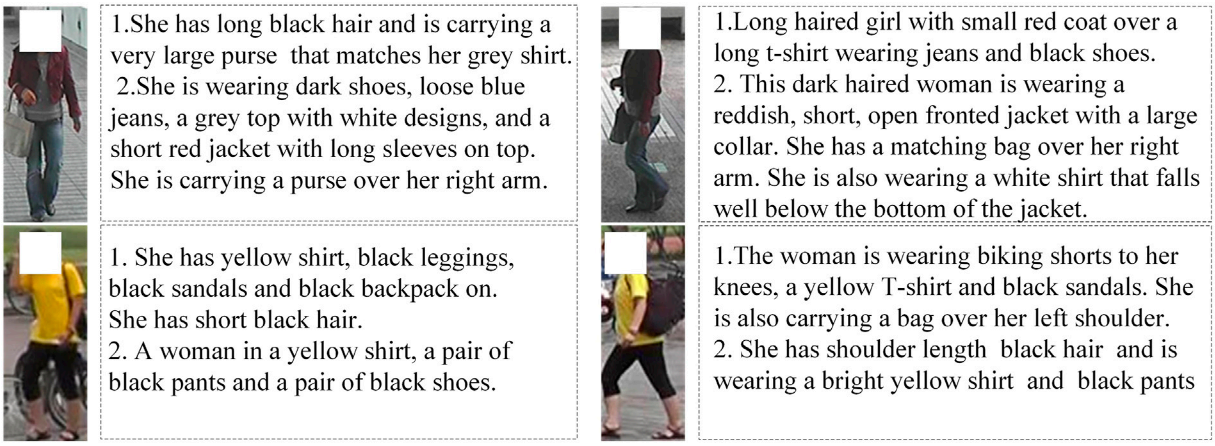
4.2 Implementation details

Our network is mainly composed of image feature extractor and text feature extractor. As with the other methods, we use ResNet-50 trained on imageNet Russakovsky et al. [34] and Bert as the backbone. The network was trained for 100 generations. Optimize network parameters using the Adam optimizer. Kingma and Ba [35]. The initial learning rate is set to 1 × 10^{-3}, and the warm-up strategy in Luo et al. [36] is used to adjust the learning rate for the first 10 epochs. At the 41st epoch, the learning rate is decayed to 10% of its current value. All images are resized to 384 × 128 × 3, and data augmentation is performed using random horizontal flipping. The batch size is set to 64, with each batch containing 64 image-text pairs. The text length is uniformly set to 64. During testing, cosine distance is used to measure the similarity between image-text pairs. The proposed model is implemented based on the PyTorch. All experiments are conducted on a single NVIDIA GeForce RTX3090 GPU device.

4.3 Comparison with state-of-the-art methods

4.3.1 Results on the CUHK-PEDES dataset

To illustrate the advantages of our method, we perform our method on the CUHK-PEDES dataset, and compare its performance with some state-of-the-art methods. The methods involved in the comparison include GNA-RNN Li et al. [29], GLA Chen et al. [33], CMPM + CMPC Zhang and Lu [31], MCCL Wang et al. [37], A-GANet Liu et al. [12], Dual-path Zheng et al. [38], MIA Niu et al. [19], PMA Jing et al. [39], TIMAM Sarafianos et al. [40], ViTAA Wang et al. [16], NAFS Gao et al. [20], DSSL Zhu et al. [32], MGEL Wang et al. [21], SSAN Ding et al. [10], TBPS(ResNet-50) Han et al. [41], and SUM Wang et al; [42]. The experimental results of different methods are shown in Table 1. It can be observed that the proposed method achieves a Rank-1 accuracy of 61.97 (%) and a



1.She has long black hair and is carrying a very large purse that matches her grey shirt.
2.She is wearing dark shoes, loose blue jeans, a grey top with white designs, and a short red jacket with long sleeves on top. She is carrying a purse over her right arm.

1.Long haired girl with small red coat over a long t-shirt wearing jeans and black shoes.
2. This dark haired woman is wearing a reddish, short, open fronted jacket with a large collar. She has a matching bag over her right arm. She is also wearing a white shirt that falls well below the bottom of the jacket.

1. She has yellow shirt, black leggings, black sandals and black backpack on. She has short black hair.
2. A woman in a yellow shirt, a pair of black pants and a pair of black shoes.

1.The woman is wearing biking shorts to her knees, a yellow T-shirt and black sandals. She is also carrying a bag over her left shoulder.
2. She has shoulder length black hair and is wearing a bright yellow shirt and black pants

FIGURE 4
Sample CUHK-PEDES dataset display.

Rank-5 accuracy of 81.01 (%), outperforming all the compared methods on the CUHK-PEDES dataset. In addition, it was found that the latest methods NAFS, SSAN and MGEL are far superior to other methods due to the use of attention mechanisms that allow the network to extract robust features adaptively. However, they do not consider the impact of the ambiguous semantic relationship between the textual and imaging descriptions of pedestrians on the matching performance, thus their performance is limited to some extent. Compared with the best-performing method TBPS in the compared methods, the proposed method achieves 0.32 (%) improvement in Rank-1 accuracy, which demonstrates the effectiveness and superiority of the proposed method over the compared methods.

4.3.2 Results on the RSTPReid dataset

In order to further verify the effectiveness of our method, we also conducted a comparative test on the RSTPReid dataset. Our proposed method is compared with five latest methods, namely, IMG-Net Wang et al. [44], AMEN Wang et al. [43], DSSl Zhu et al. [32], SSAN Ding et al. [10], and SUMWang et al.(2022c). As shown in Table 2, the latest method SSAN achieves the best performance with 43.50(%), 67.80(%) and 77.15 (%) accuracy for rank-1, rank-5 and rank-10, respectively. In contrast, the proposed method achieves significantly higher performance with 43.88(%), 76.60(%), and 80.20 (%) accuracy for rank-1, rank-5, and rank-10, respectively, exceeding the performance of SSAN. These experiments further validate the effectiveness of our method.

4.4 Ablation study

The proposed method in this paper mainly consists of two parts: Explicit Semantic Feature Extraction (ESFE), and Information Supplementation Network (ISN). In this paper, we use the model obtained by pre-training ResNet50 and Bert under the constraint of loss function as Baseline, and in order to verify the effectiveness of each module, different modules are added to Baseline gradually to observe the change of matching performance. In this process, the model obtained by adding ESFE to Baseline is named “Baseline + ESFE”; the model obtained by adding ISN to Baseline is named “Baseline + ISN″ The model after adding ISN to “Baseline + ESFE” is “Baseline + ESFE + ISN”. All experiments were conducted on the CUHK-PEDES dataset, and the experimental results are shown in Table 3.

4.4.1 The effectiveness of ESFE

In this paper, ESFE is mainly used to address the problem of semantic mismatch between textual features and their corresponding visual objects. As shown in Table 3, without using ESFE, the performance of the Baseline model on rank-1 accuracy is only 55.14 (%). When ESFE is added to the Baseline model, the performance of Baseline + ESFE is improved from 55.14 (%) to 58.42 (%), with an increase of 3.28 (%). This is mainly because ESFE can effectively address the issue of misalignment between features.

4.4.2 The effectiveness of ISN

To supplement local features, the ISN method is proposed in this paper. In this process, various local features are fused by self-attention mechanism to obtain comprehensive features. As shown in Table 3, without using ISN, the performance of the Baseline model on rank-1 accuracy is only 55.14 (%). When ISN is added to the Baseline model, the performance of Baseline + ISN is improved from 55.14 (%) to 59.20 (%), with an increase of 4.06 (%). This is mainly because ISN can effectively supplement missing information in features and improve the comprehensiveness of features.

4.4.3 The effectiveness of ESFE + ISN

Table 3 shows the effectiveness of adding ISN to Baseline + ESFE after rough alignment of local features. It can be seen that supplementing information between roughly aligned local features is more effective than directly supplementing information on the baseline. Rank-1 is improved from 59.20 (%) to 61.97 (%), with an increase of 2.77 (%). This indicates that supplementing information on relatively good features can result in more robust features.

TABLE 1 Comparative experiments on CUHK-PEDES dataset. Where the optimal results are shown in bold.

Methods	References	Rank-1	Rank-5	Rank-10
GNA-RNN	CVPR'17	19.05	—	53.64
Li et al. [29]				
GLA	ECCV'18	43.58	66.93	76.26
Chen et al. [33]				
CMPM + CMPC	ECCV'18	49.27	—	79.27
Zhang and Lu [31]				
MCCL	ICASSP'19	50.58	—	79.06
Wang et al. [37]				
A-GANet	ACM MM'19	53.14	74.03	81.95
Liu et al. [12]				
Dual-path	TOMM'20	44.4	66.26	75.07
Zheng et al. [38]				
MIA	TIP'20	53.10	75.00	82.9
Niu et al. [19]				
PMA	AAAI'20	53.81	73.54	81.23
Jing et al. [39]				
TIMAM	ICCV'20	54.51	77.56	84.78
Sarafianos et al. [40]				
ViTAA	ECCV'20	55.97	75.84	83.52
Wang et al. [16]				
NAFS	arXiv'21	59.94	79.86	86.7
Gao et al. [20]				
DSSL	ACMMM'21	59.98	80.41	87.56
Zhu et al. [32]				
MGEL	IJCAI'21	60.27	80.01	86.74
Wang et al. [21]				
SSAN	arXiv'21	61.37	80.15	86.73
Ding et al. [10]				
TBPS(ResNet-50)	arXiv'21	61.65	80.98	86.78
Han et al. [41]				
SUM	KBS'22	59.22	80.35	87.51
Wang et al. [42]				
Our(Proposed)	This paper	**61.97**	**81.01**	**87.82**

4.4.4 Ablation experiments Visualization

Figure 5 presents the effectiveness of each module. It can be observed from Figure 6 that the matching accuracy is improved when ESFE and ISN are added separately to the "Baseline", which demonstrates the effectiveness of the proposed ESFE and ISN. However, the best performance is not achieved, indicating that the current model cannot distinguish finer-grained features. When ISN is added to "Baseline + ESFE", it can be seen that information supplementation on the roughly aligned features can better explore finer-grained features. As shown in Figure 6, the model can not only distinguish large-scale features such as "A short-sleeved red top" and "Short skirt" but also better distinguish finer-grained features such as "Thick heel" and "White logo". This proves the effectiveness of the proposed "Baseline + ESFE + ISN". The above conclusions are consistent with those obtained from Table 3.

TABLE 2 Comparative experiments on RSTPReid dataset, and the best result is shown in bold.

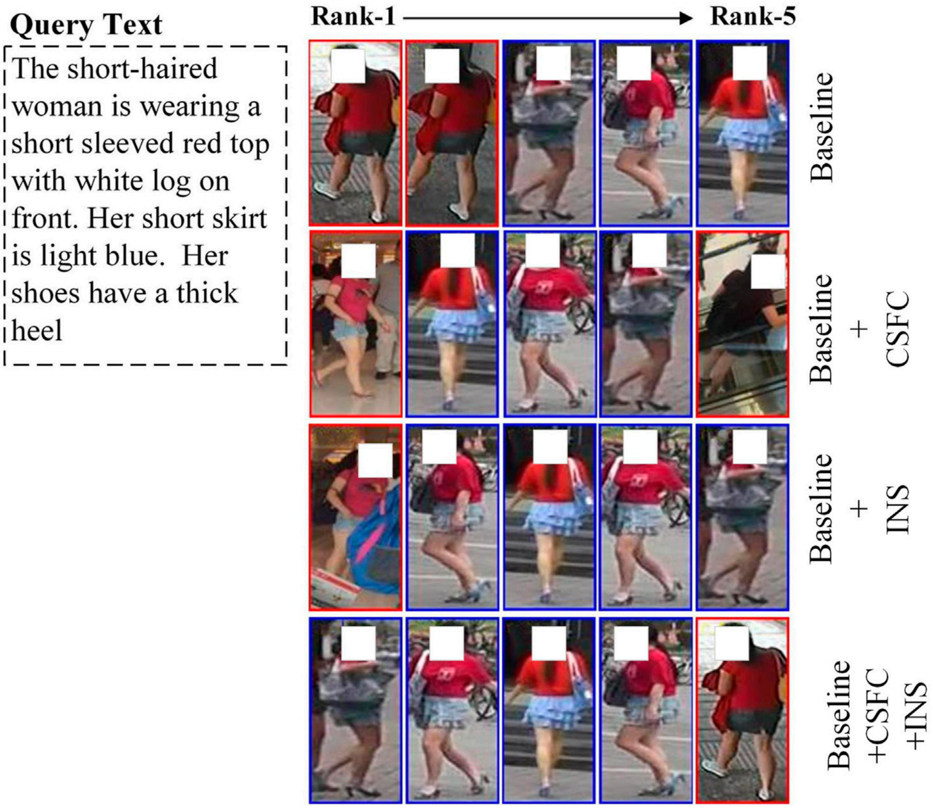
Methods	References	Rank-1	Rank-5	Rank-10
IMG-Net	JEI'20	37.60	61.15	73.55
Wang et al. [44]				
AMEN	PRCV'21	38.45	62.40	73.80
Wang et al. [43]				
DSSL	ACMMM'21	39.05	62.60	73.95
Zhu et al. [32]				
SSAN	arXiv'21	43.50	67.80	77.15
Ding et al. [10]				
SUM	KBS'22	41.38	67.48	76.48
Wang et al. [42]				
Our(Proposed)	This paper	**43.88**	**76.60**	**80.20**

TABLE 3 Ablation experiment.

Methods	Rank-1	Rank-5	Rank-10
Baseline	55.14	76.64	84.48
Baseline + ESFE	58.42	79.76	85.78
Baseline + ISN	59.20	79.29	85.63
Baseline + ESFE + ISN	61.97	81.01	87.82

4.4.5 Analysis of the loss function

Table 4 shows the effectiveness of the loss function. We find that the Rank-1 of L^1_{cmpm}+L^2_{cmpm} reaches 60.33 (%) and that of L^1_{cmpm}+L^g_{cmpm} reaches 60.12 (%). However, the Rank-1 of L^1_{cmpm} is only 58.42 (%), which we believe is because L^2_{cmpm} and L^g_{cmpm} can train the ISN network better and make the feature information more complete. The Rank-1 of L^1_{cmpm}+L^2_{cmpm}+L^g_{cmpm} reaches 61.97 (%) and the Rank-1 of L^1_{cmpm}+L^2_{cmpm} reaches 60.33 (%), which is 1.64 (%) higher, because L^g_{cmpm} constrains the global features and ensures the two modalities consistency of the global features between the two modalities. The Rank-1 of L^1_{cmpm} + L^g_{cmpm} is 60.12 (%), which is lower than the best result. This is because L^2_{cmpm}



Query Text

The short-haired woman is wearing a short sleeved red top with white log on front. Her short skirt is light blue. Her shoes have a thick heel

FIGURE 5
Visualize the retrieval results of baseline and our method. The image on the red edge indicates that the query is wrong, and the blue edge indicates that the query is correct.

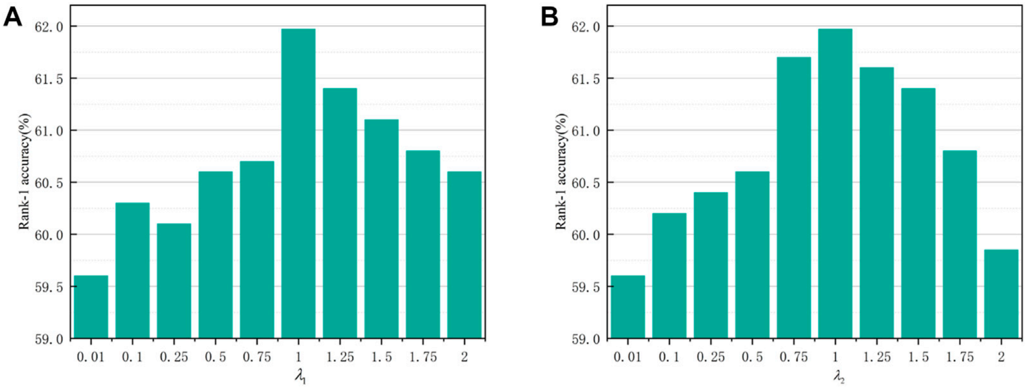


FIGURE 6
Effect analysis on hyperparameters.

constrains each local feature to ensure that each local feature has different semantics within the modality and consistent semantic information between the modalities. Thus L^1_{cmpm}+L^2_{cmpm}+L^g_{cmpm} not only ensure that the local features are discriminative, but also ensure that the global features are consistent. This also shows that it is reasonable for us to use L^1_{cmpm}+L^2_{cmpm}+L^g_{cmpm} to train the network.

4.5 Parameter selection and analysis

The three main hyperparameters involved in our approach are λ_1, λ_2 and N. In the parametric analysis, we fix two parameters to analyze the effect of another parameter on the results. All our experiments for the parameter analysis were performed on the CUHK-PEDES dataset.

TABLE 4 Analysis of the loss function.

Methods	Rank-1	Rank-5	Rank-10
L^1_{cmpm}	58.42	79.76	85.78
L^1_{cmpm}+L^2_{cmpm}	60.33	80.82	86.82
L^1_{cmpm}+L^g_{cmpm}	60.12	80.30	86.43
L^1_{cmpm}+L^2_{cmpm}+L^g_{cmpm}	61.97	81.01	87.82

The influence of λ_1. In Eq. 13, the hyperparameter λ_1 mainly regulates the role played by L^1_{cmpm}. This loss term is used to ensure the initial alignment of each local feature semantics. Figure 6A shows the effect on Rank-1 for different values of λ_1 in the CUHK-PEDES task. From this, we can find that there is an overall improvement in the Rank-1 recognition accuracy of our algorithm on the CUHK-PEDES task when $\lambda_1 \in [0.01, 1]$, the Rank-1 on CUHK-PEDES task decreases when $\lambda_1 \in [1, 2]$ Therefore, $\lambda_1 = 1$ is the optimal choice

The influence of In Eq. 13, the hyperparameter λ_2 mainly regulates the role played by L^2_{cmpm}. This loss term is used to ensure that the ISN can adaptively extract the relationship to each local feature and fuse the features in this way. We fix the hyperparameter $\lambda_1 = 1$, and λ_2 takes values in the range [0,2]. On CUHK-PEDES, the variation of Rank-1 for different values of λ_2 is shown in Figure 6B. It can be seen that when λ_2 is 1, the method in this paper can obtain the optimal performance on CUHK-PEDES, so it is reasonable to set λ_2 to 1.

The influence of *N*. In ESFE, for the image modality, we divide the image features into N local features with different semantics by PCB, and for the text modality, we generate N local features with different semantics by SA network. To verify the effect of different values of *N* on the model performance, we manually set *N* to 2, 3, 4, 6, 8, and12. From which we select the optimal N value for the model performance. This experiment was conducted on the CUHK-PEDES dataset. Table] 5 shows the experimental results of the effect of taking different values on the performance of the model. It can be seen that *N* of 6 achieves the best results.

5 Conclusion

This paper proposes a text-based framework for pedestrian image retrieval. Firstly, the ESFE method is utilized to provide clear semantic information for the text and achieve rough alignment between text and image features. In order to further enhance the representation of features, the ISN method is proposed to model the relationships among local features, fuse the features according to the underlying relationships. Finally global features are concatenated by refined local features. This improves the comprehensiveness of the features and effectively alleviates the matching difficulties caused by incomplete features. Compared with existing methods, the proposed model achieves good results on the CUHK-PEDES and RSTPReid datasets. Through ablation study, the contribution of different modules is investigated. The results show that this model is suitable for text-based pedestrian image retrieval. It is worth noting that in our study, sample diversity has a great impact on this task. For future work, we will study how to solve the problem of sample diversity.

TABLE 5 The influence of *N*.

N	Rank-1	Rank-5	Rank-10
2	59.14	78.44	86.78
3	60.04	78.89	86.91
4	60.33	79.75	87.44
6	61.97	81.01	87.82
8	61.53	80.93	87.32
10	61.11	79.61	87.14
12	60.47	78.01	86.92

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

HZ responsible for paper scheme design, experiment and paper writing. FL guide the paper scheme design and revision. YH guide to do experiments and write papers. XT guide the paper scheme design and revision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Fund Projects: Science and Technology Program of Science and Technology Department of Yunnan Province (Basic Research Special Project) (202101AT070136), National Natural Science Foundation of China (62161015), Key Science and Technology Special Project of Yunnan Province (202002AD080001).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Li H, Xu J, Yu Z, Luo J. Jointly learning commonality and specificity dictionaries for person re-identification. In: IEEE Transactions on Image Processing (2020). p. 7345–58.

2. Zhang L, Li K, Qi Y. Person re-identification with multi-features based on evolutionary algorithm. In: IEEE Transactions on Emerging Topics in Computational Intelligence (2021). p. 509–18.

3. Wang S, Liu R, Li H, Qi G, Yu Z. Occluded person re-identification via defending against attacks from obstacles. IEEE Transactions on Information Forensics and Security (2022).

4. Zhang Y, Wang Y, Li H, Li S. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia (ACMMM) (2022). p. 3347–55.

5. Li H, Yan S, Yu Z, Tao D. Attribute-identity embedding and self-supervised learning for scalable person re-identification. In: IEEE Transactions on Circuits and Systems for Video Technology (2020). p. 3472–85.

6. Zhu Z, Luo Y, Chen S, Qi G, Mazur N, Zhong C, et al. Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification. *J Vis Commun Image Representation* (2021) 80:103303. doi:10.1016/j.jvcir.2021.103303

7. Li H, Kuang Z, Yu Z, Luo J. Structure alignment of attributes and visual features for cross-dataset person re-identification. *Pattern Recognition* (2020) 106:107414. doi:10.1016/j.patcog.2020.107414

8. Lingli L, Minghong X, Fan L, Yafe Z, Huafeng L, Tingting T. *Unsupervised domain adaptive person re-identification guided by low-rank priori*. Chongqing, China: Chongqing University (2021).

9. Li H, Liu M, Hu Z, Nie F, Yu Z. Dupilumab use in non-atopic chronic hand eczema: Two cases and a review of the literature. *IEEE Trans Circuits Syst Video Technol* (2023) 1–3. doi:10.25259/IJDVL_721_2022

10. Ding Z, Ding C, Shao Z, Tao D. *Semantically self-aligned network for text-to-image part-aware person re-identification* (2021). *arXiv preprint arXiv:2107.12666*.

11. Zhang S, Long D, Gao Y, Gao L, Zhang Q, Niu K, et al. *Text-based person search in full images via semantic-driven proposal generation* (2021). *arXiv preprint arXiv:2109.12965*.

12. Liu J, Zha ZJ, Hong R, Wang M, Zhang Y. Deep adversarial graph attention convolution network for text-based person search. In: Proceedings of the 27th ACM International Conference on Multimedia (ACMMM) (2019). p. 665–73.

13. Zheng K, Liu W, Liu J, Zha ZJ, Mei T. Hierarchical gumbel attention network for text-based person search. In: Proceedings of the 28th ACM International Conference on Multimedia (ACMMM) (2020). p. 3441–9.

14. Chen Y, Zhang G, Lu Y, Wang Z, Zheng Y. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing* (2022) 494: 171–81. doi:10.1016/j.neucom.2022.04.081

15. Jing Y, Si C, Wang J, Wang W, Wang L, Tan T. *Pose-guided joint global and attentive local matching network for text-based person search*. New York, NY: AAAI Conference on Artificial Intelligence (AAAI) (2020).

16. Wang Z, Fang Z, Wang J, Yang Y. Vitaa: Visual-textual attributes alignment in person search by natural language. *European conference on computer vision (ECCV)*. Springer) (2020). p. 402–20.

17. Aggarwal S, Radhakrishnan VB, Chakraborty A. Text-based person search via attribute-aided matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020). p. 2617–25.

18. Wang C, Luo Z, Lin Y, Li S. Improving embedding learning by virtual attribute decoupling for text-based person search. *Neural Comput Appl* (2022) 34:5625–47. doi:10.1007/s00521-021-06734-9

19. Niu K, Huang Y, Ouyang W, Wang L. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Trans Image Process* (2020) 29:5542–56. doi:10.1109/tip.2020.2984883

20. Gao C, Cai G, Jiang X, Zheng F, Zhang J, Gong Y, et al. *Contextual non-local alignment over full-scale representation for text-based person search* (2021). *arXiv preprint arXiv:2101*.03036.

21. Wang C, Luo Z, Lin Y, Li S. Text-based person search via multi-granularity embedding learning. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI) (2021). p. 1068–74.

22. Li S, Cao M, Zhang M. *ICASSP 2022-2022 IEEE international conference on acoustics*. Speech and Signal Processing (ICASSP) (IEEE) (2022). p. 2724–8.Learning semantic-aligned feature representation for text-based person search.

23. Li H, Chen Y, Tao D, Yu Z, Qi G. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Trans Inf Forensics Security* (2021) 16:1480–94. doi:10.1109/tifs.2020.3036800

24. Liu X, Cheung YM, Hu Z, He Y, Zhong B. Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval. *IEEE Trans Emerging Top Comput Intelligence* (2020) 5:607–19. doi:10.1109/tetci.2020.3007143

25. Tang H, Li Z, Peng Z, Tang J. Blockmix: Meta regularization and self-calibrated inference for metric-based meta-learning. In: Proceedings of the 28th ACM international conference on multimedia (2020). p. 610–8.

26. Tang H, Yuan C, Li Z, Tang J. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition* (2022) 130:108792. doi:10.1016/j.patcog.2022.108792

27. Zha Z, Tang H, Sun Y, Tang J. Boosting few-shot fine-grained recognition with background suppression and foreground alignment. In: IEEE Transactions on Circuits and Systems for Video Technology (2023).

28. Li Z, Tang H, Peng Z, Qi GJ, Tang J. Knowledge-guided semantic transfer network for few-shot image recognition. In: IEEE Transactions on Neural Networks and Learning Systems (2023).

29. Li S, Xiao T, Li H, Zhou B, Yue D, Wang X. Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017). p. 1970–9.

30. Li S, Xiao T, Li H, Yang W, Wang X. Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017). p. 1890–9.

31. Zhang Y, Lu H. Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018). p. 686–701.

32. Zhu A, Wang Z, Li Y, Wan X, Jin J, Wang T, et al. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia (ACMMM) (2021). p. 209–17.

33. Chen D, Li H, Liu X, Shen Y, Shao J, Yuan Z, et al. Improving deep visual representation for person re-identification by global and local image-language association. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018). p. 54–70.

34. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis* (2015) 115:211–52. doi:10.1007/s11263-015-0816-y

35. Kingma DP, Ba J. *Adam: A method for stochastic optimization* (2014). *arXiv preprint arXiv:1412.6980*.

36. Luo H, Gu Y, Liao X, Lai S, Jiang W. Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019). p. 0.

37. Wang Y, Bo C, Wang D, Wang S, Qi Y, Lu H. Language person search with mutually connected classification loss. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE) (2019). p. 2057–61.

38. Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen YD. Dual-path convolutional image-text embeddings with instance loss. *ACM Trans Multimedia Comput Commun Appl (Tomm)* (2020) 16:1–23. doi:10.1145/3383184

39. Jing Y, Si C, Wang J, Wang W, Wang L, Tan T. Pose-guided multi-granularity attention network for text-based person search. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 34 (2020). p. 11189–96. doi:10.1609/aaai.v34i07.6777

40. Sarafianos N, Xu X, Kakadiaris IA. Adversarial representation learning for text-to-image matching. In: Proceedings of The IEEE/CVF International Conference on Computer Vision (ICCV) (2019). p. 5814–24.

41. Han X, He S, Zhang L, Xiang T. *Text-based person search with limited data* (2021). *arXiv preprint arXiv:2110.10807*.

42. Wang Z, Zhu A, Xue J, Jiang D, Liu C, Li Y, et al. Sum: Serialized updating and matching for text-based person retrieval. *Knowledge-Based Syst* (2022) 248:108891. doi:10.1016/j.knosys.2022.108891

43. Wang Z, Xue J, Zhu A, Li Y, Zhang M, Zhong C. Amen: Adversarial multi-space embedding network for text-based person re-identification. In: *Chinese conference on pattern recognition and computer vision (PRCV)*. Springer) (2021). p. 462–73.

44. Wang Z, Zhu A, Zheng Z, Jin J, Xue Z, Hua G. Img-net: Inner-cross-modal attentional multigranular network for descriptionbased person re-identification. *J Electron Imaging* (2020) 29:043028. doi:10.1117/1.jei.29.4.043028