Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 10, No. 3, 2024

A Review of Text-Based Pedestrian Retrieval Methods

Xiangmian Qiu, Yali Zhang *, Yichen Zhao, Jinzhao Li

School of Information Cyber Security, People's Public Security University of China, Beijing, China
*** Corresponding author:** Yali Zhang (Email: zhangyl_mail@163.com)

Abstract: Text-based pedestrian retrieval task uses textual descriptions as query inputs to retrieve pedestrians in image gallery, which is crucial for social security and investigation. By combing the relevant working literature, we summarize the main methods in this field, which are grouped into five categories of methods based on feature matching, based on multi-granularity information, based on adversarial ideal, based on cross-modal attention, and based on visual text pre-training models, and this paper compare and analyze the design ideas, method features, advantages and disadvantages of classical model of each category. The performance of each model is compared using the commonly used datasets and evaluation metrics (TPR, mAP) for this task, the problems faced in field are discussed and future development trends are envisioned. Recently, the method based on VLP has become a mainstream, which can achieve higher precision retrieval, but still faces the problems of large number of model parameters and high difficulty of training, so it needs to explore the lightweight solution in the future.

Keywords: Pedestrian Retrieval; Cross-modal Retrieval; Image Text Matching; Visual Language Pre-training Models.

1. Introduction

Text-based pedestrian retrieval is closely related to the tasks of pedestrian re-identification and image retrieval. Practically, we face the problem of lack of target pedestrians' image or pedestrians real-time clothing does not match the image gallery. In contrast, text-based retrieval methods have a greater potential for application because they use more accessible text as query. To better understand the content of this paper, we list the definition of pedestrian re-identification, person search and text-based pedestrian retrieval, and the main differences can be seen in Table 1, the process shown in Figure 1.

Table 1. Differences between Pedestrian Re-identification, Person Search and Text-based Pedestrian Retrieval Tasks

Task	Query Input	Application scene	Goal
Pedestrian re-identification	imagery	Cross-device, cross-scene	Pedestrian Tracks
Pedestrian Search	imagery	Single device, single scene	Specific pedestrian locations
Text-based Pedestrian Retrieval	copies	Cross-device, cross-scene	Pedestrian location and tracking

1) Pedestrian Re-ID (Pedestrian Re-ID): Aims at articulating the tracking of target pedestrian in different monitoring areas and Reid pedestrian tracking across time, location and devices. Its input query is generally an image.

2) Pedestrian Search (Person Search): It is to retrieve and match a target pedestrian in a complex scene given a target pedestrian image, using image as a query.

3) Text-based Pedestrian Retrieval (Text-based Pedestrian Retrieval): The target pedestrian is retrieved from the image gallery by textual description of the pedestrian.

In this paper, we have comprehensively sorted out the research related to text-based pedestrian retrieval technology since 2017, summarized the current research status at home and abroad, explored the main challenges and difficulties faced by text-based pedestrian retrieval technology, analyzed the advantages and disadvantages of various methods, and looked forward to the future development trend, and the main contributions of this paper:

(1) The text-based pedestrian retrieval methods are systematically sorted out and summarized into five categories: global matching, multi-granularity information, adversarial learning, cross-modal attention mechanism, visual-textual pre-training model, and the technical routes, model structures, advantages and disadvantages of each type of methods are analyzed.

(2) Based on the CUHK-PEDES, ICFG-PEDES and RSTPReid datasets, the performance of the classical models was compared using TPR.

(3) The future development trend of text-based pedestrian retrieval technology is envisioned to provide reference for subsequent research.

(a) Pedestrian re-identification

(b)Person search

(c)Text Based Pedestrian Retrieval

Fig 1. Brief schematic diagram of pedestrian re-identification, pedestrian search, and Text Based Pedestrian Retrieval process

2. Overview of the Development of Text-Based Pedestrian Retrieval

The core challenge of text-based pedestrian retrieval is how to effectively bridge the modal differences between images and texts. In this paper, we categorize the existing methods into the following types: based on global feature matching, based on multi-granularity information, based on adversarial learning, based on cross-modal attention mechanism, and based on visual-text pre-training model.

1) Global feature matching-based approach

The core idea of this type of method is to use CNN and RNN to extract image and text features respectively, map the extracted global features into a common feature space, bring similar features closer, and push away the features with larger differences. CMPC+CMPM uses ResNet-50 and LSTM network to extract image features, and designs the CMPM loss function to map the cross-modal features into the KL scatter chain to bridge the modal gap; in addition, CMPC projects the features onto the other modality to strengthen the image-text pairs and the CMPM loss function map the cross-modal features into the KL scatter chain to bridge the modal gap; in addition, CMPC projects the features onto another modality to strengthen the intrinsic connection between the image and text pairs.

2) Methods based on multi-granularity information

The main idea is to slice the image horizontally divide into fine-grained regions such as head, upper body, lower body, foot, hand, and so on, and then realize the level-by-level fine-grained alignment of the image patch with words, image regions with phrases, and the whole image with text sentences, this type of method can effectively improve the retrieval accuracy. For example, ViTAA segments the image according to specific pedestrian attributes through the semantic segmentation layer of the lightweight MaskHead architecture to form a fine-grained segmentation map; the text part is parsed and extracted by Stanford POS tagger to extract words and phrases related to the attributes, and then LSTM is used to extract the global and local textual features, and a fine-grained segmentation map is made by establishing correspondences between words, phrases and segmentation maps. The fine-grained image feature alignment is performed by establishing the correspondence between words, phrases and segmentation graphs.

3) Methods based on adversarial learning

Since GAN (Generative Adversarial Network) was proposed, it has received a lot of attention. Researchers in the field of text-based pedestrian retrieval have also borrowed the adversarial idea of GAN and designed models such as A-GANet and TIMAM. The core idea of this class of methods is to use generators to generate features that are as simulated as possible, while the discriminator determines the modal class of the input features and identifies whether they are generated by the generator, with the main advantage that the model is simple in structure and does not require additional modules or plug-ins.

4) Methods based on cross-modal attention mechanisms

Representative models include DSSL, NAFS, SAF, ACSA, etc. The core idea of such methods is to pay attention to each pixel in the feature image through the multi-head self-attention mechanism, and focus the attention on more important regions, extract more fine-grained features, and calculate the similarity between each part of the image and the text. Therefore, the attention mechanism approach can further improve the retrieval accuracy. Just as the ACSA approach uses Swin Transformer and BERT to extract image features separately, Swin Transformer hierarchically focuses on each pixel point in the image and understands the textual contextual semantic information through BERT to perform deeply alignment. However, since each pixel point is involved in the computation, such methods have higher computational and slower retrieval speed.

5) Methods based on visual text pre-training models

The main idea of such methods is to utilize pre-trained VLP models, such as the IRRA model using CLIP as the backbone network, to achieve efficient alignment of image features with the help of the VLP model's strong cross-modal learning ability, so that the model understands in depth that the image and text pairs contain semantic information, which makes it easy for the model to make the correct prediction in the retrieval. Carrying out such methods to become the mainstream methods in the field, significantly improving the retrieval accuracy in the field.

3. Dataset and Evaluation Metrics

3.1. Dataset

Text-based pedestrian retrieval is an emerging research problem, and the commonly used datasets in this area are relatively limited, mainly including CUHK-PEDES dataset[1] , ICFG-PEDES dataset[2] and RSTPReid dataset [3] . Table 2 details the information about these three datasets.

Table 2. Common datasets for text-based pedestrian retrieval

Dataset	Number of pedestrians	Number of images	Number of texts	Average length of text
CHUK-PEDES	13003	40206	80440	23.5
IGFG-PEDES	4102	54522	54522	37.2
RSTPReid	4101	20505	41010	23

3.2. Evaluation Metrics

3.2.1. TPR

True Positive Rate (TPR) or Recall indicates the proportion of positive samples to the total positive samples in the retrieval results, reflecting the number of samples that can be successfully retrieved by the model among all the positive samples. The TRP value can be calculated by using equation (1).

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

where TP is the sample that is positive and is predicted to be positive, and FN represents the sample that is negative but is incorrectly predicted to be negative by the model.

3.2.2. mAP

The mean accuracy value mAP is the average of the maximum accuracy at different recall rates and is commonly

used to measure the performance of the target detector. mAP value can be calculated by using equation (2).

$$mAP = \frac{1}{S}\sum_{i=1}^{S} AP_s \tag{2}$$

where S represents the total number of samples and APS represents the average accuracy of the prediction for the sample set S.

4. Methods of Text-based Pedestrian Retrieval

Text-based pedestrian retrieval aims to retrieve images of specific pedestrians through textual descriptions, and its core challenge lies in bridging the inherent gap between the two modalities and realizing cross-modal alignment between image features and text features. In this paper, text-based pedestrian retrieval methods are categorized as follows.

4.1. Method based on Global Feature Matching

Li et al.[1] first proposed text-based pedestrian retrieval as a cross-modal task, and designed GNA-RNN, an RNN network framework with a gate-attention mechanism, which is shown in Fig. 2. Li et al. designed GNA-RNN by dividing the network into a visual sub-network based on the VGG-16 and a textual sub-network based on the LSTM network[4] .

In the visual subnet two 512-dimensional fully connected layers have added for generating visual units on top of the drop7 layer of VGG-16 to recognize the presence or absence of a specific pedestrian appearance pattern in the pedestrian images by visual units. The visual subnet is first pre-trained on the CHUK-PEDES dataset assembled by the authors, then only the parameters of the added fully connected layer are updated during the joint training with the text subnet, and regions with higher similarity to the text are given higher weights; meanwhile, a scalar gate mechanism is designed according to the trait that different words have different contribution for image-text matching, so that the text subnet learns the weights in the word dimensions, measures the contribution of different words' contribution and focus on words with high contribution.

However, GNA-RNN also has many limitations. First, GNA-RNN processes text word by word via LSTM, ignoring the continuity between words, and is insensitive to the spatial location of key pedestrian attributes. For example, a pedestrian image containing the phrase "wearing a red jacket and white short sleeves" may have a higher similarity to the text describing "red short sleeves". Secondly, the graphical similarity is obtained by simply adding the similarity of each word and image, and the length of the sentence will significantly affect the final similarity score. In addition, the Top 1 accuracy of GNA-RNN is only 19.05, needs to be improved.

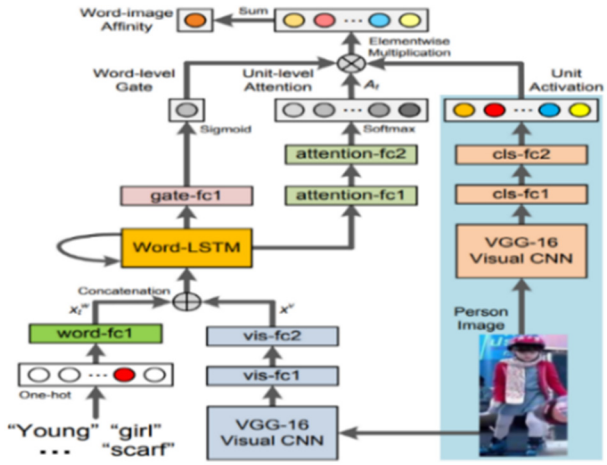

Fig 2. Framework of GNA-RNN

After GNA-RNN, Li et al.[5] noticed that the labeling information with pedestrian identities was underutilized and proposed a two-stage network for IATV with identity-awareness capability. The first stage network of IATV consists of a CNN-LSTM network, which is responsible for the extraction of graphic features and initial alignment. In this stage, Li et al. designed a new cross-modal cross-entropy loss function CMCE, which introduces a feature buffer mechanism to simplify the screening process of difficult negative samples and enables the model to efficiently calculate the similarity between samples and other identities, providing the basis for the second stage network. In the second stage, Li et al. introduced a potential collaborative attention mechanism and a spatial attention mechanism for aligning texts with different sentence structures and reducing the impact of sentence differences on the model.

The introduction of the CMCE loss function significantly advances the development of text-based pedestrian retrieval tasks. However, IATV uses a two-stage network framework in pursuit of higher accuracy, and the training process needs to be carried out separately, while the addition of multiple attention mechanisms leads to a significant increase in the model structure and computational complexity, which prevents end-to-end training.

Chen et al.[6] proposed an improvement idea to address the limitations of other methods in the field, especially GNA-RNN. Chen et al. proposed an adaptive thresholding image block and word matching network architecture, PWM (Patch-Word Matching). When performing image block and word matching, PWM assigns the highest matching score to an image block only when two semantically consecutive key words correspond to that image block at the same time, thus effectively solving the problem of spatial location insensitivity of the key attributes of the characters in the text processed by the GNA-RNN model.

The adaptive threshold mechanism predicts a matching threshold for each word in advance, and when the matching score is lower than the threshold, it is determined as a negative

sample, and no operation is performed; when the score is higher than the threshold, it determines that the image block matches the word, and automatically adjusts the matching score to the vicinity of the threshold, to avoid erroneous judgment caused by the matching score of the image block and the word being too high.

PWM overcomes the problems in the GNA-RNN method by virtue of the adaptive thresholding mechanism of the continuous matching mechanism of the image and text and adopts a hierarchical computational approach to compute the overall matching scores of the image and text layer by layer, which significantly improves the SOTA (State-of-the-Art) level of the text-based pedestrian retrieval task and pushes forward the further development of the field.

Zhang et al.[7] agree with IATV that individual identity information must be fully utilized, but the CMCE loss function proposed by Li et al. requires additional feature buffer allocation, which leads to high memory consumption when processing a large number of targets. To solve this problem, Zhang et al. proposed the cross-modal mapping matching loss function CMPM and the cross-modal mapping classification loss function CMPC. the network schematic is shown in Fig. 3.

The visual part is based on MobileNet as the backbone network and the text part is a bidirectional LSTM network. Unlike previous work after inputting the images, CMPM considers all the candidate texts in a small batch (mini batch) and incorporates them into the computation of the classification loss function, avoiding the cumbersome sampling process in the traditional bidirectional LSTM model. For the classification task with identity annotation information, CMPC projects the feature vectors onto the matching features in another modality and then classifies them, while integrating the similarity scores of the graphic sample pairs into the classification task to strengthen the connection between the graphic sample pairs. The CMPM+CMPC model demonstrates great stability and superiority, which significantly strengthens the match between the features and the individual categories. further advancing cross-modal learning in text-based pedestrian retrieval tasks.

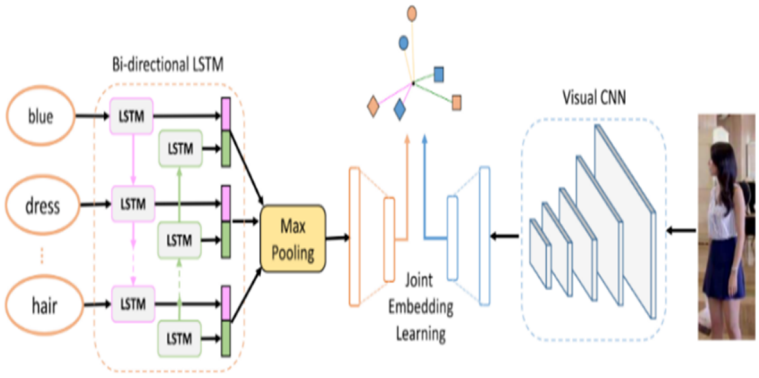

Fig 3. Schematic diagram of the CMPM+CMPC network framework

4.2. Method based on Multi-granularity

Jing et al.[8] proposed a multi-granularity textual pedestrian retrieval model PMA (Pose-Guided Multi-Granularity Attention Network) for fully exploiting human pose information. The architecture of the PMA model is shown in Fig. 4. The model adopts the PAF (Part Affinity Fields) method for human posture estimation and combines the posture confidence maps with the original images in the evaluation stage, so that the global visual features are more focused on the regions related to the human body. The role of the pose confidence maps is mainly reflected in two aspects: on the one hand, the confidence maps work together with the input images to enhance the representation of visual features; on the other hand, the confidence maps are used to learn the potential semantic alignment relationship between noun phrases and image regions.

PMA also contains two parts, the coarse-grained alignment network and the fine-grained network alignment network. The coarse-grained network selects image parts that are motivated by the highest correlation with the text parts, and the fine-grained network further learns the potential semantic links between noun phrases and image regions to extract the relevant image content in the phrase dimension. The PMA uses Triplet Ranking Loss and introduces the Identification Loss, which ensures that the pedestrian features can be grouped together to ensure the accurate matching at the individual level.

PMA enhances the potential semantic alignment of images and texts by fully utilizing the human posture information through the posture confidence map; learns the semantic information of different granularity in images and texts by designing coarse-grained and fine-grained alignment networks; and achieves high-precision retrieval by using Triplet Ranking and Identification Loss.

Niu et al.[9] argued that the key to cross-modal learning is adaptive local alignment and the potential relationship between local and global, and in order to address these issues, Niu et al. proposed a MIA model consisting of two parts: global and local feature extraction and graphic and textual alignment at multiple granularities. CNN networks and bidirectional gated recurrent units (Bi-GRUs) are utilized in global and local feature extraction to extract visual feature maps and encode sentences and phrases, respectively.

The multi-granularity graph alignment part contains three granularity alignment modules: the global comparison module (GC), the relation-guided global-local alignment module (RGA), and the bidirectional fine-grained matching module (BFM). In the RGA module, the feature map is uniformly segmented into multiple non-overlapping regions along the vertical direction[10] , and this segmentation method does not incur additional overhead, and is performed in parallel with the training process, which realizes end-to-end training. The MIA model is the first of its kind that

combines the local features of human body with noun phrases for the text-based pedestrian re-recognition task, which opens a new research direction in this field of textual pedestrian retrieval tasks. The MIA model is the first to combine human local features with noun phrases for the text-based pedestrian re-recognition task, opening a new research direction for the text-based pedestrian retrieval task in this area.

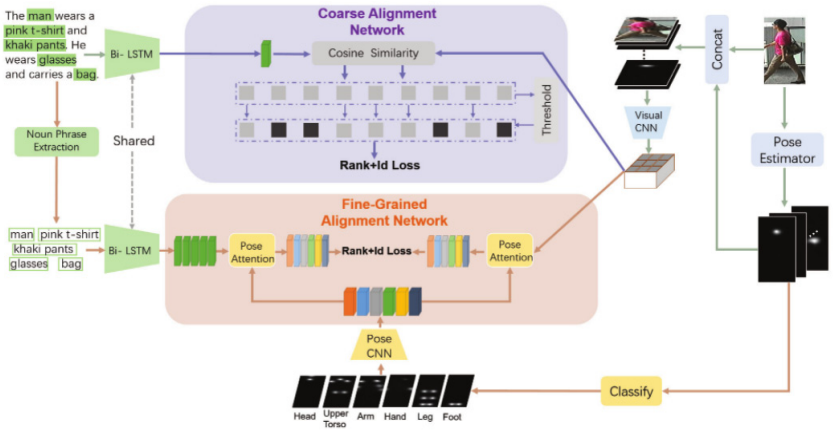

Fig 4. PMA network framework

Wang et al.[11] pointed out that previous approaches viewed the text-based pedestrian retrieval task as a holistic graphical feature matching. Wang et al. instead adopted an attribute-alignment-based perspective on the task. Based on this motivation, Wang et al. proposed the attribute feature alignment retrieval model ViTAA.

ViTAA contains a global feature branch and several local branches for generating global visual features and attribute visual features. Each local branch is also equipped with a lightweight MaskHead architecture to assist in generating segmentation maps for specific attribute categories. The text part first parses and extracts attribute-related noun phrases, then generates global and local features of the text through a bidirectional LSTM network, and further employs a contrast learning approach to achieve cross-modal alignment of the graph and text. Since the text description usually targets the appearance and dress of pedestrians, which are highly consistent with the attribute features, the mismatch due to the similarity of pedestrians' dress is effectively reduced. With these advantages, ViTAA significantly improves the SOTA scores on the CUHK- PEDES dataset.

Ding et al.[2] argued that semantic information alignment is very effective in bridging the gap between image and text modalities, and therefore proposed the Semantic Information Self-Aligning Network SSAN. For feature extraction SSAN uses ResNet-50 and bi-directional LSTM as a backbone network for extracting graphic and textual features. SSAN utilizes easy-to-align human body parts as a supervised acquisition of local features of the text, avoiding the introduction of additional tools to ensure the integrity of contextual links in the text part.

Meanwhile, to better mine the local connections in text descriptions and help the model adapt to text-to-figure pedestrian retrieval, SSAN introduces a multi-view non-local network MV-NLN. The loss function aspect uses Ding et al. to design Compound Ranking (CR) Loss as well as the commonly used ID Loss for local and global feature optimization.

Notably, Ding et al. concluded that the text annotations in the CUHK-PEDES dataset contain too much descriptive information that is not relevant to individuals, so Ding et al. created a new dataset, ICFG-PEDES. Ding et al. designed a new loss function while making full use of the semantic information and constructed a completely new dataset, which makes an excellent contribution to the text-based pedestrian retrieval field development with excellent contributions.

4.3. Methods based on Adversarial Learning

Liu et al.[12] first proposed a deep adversarial graph attention convolutional network framework, A-GANet, which consists of an image graph attention network, a text graph attention network, and an adversarial learning module.

The image graph attention network uses ResNet-50 as the backbone network, which contains five residual blocks, a visual scene graph module, and a joint embedding layer. Liu et al. designed Graph Attention Convolutional Layer (GAConv) for extracting structured semantic visual features. Finally, the joint embedding layer maps the virtual node feature values output from the visual scene graph module to the shared feature space by joining them with the global visual features.

The textual graph attention network consists of a bidirectional LSTM network, a textual scene graph module with a joint embedding layer. First bi-directional LSTM extracts text features and understands potential semantic dependencies between words, similar to the visual scene graph module, the text scene graph module detects objects present in the text and predicts the relationships between them, and at the same time, a graph attention convolutional layer is used to obtain structured text semantic features, and finally, the structured semantic text features and the initial text features are fused, which are mapped by the joint embedding layer into the shared feature space.

The adversarial learning module consists of a modal discriminator and a feature converter, which are used to optimize the features extracted by the image graph attention network and the text graph attention network, and the process is like that of GAN. Based on the training of the feature converter, the goal of the modal discriminator is to utilize the adversarial loss function to distinguish the sample modalities as much as possible, while the feature converter maps the graphic and textual modal features into a common feature space, where the recognition loss function is used to make the features discriminative, and the pairwise loss function is used to ensure that the graphic and textual features belonging to the same identity have a higher level of semantic similarity and

cross-modal invariance.

A-GANet fully mines the directed semantic scene graph to extract more discriminative graphical and textual features, which improves the retrieval accuracy without adding extra modules. The adversarial idea is firstly introduced into the field of text-based pedestrian retrieval, which provides new ideas for subsequent research. However, the introduction of graph attention mechanism increases the complexity of the training process, while the balance between modal discriminator and feature converter is harder to define.

In the same year, Sarafianos et al.[13] also proposed TIMAM, a network framework based on adversarial ideas. Sarafianos et al. argued that the lack of textual feature recognition makes the main reason for the poor performance of current computer vision methods in graphic matching tasks. For this reason, Sarafianos et al. introduced BERT[14] for the first time to the study of text-based pedestrian retrieval.

Compared with LSTM, a text feature extraction network commonly used in the field, BERT uses a large amount of data for unsupervised pre-training and can be easily migrated to various downstream tasks while possessing strong semantic understanding capabilities. Secondly, BERT adopts the attention mechanism, which can compute the relative weights of different positions in parallel, making it more efficient in processing long text. In addition, BERT adopts bi-directional encoding and a masking strategy during training, which allows it to understand contextual information in greater depth.

In addition, TIMAM introduces two loss functions for identification and cross-modal matching, i.e., Norm-SoftMax Cross Entropy Loss and CMPM Loss function proposed by Zhang et al. And the generalized features are further learned by GAN Loss to maximize the matching accuracy. The accuracy improvement brought by BERT is verified by ablation experiments. The Rank-1 accuracy comes from 49.85% to 52.97%. With the help of BERT and the adversarial representation learning paradigm, TIMAM refreshed the SOTA score, verifying the advantages of BERT over traditional LSTM networks. Since then, more and more researchers have begun to use BERT as a text feature encoder, which promotes the further development of the field.

4.4. Method based on Cross-modal Attention Mechanisms

Zhu et al.[3] argued that image information can be divided into character information and background information, which are mutually exclusive, while text descriptions are usually mainly related to character information. Therefore, algorithms that accurately distinguish character and background information and filter irrelevant words and syntactic noise in text are particularly important. To this end, the authors propose the Deep Background-Character Separation Model DSSL, which consists of the SDM module, the SPSM module, the SPFM module, the PDM module, and the SAM module.

DSSL first extracts global visual features V_G, and local visual feature matrix M_V by ResNet-50, and the text part uses bi-directional GRU to obtain global text features T_G and local text feature matrix M_T. The SDM module processes the global text features and the local feature matrix by setting a fixed number of elements in the input vectors to zero with a fixed nulling ratio, and then reconstructs the vectors in an auto-encoder manner through ternary permutation loss function optimization to obtain the text character feature vector T_P and the local text character feature matrix T_L, which fully retains the valid information while ignoring the redundant noise signals.

The SPSM module separates the character features from the background features in the image to obtain the character visual feature V_P and the environment visual feature V_S, while Zhu et al. design the mutual exclusion loss function according to the Mutually Exclusion Constraint (MEC) strategy to ensure that the character and the background do not overlap. The SPFM module, in the form of summation or concatenation, integrates the text feature T_P and background features V_S fusion, reconstructs the visual modal features V_R containing character and background information through the full connectivity layer MLP, and then aligns the rejoined visual modal features V_R with the visual global features V_G. Similarly, V_P is rewired as a textual modal feature T_R in the PDM module by means of a fully connected layer with a tanh activation function, *which* is aligned with the textual global feature.

SAM first strengthens the character information in the local visual feature matrix M_V to obtain V_L, and then aligns V_L with the textual character features T_G through the cross-modal attention mechanism to further enhance the recognition ability of the model. The fifth module, on the other hand, is like the fourth module, which mainly strengthens the character information in the local text feature matrix.

DSSL divides the text image into character part and background part, and through segmentation, the model focuses on the character information in the image and text, and filters irrelevant information and noise in the graphic text. However, DSSL requires multiple modules to differentiate characters from the environment and reconstruct graphic features, which is a cumbersome and complex process with high training overhead.

ACSA asymmetric cross-scale alignment includes a global alignment module and an asymmetric cross-attention module. The global alignment module uses the CMPM loss function proposed by Zhang et al.[7] as well as the CMPC loss function to perform global alignment of graphic and textual features and the alignment of images or local regions with text phrases to accomplish the first and second alignment, and the specific alignment process is similar to the method proposed by Zhang et al.[7] . The asymmetric cross-modal attention module is optimized by the asymmetric cross-scale alignment loss function, which is responsible for realizing the alignment between local phrases and local images, and contains two sets of inputs: the visual part connects the global image features with the regional features to obtain a set of multi-scale visual feature vectors as the input of the visual part; the text part takes a set of noun phrase feature vectors as the text input. The cross-modal attention mechanism is utilized to calculate and output the similarity scores between pairs of graphical and textual local features.

ACSA automatically learns multi-scale alignment through the cross-modal attention mechanism and is no longer limited to a specific scale range. ACSA further improves the multi-scale alignment mechanism and significantly improves the retrieval accuracy.

4.5. Methods based on VLP

Yan et al.[15] recognized the power of the CLIP model[16] in semantic learning and cross-modal knowledge extraction, and proposed CFine, a CLIP-based framework for fine-grained information extraction. The CLIP model is mainly

focused on individual-level features, which may lead to poor results when directly applied to new tasks. The CFine framework consists of two main components: a modality-specific feature extraction module and fine-grained information mining module.

In order to avoid intra-modal information distortion, the CFine model only uses the encoder of the CLIP model for the image part, while the text part directly uses the pre-trained BERT model. However, the graphical and textual features extracted by CLIP image encoder and BERT only contain individual dimensional modal information and cross-modal correspondences, but lack the crucial fine-grained information, so Yan et al. further designed the fine-grained information mining module to fully explore the fine-grained correspondences.

The fine-grained information mining module is divided into three sub-modules: first, the Multi-Granular Global Feature Learning (MGF) module is responsible for mining potential clues related to individual identity, executing the Token selection process according to the Transformer's strategy for text and images, respectively, and then inputting the Token into the global-local decoder, which utilizes the multi-head self-attention layer to make the Token information pass into the classification Token, after which the multi-head cross-attention mechanism intervenes to highlight the Token information and contextual information.

Cross-granularity feature optimization (CFR) module removes the non-modal sharing information and initially establishes the cross-modal linkage, firstly, the similarity between image-word and sentence-image blocks is evaluated by inner product, and the similarity of individual dimensions is obtained by fusion, and the weights are adaptively assigned by Softmax function during the process, and the cross-modal correspondences between the image and the text have been roughly established after passing through the CFR module.

The Fine-Grained Connection Discovery (FCD) module is used to establish fine-grained connections between modalities, after going through the MGF and CFR modules, the FCD module has made explicit the image blocks and text phrases that contain the most information, and the FCD calculates the pre-similarity between the two, and then the most relevant image or text features are selected using the Top-k operation, and the average pooling results in matching graphic and text feature pairs. To bring these graphic and text feature pairs closer together in the feature space, FCD again calculates the individual dimensional similarity by cosine similarity.

CFine avoids the problem of intra-modal information distortion that may arise from the direct use of the CLIP model, and does not introduce additional embedding layers, supports end-to-end optimization, and achieves SOTA scores on all three datasets, which verifies the advantages of the VLP model, and strengthens the foundation for the continued use of the VLP model in the domain. However, CFine only uses the image encoder of the CLIP model, which limits the cross-modal learning ability of the CLIP model and fails to give full play to the semantic understanding ability of the CLIP model through the pre-training of hundreds of millions of graphic datasets.

Jiang et al.[17] pointed out two major problems with previous approaches using the VLP model: one, the lack of necessary underlying alignment capabilities, which leads to insufficient efficiency in multi-modal data matching; and two, the reliance on too much a priori knowledge to explore explicit alignment patterns, which may trigger distortion of intra-modal information. To address these issues, Jiang et al. proposed the IRRA framework based on the CLIP model. Unlike previous approaches, IRRA utilizes the full CLIP model.

The IRR module employs MLM, a text masking strategy widely known for BERT, to optimize both image and context feature vector alignment and static vector consistency in the joint embedding space through MLM. To realize cross-modal interaction between images and texts, Jiang et al. designed a computationally more efficient multimodal interaction encoder. The encoder consists of a multi-head cross attention layer (MCA) with a four-layer Transformer block. For the input text, the text encoding features are masked according to the BERT's strategy MLM, and then the original features of the masked parts are predicted by the MLP layer classifier. Through self-attention and cross-modal attention mechanisms, textual and visual features are first fused at the cross-attention layer before being input into a single Transformer block. The multimodal interactive encoder establishes the link between image and text without additional supervision or inference overhead.

In order to explore more effective cross-modal matching, Jiang et al. designed the SDM Loss loss function to combine the cosine similarity distributions of image-text pair vectors into the KL dispersion to correlate the representations of the different modalities, and into the temperature hyperparameter to precisely control the compactness of the similarity distributions, so that the model focuses on the difficult negative samples, and at the same time, expands the variance between non-matching pairs and the correlation between matching pairs.

Compared to CFine, IRR uses the full CLIP, and the introduction of the MLM masking strategy enhances IRRA's understanding of contextual semantic information. IRRA significantly outperforms the previous SOTA model on several datasets, demonstrating superior performance.

Tan et al.[18] pointed out that the main challenge faced in the field is that the annotation process is too difficult, which leads to the small size of the existing dataset and is not conducive to deep training of models. Inspired by CLIP, Tan et al. investigated the problem of reloadability through multimodal large language models (MMLMs). First, Tan et al. prompted ChatGPT to generate multiple description templates, which prompted the MMLMs model to generate text descriptions with diversity based on the templates, solving the overfitting problem due to the single syntax.

Despite its power, the MMLMs model may still generate mismatched words when performing image naming tasks. Therefore, Tan et al. designed the Noise-Aware Masking (NAM) method, which calculates the similarity between each word and the image region, identifies the words with too little similarity, and masks them with a higher probability during training, which effectively reduces the noise generated by the automatic generation of textual descriptions by the MMLMs, while retaining the remaining useful information. Unlike traditional masked language model MMLM methods, NAM performs word masking based on similarity and does not need to re-predict the masked words.

Based on the above approach, Tan et al. selected the LUPerson dataset as the image source to generate the LUPerson-MMLM dataset containing one million images. Notably, using this large-scale dataset as well as the pre-training parameters of the previous models, the authors successfully reached SOTA scores on three benchmark

datasets. Tan et al.'s idea of generating a large-scale dataset using the MMLMs large language model remedies the problem of missing datasets in the field and significantly advances the field.

5. Conclusion

Various types of text-based pedestrian retrieval methods have their own advantages and disadvantages. The advantages and disadvantages of each type of methods are demonstrated and compared through Table 3:

Table 3. Comparison of advantages and disadvantages of various methods of text-based pedestrian retrieval techniques

Type	Advantage	Drawbacks
Global Feature Matching	Relatively simple model structure, low computational complexity, low retrieval overhead, suitable for end-to-end applications	Focus only on the global features in the graphic samples, not enough mining of fine-grained local information, relatively low retrieval accuracy
Multi-granularity information	Capable of capturing multi-granularity detail information, with strong detail learning capability, it can effectively improve the retrieval accuracy.	Additional modules are usually required to align information at different granularities, and the model structure is more complex, making it difficult to realize end-to-end training and real-time retrieval.
antagonistic thought	Competitive retrieval accuracy can be achieved while keeping the model structure relatively simple	Consistent with GAN, a balance between generator and discriminator is required, and definitional difficulties lead to difficulty in converging the model
Cross-modal attention mechanisms	Establish the connection between each word and phrase and pixel and image block to realize deep cross-modal alignment and significantly improve the model's cross-modal semantic understanding and retrieval accuracy	The increase in computational complexity and retrieval time makes it difficult for practical applications to bear such a heavy computational burden
VLP	With the powerful cross-modal learning and semantic understanding capability of VLP model, it realizes the significant improvement of retrieval accuracy	VLP models require large amounts of data and resources for training and fine-tuning, and are difficult and expensive to use and reproduce

The performance of the various methods on the CUHK-PEDES, IGFG-PEDES and RSTPReid datasets are compared in Table 4. Text-based pedestrian retrieval techniques are an emerging field that still faces many challenges:

1) Cross-modal feature alignment: cross-modal feature alignment is still a core problem in the field. It is difficult for existing methods to improve the retrieval accuracy while considering the computational efficiency and speed, and the balance between accuracy and speed has not been effectively solved in large-scale data retrieval tasks.

2) Pedestrian Occlusion Problem: Pedestrian occlusion is common in practical applications, and the occlusion problem has not been fully considered in the current domain. The lack of occlusion-related samples and datasets leads to the lack of reliability of the main methods in occlusion scenarios.

3) The problem of the number of model parameters: model lightweighting is another challenge in the field. Existing methods based on cross-modal attention mechanism and visual language pre-training models have significant improvement in accuracy, but their huge number of model parameters and extremely high requirements on hardware equipment make it difficult to retrieve them in real time in practical applications, which limits their practical deployment.

4) Dataset problem: In the privacy protection and other factors, it is very difficult to construct high-quality and large-scale datasets. Currently, it mainly relies on three datasets for model training and performance evaluation, which limits the diversity and innovation of algorithms in the field to some extent.

In summary, although text-based pedestrian retrieval technology has made some progress in the past few years, it still faces the problems of cross-modal feature pairs alignment, pedestrian occlusion, number of model parameters and datasets. Combining the existing research status and domestic and international research hotspots, this paper looks forward to the future development direction of text-based pedestrian retrieval technology and explores possible breakthrough paths.

1) Cross-modal alignment with the help of additional information and modules. To reduce the inherent differences between image and text modalities, the cross-modal learning ability of the model can be enhanced with the help of additional information to better achieve feature alignment can be explored. For example, the LapsCore method proposed by Wu et al[19] to design a color inference module can be directly attached to the mainstream baseline model, enabling the model to make full use of the color features of the pedestrian's clothing to improve the accuracy and efficiency of retrieval. In addition, behavioral and pose information can also provide strong support to the model, especially in scenarios where the target pedestrians are occluded by obstacles, and this information can significantly improve the robustness of the model. Combining these additional information, text-based pedestrian retrieval is expected to be a very promising development direction.

2) Constructing large-scale datasets. Compared with other popular fields, the number of text-based pedestrian retrieval datasets is small, which limits the further development of the field. Tan et al.[18] utilize the MMLMs large language model to automatically annotate the images to form a large-scale dataset; Wu et al.[20] construct datasets more in line with the existence of occlusion problems in real scenes by designing an occlusion module. Referring to previous studies, collecting relevant images from pedestrian re-identification and video

pedestrian retrieval datasets is a feasible expansion method that can further promote the progress of text-based pedestrian retrieval technology.

3) Lightweight design and improvement. The current mainstream methods, with huge computational overheads, are difficult to be directly applied to real-world scenarios. Future research should focus on the lightweight improvement of the model or designing new lightweight models to reduce the demand for computational resources and achieve more efficient deployment and retrieval.

4) Focus on the research hotspots in related fields. At present, many mainstream methods draw on the technology and experience in the field of graphic matching or pedestrian re-identification. It should continuously pay attention to the progress in related fields to discover new methods and ideas. VLP model, as a research hotspot in recent years, plays an important role in text-based pedestrian retrieval. Currently, CLIP model is mainly utilized as the backbone network in the field, which can be mined to be more suitable for the field of VLP model, e.g., Bai et al.[21] based on the ALBEF model proposed the RaSa model to further advance the domain development. Therefore, in-depth exploration of the application and optimization of VLP models is an important direction for future research.

Table 4. Comparison of scores on the CHUK-PEDES, ICFG-PEDES, and RSTPReid datasets

category	method	Image encoder	Text encoder	Rank-1			Rank-5			Rank-10		
				CHUK-PEDES	ICFG-PEDES	RSTPReid	CHUK-PEDES	ICFG-PEDES	RSTPReid	CHUK-PEDES	ICFG-PEDES	RSTPReid
global feature	GNA-RNN	VGG16	LSTM	19.05	N/A	N/A	N/A -	N/A	N/A	53.64	N/A	N/A
	IATV	VGG16	LSTM	25.94	N/A	N/A	N/A	N/A	N/A	60.48	N/A	N/A
	Dual-Path	ResNet-50	ResNet-50	N/A	38.99	N/A	N/A	59.44	N/A	N/A	68.41	N/A
	PWM	VGG16	LSTM	27.14	N/A	N/A	49.45	N/A	N/A	61.02	N/A	N/A
	CMPM+CMPC	MobileNet	LSTM	49.37	43.51	N/A	N/A	65.44	N/A	79.27	74.26	N/A
multi-granularity	PMA	VGG16/ ResNet-50	LSTM	53.81	N/A	N/A	73.54	N/A	N/A	81.23	N/A	N/A
	TVFR	MobileNet	LSTM	53.87	N/A	N/A	75.25	N/A	N/A	83.47	N/A	N/A
	MIA	ResNet-50	Bi-GRU	53.10	46.49	N/A	75.00	67.14	N/A	82.90	75.18	N/A
	ViTAA	ResNet-50	LSTM	55.97	50.98	N/A	75.84	68.79	N/A	83.52	75.78	N/A
	TIPCB	ResNet-50	BERT	63.63	54.96	N/A	82.82	74.72	N/A	89.01	81.89	N/A
	CAIBC	ResNet-50	Bi-GRU	64.43	N/A	47.35	82.87	N/A	69.55	88.37	N/A	79.00
	MANet	ResNet-50	BERT	65.64	59.44	N/A	83.01	76.80	N/A	88.78	82.75	N/A
	SSAN	ResNet-50	LSTM	61.37	54.23	43.50	80.15	72.63	67.80	86.73	79.53	77.15
adversarial learning	A-GANet	ResNet-50	LSTM	39.52	N/A	N/A	69.91	N/A	N/A	80.91	N/A	N/A
	TIMAM	ResNet-101	BERT	54.51	N/A	N/A	77.56	N/A	N/A	84.78	N/A	N/A
cross-modal attention	DSSL	ResNet-50	Bi-GRU	62.33	N/A	39.05	82.11	N/A	62.60	88.01	N/A	73.95
	NAFS	ResNet-50	BERT	61.50	N/A	N/A	81.19	N/A	N/A	87.51	N/A	N/A
	ACSA	Swin Transformer	BERT	68.67	N/A	N/A	85.61	N/A	N/A	90.66	N/A	N/A
	CCMF	ViT	BERT	73.81	N/A	57.35	88.89	N/A	77.50	92.77	N/A	85.50
VLP	PSLD	ResNet-50	CLIP	64.08	N/A	N/A	81.73	N/A	N/A	88.19	N/A	N/A
	CFine	CLIP	BERT	69.57	60.83	50.55	85.93	76.55	72.50	91.15	82.42	81.60
	CSKT	CLIP	CLIP	69.70	58.90	57.75	86.92	77.31	81.30	91.80	83.56	88.35
	IRRA	CLIP	CLIP	73.38	63.46	60.20	89.93	80.25	81.30	93.71	85.82	88.20
	RaSa	ALBEF	ALBEF	76.51	65.28	66.90	90.29	80.40	86.50	94.25	85.12	91.35
	AUL	Swin Transformer	BERT	77.23	69.16	71.65	90.43	83.32	87.55	94.41	88.37	92.05
	SAP-SAM	SAM-ViT	BERT	75.05	63.97	62.85	89.93	80.84	82.65	93.73	86.17	89.85

Note: Bolded text indicates optimal values, "N/A" indicates that the relevant data are not available.

Acknowledgments

This work is supported by The Double First-class Special Project of Security and Prevention Project of People's Public Security University of China (2023SYL08).

References

[1] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person Search with Natural Language Description," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21-26 July 2017 2017: IEEE, pp. 5187-5196, doi: 10.1109/CVPR.2017.551.

[2] Z. Ding, C. Ding, Z. Shao, and D. Tao, "Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification," p. arXiv:2107.12666doi: 10.48550/arXiv. 2107. 12666.

[3] A. Zhu et al., "DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval," p. arXiv: 2109. 05534 doi: 10.48550/arXiv.2109.05534.

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[5] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-Aware Textual-Visual Matching with Latent Co-attention," p. arXiv:1708.01988doi: 10.48550/arXiv.1708.01988.

[6] T. Chen, C. Xu, and J. Luo, "Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 12-15 March 2018 2018: IEEE, pp. 1879-1887, doi: 10.1109/WACV.2018.00208.

[7] Y. Zhang and H. Lu, "Deep Cross-Modal Projection Learning for Image-Text Matching," in European Conference on Computer Vision, Cham, 2018: Springer International Publishing, in Computer Vision – ECCV 2018, pp. 707-723, doi: 10.1007/978-3-030-01246-5_42.

[8] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 11189-11196, 04/03 2020, doi: 10.1609/aaai. v34i07.6777.

[9] K. Niu, Y. Huang, W. Ouyang, and L. Wang, "Improving Description-based Person Re-identification by Multi-granularity Image-text Alignments," p. arXiv:1906.09610doi: 10. 48550/arXiv.1906.09610.

[10] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline)," p. arXiv:1711.09349doi: 10.48550/arXiv.1711.09349.

[11] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language," p. arXiv:2005.07327doi: 10.48550/arXiv. 2005. 07327.

[12] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, "Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search," presented at the Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 2019. [Online]. Available: https://doi.org/10. 1145/ 3343031. 3350991.

[13] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial Representation Learning for Text-to-Image Matching," p. arXiv:1908.10534doi: 10.48550/arXiv.1908.10534.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," p. arXiv:1810.04805doi: 10.48550/arXiv. 1810. 04805.

[15] S. Yan, N. Dong, L. Zhang, and J. Tang, "CLIP-Driven Fine-grained Text-Image Person Re-identification," p. arXiv:2210. 10276doi: 10.48550/arXiv.2210.10276.

[16] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," p. arXiv:2103.00020doi: 10. 48550/ arXiv.2103.00020.

[17] D. Jiang and M. Ye, "Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval," p. arXiv: 2303.12501doi: 10.48550/arXiv.2303.12501.

[18] W. Tan, C. Ding, J. Jiang, F. Wang, Y. Zhan, and D. Tao, "Harnessing the Power of MLLMs for Transferable Text-to-Image Person ReID," p. arXiv:2405.04940doi: 10.48550/arXiv. 2405.04940.

[19] Y. Wu, Z. Yan, X. Han, G. Li, C. Zou, and S. Cui, "LapsCore: Language-guided Person Search via Color Reasoning," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 10-17 Oct. 2021 2021: IEEE, pp. 1604-1613, doi: 10.1109/ ICCV48922.2021.00165.

[20] X. Wu, W. Ma, D. Guo, T. Zhou, S. Zhao, and Z. Cai, "Text-Based Occluded Person Re-identification via Multi-Granularity Contrastive Consistency Learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 6, pp. 6162-6170, 03/24 2024, doi: 10.1609/aaai. v38i6.28433.

[21] Y. Bai et al., "RaSa: relation and sensitivity aware representation learning for text-based person search," presented at the Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, P.R.China, 2023. [Online]. Available: https://doi.org/ 10.24963/ ijcai. 2023/62.