A Review of Pedestrian Information Retrieval Research

Yan Xie[1] and Jian Qu[2*]
[1,2]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand
E-mail: 6572100049@stu.pim.ac.com, jianqu@pim.ac.com[*]

Received: January 4, 2024 / Revised: July 5, 2024 / Accepted: July 10, 2024

***Abstract*—Pedestrian Information Search (PIS) has gained attention for its wide range of practical applications. The main objective of PIS is to find a matching object in a set of scene images or videos. Early work on PIS focused on image-based search. With the advent of deep neural networks, PIS can be freed from the limitations of the search source. Therefore, a systematic study of PIS is necessary. In this paper, we review the research results of PIS based on different modalities in terms of the origin of the PIS task, the development history of PIS, and the methods of training and evaluation of PIS models. We selected the better-performing models for experiments. We summarize and comparatively evaluate the experimental results. Finally, we discuss some of the present problems of PIS and some meaningful future research directions.**

***Index Terms*—Pedestrian Information Search, Pedestrian Detection, Pedestrian Re-Identification, Deep Learning, Neural Network Models**

I. Introduction

Pedestrian Information Search (PIS) is an important and challenging task in computer vision, especially in human-targeted tasks. PIS aims to achieve an effective search for target pedestrian information in a variety of search scenarios. PIS has great potential for application in real-world search scenarios of surveillance videos. Therefore, this paper presents a comprehensive survey of work related to PIS.

PIS is an end-to-end technique for Pedestrian Detection (PD) [1] and Pedestrian Re-Identification (PReI) [2] in panoramic images. PIS needs to accurately derive the coordinate position information and identity information of pedestrians in the image. Considering the actual pedestrian information retrieval function, the PIS can be divided into the joint execution of PD and PReI tasks.

Since 2004, PD has received extensive attention and research [3]. PD, as a kind of target detection, mainly extracts features by manually designed feature extraction methods [4]. For example, the study in [5] achieved PD by designing a unified framework approach. However, the research of Enzweiler and Gavrila [6] showed that the accuracy of manual feature extraction is not high. Meanwhile, the emergence of deep neural networks has brought a new development direction for PD technology. Liu and Stathaki used CNN networks to complete the detection task in their research [7]. The research results of Zhai *et al.* [8] show that deep neural networks can greatly improve accuracy.

PReI was proposed for this task as early as 1996 [9]. However, PReI gained widespread attention after being reintroduced in 2006 at the CVPR (International Conference on Computer Vision and Pattern Recognition). PReI is a technique for determining the presence or absence of specific target pedestrian information in an image or a video sequence [10]. PReI has been reintroduced for two main reasons: 1) the lack of acquisition of pedestrian information in the research conducted at that time [11] and 2 the deep neural networks have been applied so that higher-level features can be acquired [12]. PReI can acquire more and deeper pedestrian information. In PD systems, pedestrian sample information is always ignored. Therefore, better accuracy and efficiency can be achieved by linking PD and PReI tasks [13]. Researchers have indicated that the combination of PD and PReI techniques can indeed enhance the performance of pedestrian information retrieval [14]. Unlike PD and PReI, the main challenge of PIS is to query the gap between people. PIS needs to deal with extra details. PIS can be categorized into Image-based Pedestrian Information Search (IPIS) and Natural Language-based Pedestrian Information Search (NLPIS) according to the search source.

Image-based Pedestrian Information Search (IPIS) is performed using the detection image as the search source. IPIS benefits from the fact that there is the source for conducting the search is explicit. In the study of Sun *et al.* [15], they successfully implemented pedestrian information retrieval using images by training a CNN model. However, IPIS suffers from the limitation of search sources [15]. IPIS

cannot be applied in many scenarios. When the search source image is not available, free-form natural language-based character search is very convenient [16].

NLPIS is another major PIS category that uses free-form natural language as a search query. NLPIS is more challenging than the IPIS problem. In practice, the effects of factors such as morphology, occlusion, resolution, and background responsibility can make the PIS task more challenging [17]. Therefore, NLPIS requires that discriminative features need to be learned first before text character matching.

The feature extraction methods of NLPIS are mainly divided into manual feature-based methods and deep feature-based methods. For handcrafted feature-based pedestrian information retrieval methods, a common approach will use handcrafted features such as Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Color Histograms [18] to obtain information such as shapes, colors, and textures. Hand-designed features are usually interpretable and understandable because they are constructed based on domain expertise. However, hand-designed features are relatively sensitive to changes and deformations in the data and may be less adaptable to changes in lighting, viewing angle, and background. For deep feature-based approaches, use deep learning models to learn higher-level feature representations from images. The deep features can be selected from the deep learning models suitable for pedestrian information retrieval. Deep features are acquired with more freedom compared to manual features [19]. Deep neural networks commonly used for deep features are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) [20], etc. The application of deep neural networks also foresees the need for data preprocessing and training of deep neural network models. The pre-processing workload for deep feature extraction is relatively high.

While manual feature-based methods are still effective in some scenarios [21], manual feature-based pedestrian information retrieval methods do not perform well when dealing with complex scenarios and highly variable conditions [22]. Manually designed features cannot capture all the variations and information in the data. Deep learning models can automatically learn the features in the data to better adapt to different scenes and changes. The rise of deep learning methods in recent years has gradually replaced some of the manual feature-based methods. The advantages of deep learning methods are more significant when dealing with large-scale and complex datasets [23]. Overall, deep feature-based PIS methods have achieved significant results in many applications [24].

NLPIS will also involve the matching relationship between two modalities of information, text and image. In the early stage of research on PIS, PIS used unimodal learning. Single modality learning in PIS refers to the process in which the model learns information from image perception modalities only [25]. Single-modality learning is relatively simpler and intuitive because the model only has to process data from one perceptual modality. Therefore, single modality learning is computationally efficient and has a relatively single task. With the improvement of PIS requirements, PIS is no longer a simple determination of whether there is a human or not. Therefore, the necessity of cross-modal pedestrian information retrieval has been emphasized in more studies [26].

Cross-modal refers to the process of information interaction or learning between different perceptual modalities. The cross-modal approach allows the system to acquire rich information from different perceptual modalities, which helps to improve the comprehension and representation of the input data by the system. Cross-modal learning helps to improve the generalization ability of pedestrian information retrieval models as it can learn more abstract and generic representations from multiple modalities, rather than just representations specific to one modality. Overall, unimodal learning may not be able to handle the complex relationships of multimodal information in real-world applications involving multimodal inputs, thus limiting its use in these scenarios. Single-modality learning is suitable for specific tasks and data contexts but has some limitations in processing multimodal information and improving generalization [27]. With the research and development of multimodal learning, more and more approaches are exploring how to effectively use information from different modalities to improve system performance.

Some relevant datasets and evaluation metrics are also investigated. Common quantitative evaluation metrics in the field of pedestrian information search: Precision, recall, F1 score, rank-k accuracy, and Mean Average Precision (MAP) score. Meanwhile, we selected the latest PIS research projects for replication [28]. We will use rank-k accuracy and mAP to evaluate the existing models.

II. Related Studies

In this section, we introduced the tasks associated with the PIS task.

A. Pedestrian Detection

Pedestrian Detection (PD) is the detection of traveling people in the input image. PD needs to locate the position of pedestrians. The application scenarios of PD are extremely wide, including but not limited to pedestrian retrieval in surveillance and automated driving. The main step in the study of PD as a target detection task is to select the region by

traversing. Feature extraction is performed during PD with manually designed feature extraction methods. Finally, PD can be achieved by classifying the extracted features. Cao et al. [29] proposed a detection framework that relies on handcrafted features and linear classifiers to achieve PD. Following the 2004 PD, the detectors were improved based on the research of Ribeiro et al. [30], and ICF, ACF, LDCF, and SCF were proposed [31].

However, the development of pedestrian detection methods has also taken a turn for the worse with the emergence of deep neural networks. The emergence of deep neural networks has brought a new direction to pedestrian detection techniques. Researchers have combined artificial features with stronger classifiers. Ribeiro *et al*. [30] used SCF as a detector combined with a deep neural network as a classifier for the detection task. Experiments have shown that access to deep neural networks has improved the accuracy of pedestrian retrieval substantially. Byeon and Kwak [31] used an ACF detector and trained an R-CNN-type neural network to generate pedestrian candidates. The study of Sheng *et al.* [32] implements the subdivision of pedestrian detection into pedestrian attributes and scene attributes [33]. Combine filtered channel features with CNN networks, following the traditional idea of manually designing a feature convolution kernel [34]. Propose an algorithm for learning complexity-aware cascades by seamlessly integrating manual and CNN features into a unified detector. Ma and Gao [35] use LDCF detectors and CNN models to construct part pools for local detection to deal with occlusion problems. Cai *et al.* [34] achieve the best trade-off between accuracy and speed. Meanwhile, [36] shows that deep neural networks can automatically learn high-level features of the target object without relying on manually designed feature extraction methods. Deep neural networks can extract robust features that are independent of the environment, increasing the robustness of detection.

In the early stage of pedestrian detection, researchers use R-CNN to generate candidate suggestions first, and then apply classification and regression algorithms to filter the candidate suggestions [37]. Based on R-CNN, Fast R-CNN, proposed by Zhang *et al*. [38] achieves further improvement in detection time and performance.

B. Pedestrian Re-Identification

Pedestrian Re-Identification (PRI) is also known as Pedestrian Re-Identification. PRI is a technique for determining the presence or absence of a target pedestrian in an image or video sequence [10]. The current research direction of the PRI technique can be roughly divided into feature extraction and metric learning.

In terms of feature extraction, most of the PRI research uses a combination of manual features and deep features. PRI researchers will first extract distinguishable features using manually designed feature extraction, and then learn higher-level features through deep learning neural networks. More PRIs innovate in structure to improve performance. For example, [39] designed two new convolutional layers to obtain the relationship between pairs of pedestrian images whose inputs have been aligned and cropped. Chen *et al.* [40] designed four convolutional layers and 2 fully connected layers to extract feature information from pedestrian images. In metric learning, PRI solves the problem of PRI by learning a distance metric [41]. Proposed KISSME, which uses likelihood ratios to determine similarity using statistical inference, proposed the null space to solve the problem of small sample sizes encountered in metric learning [42].

Traditional deep learning methods mainly use pairwise or ternary distance loss functions to supervise the training process [43], [39] input a pair of cropped pedestrian images into the network. Utilized ternary samples and managed to make the feature distances between pedestrian samples of the same identity as close as possible. Another approach is to consider the PRI problem as a multiclassification problem. Similarly, as the number of categories increases, using the Softmax loss function for PRI makes the process very slow or even fails to converge [44].

From the development history of PRI, we can find that after 2014, deep learning-based PRI task models have gradually gained the favor of most researchers. However, due to the size and singularity of the dataset, deep learning has not achieved as much success in PRI as other computer vision techniques. There is still much space for improving the performance and scene applicability of PRI tasks.

C. Pedestrian Information Search

Pedestrian Information Retrieval (PIS) is an end-to-end technique for detecting and recognizing pedestrians in panoramic images. PIS outputs information about the coordinates of the position of pedestrians in the image, as well as information about their identity. In terms of actual functionality, PIS can be considered as a joint task of PD and PRI. However, simply connecting the two tasks together cannot obtain good accuracy and efficiency.

Xu *et al.* [45] achieved PIS by modeling the common and unique characteristics of pedestrians through a sliding window search strategy. However, their research results show that simply connecting the two tasks cannot achieve good accuracy and efficiency. Proposed an end-to-end single CNN PIS framework [46]. Xiao *et al.* achieved simultaneous processing of two tasks in a single CNN. Meanwhile, proposed an Online Instance Matching (OIM) loss function to effectively improve the performance of neural networks. Investigated the overall impact of

the performance of the pedestrian detection component in the pedestrian search task. Used a two-stage strategy to implement PIS with a pedestrian detection network cascaded with a pedestrian re-identification network [46], [47]. Liu *et al.* [48] used a recurrent neural network to correct the pedestrian position in the panoramic image step by step and match the pedestrians. In traditional PIS studies, the dataset only contains manually cropped pedestrian frames. Contributed a new large-scale dataset, PRW, for pedestrian search. These works are aimed at making PIS applicable in integrated scenarios. These works aim to incorporate PD and PRI into a complete framework to reduce the mutual influence of the errors of the two otherwise independent networks [49].

The proposed PIS has made its application scenarios to become more leading with higher market requirements for PIS. PIS in severe non-aligned scenarios is a typical scenario. The focus of severe non-aligned scenarios is to utilize multiple features of the pedestrian images to achieve a reliable search of target pedestrians. Specifically, the facial features of the target will first be used to expand the target pedestrian samples to indirectly search for target pedestrians with large differences in body shape and appearance in the image. Then, the search results will be used to reverse search the sample with face information, and the similarity between the face features of the sample and the target face features will be used to filter the search results to obtain the final search results. The non-aligned scenario is closer to the actual application scenario, and the research is very significant. For example, in the study of Zheng, Gong, and Xiang, they proposed a Probabilistic Relative Distance Comparison (PRDC) model to reduce the distance between true matches and false matches [50]. Used the Deformable Part Model (DPM) to generate the Market-1501 dataset. Zheng *et al.* proposed a BoW descriptor method to try to bridge the gap between image searches [51]. Sun et al. proposed a Part-based Convolutional Baseline (PCB) to learn the features noted by parts. The PCB uses a simple uniform partitioning method to assemble some of the informative features into convolutional descriptors. However, their research requires images that have been used as a search source. It also makes PIS limited in many scenarios. Therefore, free natural language-based PIS is being seen in more and more studies [52].

Li *et al.* [16] proposed a recurrent neural network (GNA-RNN) with a gated neural attention mechanism using a recurrent neural network structure to solve the problem of affinity between textual descriptions and images of people. Krizhevsky *et al.* [53] used 1.3 million high-resolution images from the LSVRC-2010 ImageNet training set to train a large deep convolutional neural network to implement image-based PIS [54]. Modified the pre-trained BERT network by introducing Sentence-BERT (SBERT). BERT is a well-trained network for NLP [55], [56]. Reimers et al implemented text-based PIS by using concatenated and ternary network structures. However, the effect of unimodal PIS is limited. Then, a cross-modality-based feature extraction approach was verified to enhance the performance of PIS [57]. Multiple feature extraction requires multiple types of information to be concentrated in a single system [28]. Proposed an end-to-end learning framework, TIPCB. Using a multi-stage cross-modal matching approach, visual and textual representations are matched at multiple levels. Zhang *et al.* [58] proposed two losses, Cross-Modal Projection Matching (CMPM) and Cross-Modal Projection Classification (CMPC), which achieve image-text cross-modal feature extraction [57]. Proposed an end-to-end Simple and Robust Correlation Filtering (SRCF) framework to extract key information and adaptively align local features without the need for auxiliary tools. In summary, multimodal-based PIS performs better than unimodal PIS. A schematic of a typical cross-modal PIS model is shown in Fig. 1.

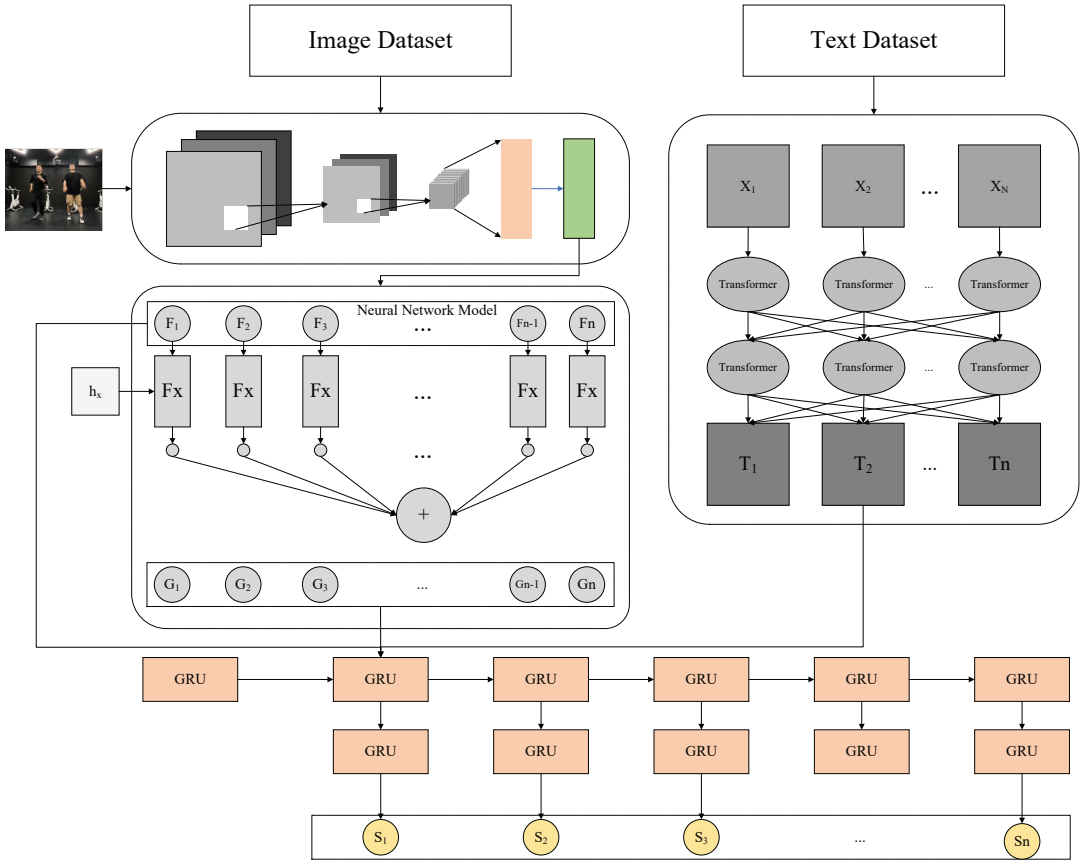

Fig. 1. A typical deep neural network model for cross-modal pedestrian information retrieval

Meanwhile, the research of Ye *et al.* [57] shows that a suitable sample set can effectively improve the performance of PIS. Existing PIS large-scale image datasets that are heavily used are the MS COCO dataset, the ImageNet dataset, and the CUHK-PEDES dataset. The MS COCO dataset was first created and released by Microsoft Research in September 2014. MS COCO, as a large-scale image dataset designed for the task of computer vision, has become one of the most important benchmarks in the field of computer vision. One of the important benchmarks in the field of computer vision. The data in the MS COCO dataset is rich and diverse. The richness and diversity of MS COCO make it an ideal data source for computer vision tasks. At the same time, the MS COCO dataset has gone through several updates to meet the evolving needs of computer vision research. The MS COCO dataset has been progressively added with more images, detailed annotations, and support for different tasks. The MS COCO dataset updating process not only increases the size of the dataset, but also improves the coverage of the dataset in terms of complex scenarios and a wide variety of objects. The MS COCO data offers researchers have more challenging and varied data sources for computer vision tasks. researchers with more challenging and practically relevant data support [59].

The ImageNet dataset is a large-scale image database created by Stanford University [60]. The ImageNet dataset is designed to facilitate research in the field of computer vision. The ImageNet dataset contains more than 14 million images. The images in the ImageNet dataset cover more than 20,000 different categories of objects. The categories of the ImageNet dataset cover a wide variety of objects, ranging from animals and plants to everyday objects. The image annotation of the ImageNet dataset is the addition of relevant labels and annotations to each image. The categories of the ImageNet dataset cover a wide range of objects, from animals and plants to everyday objects. The image annotation of the ImageNet dataset is the addition of relevant labels and comments to each image. The labels of the ImageNet dataset describe the main objects or scenes that appear in the image. The comprehensiveness and the wide range of applications of the ImageNet dataset have made it an important part of the image classification task.

The CUHK-PEDES dataset was created and released by the Research Institute of the Chinese University of Hong Kong [16]. The CUHK-PEDES dataset is a rich pedestrian dataset with annotations. The CUHK-PEDES dataset is the first dataset created specifically for PIS. The CUHK-PEDES dataset

aggregates images from five existing pedestrian re-identification datasets, including CUHK03, Market-1501, SSM, VIPER, and CUHK01, resulting in a large dataset of 40,206 images containing more than 13,003 individuals. Each image in the CUHK-PEDES dataset has been carefully annotated by staff [16]. CUHK-PEDES uses two textual descriptions to provide exhaustive details on the appearance, movements, and poses of the characters. The textual descriptions are rich in information, making the dataset more challenging and complex for practical applications [61]. Overall, all the above datasets play an important role in computer vision research. Different datasets provide rich and extensive data resources for model training and evaluation for image classification, target detection, and other related tasks.

The PIS research process is shown in Fig. 2. All the experiments were conducted based on this training-validation-testing division, which helps the researcher to compare and evaluate in a standard experimental setup. For performance evaluation, the researcher uses different metrics to evaluate the performance of the PIS task. The PIS task acts as a retrieval task. For a given target pedestrian image, after comparing it with the features of the samples in the candidate set, the similarity between the query pedestrian image and all the samples in the candidate set is calculated, and finally, all the candidate targets are sorted according to the similarity from highest to lowest. When the similarity between the query image and the matching image is higher, it means that the performance of the pedestrian search model is better. In the field of PIS, the commonly used quantitative evaluation metrics are Rank-K accuracy [62] and mean Average Precision (mAP) score [63]. In our study, we used Top1, Top5, and Top10 in the training model to evaluate the performance of the model. mAP is the average of the mean accuracy of all samples in the query set. This division provides researchers with a balanced and diverse dataset for model training and performance evaluation.

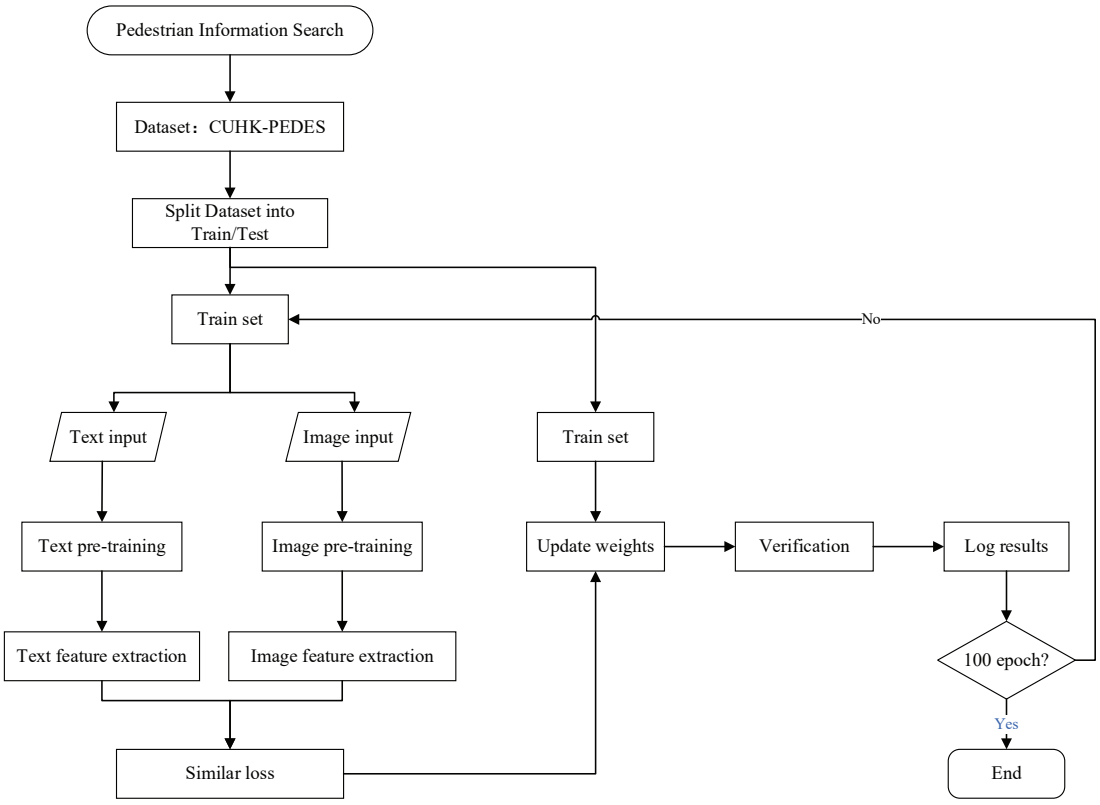

Fig. 2. Training process of a deep neural network model for cross-modal pedestrian information retrieval

III. Comparing Model Methods

In this section, we have selected different models and datasets for the training of PIS models. For the PIS task, the main choices are the selection of the PIS model and the selection of the training dataset. For model selection, we chose the cross-modal pedestrian information retrieval model, the Res50Bert model. Res50Bert model is a combination of ResNet50 and bert-base-uncased. Res50Bert model will train both natural language feature information and image feature information to improve the performance of PIS. Meanwhile, the performance of the Res50Bert model is better than the performance of other models in existing PIS research [28]. Therefore, we choose the Res50Bert model as the basic model for our validation. For dataset selection, we chose the

CUHK-PEDES dataset as the dataset designed for the PD task. The detailed description of pedestrians in the CUHK-PEDES dataset helps the model to better understand the semantic information in the images, which in turn improves its performance in real scenarios. The CUHK-PEDES dataset is divided into three non-overlapping subsets, which are used for training, validation, and testing, ensuring that individuals with the same identity do not appear in different sets. In addition, for performance evaluation, we adopted top-k accuracy as the primary metric for the person retrieval task. At the same time, we also calculated the mAP to analyze the overall performance of the model. Meanwhile, the study of Suo et al. [64] shows that the CUHK-PEDES dataset has better PIS performance. Therefore, we choose the CUHK-PEDES dataset as the training dataset.

In the process of studying the PIS task, we found that among the image neural network models, the performance of the EfficientNet neural network is improved in both efficiency and accuracy compared to the ResNet neural network model. Meanwhile, in the process of understanding text models, we also found several text models applicable to pedestrian information retrieval: paraphrase-multilingual-MiniLM-L12-v2 [65], distiluse-base-multilingual-cased-v2 [66], bert-base-nli-mean-tokens [67], all-mpnet-base-v2 [68], MiniLM-L12-H384-uncased [69]. Thus, we trained the base model and its different combinations as shown in Table I.

TABLE I
CROSS-MODAL PIS MODEL TRAINING RESULTS

Pedestrian Information Retrieval Model	Image Model	Text Model	Rank1	Rank5	Rank10	mAP
Res50Bert	ResNet50	best-base-uncased	0.595992	0.800582	0.867647	0.507627
Res50PMML12V2	ResNet50	paraphrase-multilingual-MiniL M-L 12-v2	0.577085	0.789754	0.860698	0.493230
Res50DBMCV2	ResNet50	distiluse-base-multilingual-cased-v2	0.584034	0.792178	0.863445	0.497706
Res50BBNMT	ResNet50	bert-base-nli-mean-tokens	0.573368	0.791370	0.864092	0.490792
Res50AMBV2	ResNet50	all-mpnet-base-v2	0.588235	0.789916	0.868778	0.500985
Res50MLH384U	ResNet50	MiniL M-L 12-H384-uncased	0.578539	0.786316	0.837750	0.497329
EB1Bert	EfficientNet B1	best-base-uncased	0.605992	0.805582	0.869647	0.509627
EB1PMML12V2	EfficientNet B1	paraphrase-multilingual-MiniL M-L 12-v2	0.314156	0.554299	0.665482	0.264532
EB1DBMCV2	EfficientNet B1	distiluse-base-multilingual-cased-v2	0.526503	0.754848	0.840175	0.450074
EB1BBNMT	EfficientNet B1	bert-base-nli-mean-tokens	0.499199	0.742400	0.826811	0.425028
EBI AMBV2	EfficientNet B1	all-mpnet-beste-v2	0.577085	0.789754	0.860698	0.493230
EBIMLH384U	EfficientNet B1	MiniL M-L 12-H384-uncased	0.080478	0.214447	0.316742	0.076245

From Table I, we found that the basic model reaches about 60%. The experimental results indicate that both different text models and image models affect the performance of the PIS model. At the same time, the experimental results also illustrate that there is a possibility that the combination of different text models and image models can enhance the performance of PIS.

IV. CONCLUSION AND FUTURE WORK

Our study focuses on a systematic review of pedestrian information retrieval. Firstly, we introduce the pedestrian search task and the pedestrian re-identification task. After the introduction of pedestrian search and pedestrian re-identification focus is on the pedestrian information retrieval task, which is a combination of the two tasks. In this paper, we focus on the feature extraction method, modality, and dataset of PIS. We summarize the performance of unimodal and multimodal pedestrian information retrieval tasks. Existing research shows that multimodality is fully capable of implementing free-form natural language-based PIS. The implementation of natural language-based pedestrian information retrieval also heralds the possibility of a higher degree of freedom in human-computer interaction. Meanwhile, we validate the current best-performing model using experimental replication. The experimental results show that although the PIS performance has been improved, the accuracy still cannot reach a high level. Therefore, natural language-based PIS still deserves more time and effort.

At the same time, the existing studies are incomplete. First of all, the existing studies have been conducted in English. There is a lack of research on other languages. Other languages, such as Chinese, which has a large number of speakers in the world, and Thai and Japanese, which are widely spoken, are also small languages that are worth studying. Correspondingly, the existing datasets are all labeled and annotated in English. There are differences in the way languages are expressed and used in practice. It is also meaningful to establish exclusive pedestrian information retrieval datasets for different languages.

Secondly, existing research is realized by strong research teams and companies. The implementation of low-performance hardware to train high-performance models can be an important research direction in the future. We can consider how we can utilize the extreme performance of computers. This will make AI reachable.

REFERENCES

[1] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3820-3834, Jan. 2020.

[2] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Proc. IEEE CVPR*, 2018, pp. 6781-6789.

[3] C. J. Pai, H. R. Tyan, Y. M. Liang, and H. Y. Mark Liao, "Pedestrian detection and tracking at crossroads," *Pattern Recognit.*, vol. 37, no. 5, pp. 1025-1034, Jan. 2003.

[4] M. You, Y. Zhang, C. Shen, and X. Zhang, "An extended filtered channel framework for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1640-1651, May 2018.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743-761, Apr. 2012.

[6] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179-2195, Oct. 2009.

[7] T. Liu and T. Stathaki, "Faster R-CNN for robust pedestrian detection using semantic segmentation network," *Front. Neurorobot.*, vol. 64, no. 12, pp. 1-10, Oct. 2018.

[8] S. Zhai, S. Dong, D. Shang, and S. Wang, "An improved Faster R-CNN pedestrian detection algorithm based on feature fusion and context analysis," *IEEE Access*, vol. 8, pp. 138117-138128, Jul. 2020.

[9] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780-785, Jul. 1997.

[10] L. Zheng et al. "Mars: A video benchmark for large-scale person re-identification," *Computer Vision-ECCV*, vol. 9910, pp. 868-884, Sep. 2016.

[11] D. Figueira, M. Taiana, A. Nambiar, J. Nascimento, and A. Bernardino, "The HDA+ dataset for research on fully automated re-identification systems," Cham, CH: Springer, 2015, pp. 241-255.

[12] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. ECCV*, 2018, pp. 486-504.

[13] S. Salehian, P. Sebastian, and A. B. Sayuti, "Framework for pedestrian detection, tracking and re-identification in video surveillance system," in *Proc. IEEE ICSIPA*, 2019, pp. 192-197.

[14] S. Zhang, D. Chen, J. Yang, and B. Schiele, "Guided attention in CNNs for occluded pedestrian detection and re-identification," *Int. J. Comput. Vis.*, vol. 129, pp. 1875-1892, Apr. 2021.

[15] 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)," *ECCV*, vol. 11208, pp. 480-496, Oct. 2018.

[16] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE CVPR*, 2017, pp. 5187-5196.

[17] X. Han, S. He, L. Zhang, Q. Ye, and J. Sun, "Text-based person search with limited data," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 1-20.

[18] D. K. Panda and S. Meher, "Dynamic background subtraction using local binary pattern and histogram of oriented gradients," in *Proc. ICIIP*, 2015, pp. 306-311.

[19] C. Wang, Z. Luo, Y. Lin, and S. Li, "Text-based person search via multi-granularity embedding learning," in *Proc. IJCAI*, 2021, pp. 1068-1074.

[20] H. Liu, J. Feng, M. Qi, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492-3506, Jul. 2017.

[21] P. Srivastava and A. Khare, "Utilizing multiscale local binary pattern for content-based image retrieval," *Multimedia Tools Appl.*, vol. 77, pp. 12377-12403, Jun. 2017.

[22] K. Chen, X. Song, X. Zhai, B. Zhang, B. Hou, and Y. Wang, "An integrated deep learning framework for occluded pedestrian tracking," *IEEE Access*, vol. 7, pp. 26060-26072, Feb. 2019.

[23] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proc. IEEE/CVF*, 2023, pp. 2787-2797.

[24] J. S. J. Rani and M. G. Augasta, "PoolNet deep feature based person re-identification," *Multimedia Tools Appl.*, vol. 82, no. 16, pp. 24967-24989, Jan. 2023.

[25] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *J. LaTeX Class. File*, vol. 14, no. 8, pp. 1-20, Aug. 2016.

[26] Z. Wang, Z. Wang, Y. Zheng, and Y. Y. Chuang, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proc. IEEE/CVF CVPR*, 2020, pp. 618-626.

[27] Y. Sun, L. Zheng, W. Deng, and S. Wang., "SVDNet for Pedestrian Retrieval," in *Proc. 2017 IEEE ICCV*, 2017, pp. 3820-3828.

[28] Y. Chen, G. Zhang, Y. Lu, Z. Wang, Y. Zheng, and R. Wang "TIPCB: A simple but effective part-based convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171-181, 2002.

[29] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 44, no. 9, pp. 4913-4934, Apr. 2021.

[30] D. Ribeiro, J. C. Nascimento, A. Bernardino, and G. Carneiro, "Improving the performance of pedestrian detectors using convolutional learning," *Pattern Recognit*., vol. 61, pp. 641-649, 2017.

[31] Y. H. Byeon and K. C. Kwak, "A performance comparison of pedestrian detection using faster RCNN and ACF," in *Proc. 2017 6th IIAI Int. Congr. Adv. Appl. Inform.*, 2017, pp. 858-863. https://doi.org./10.1109/IIAI-AAI.2017.196

[32] B. Sheng, Q Hu, J. Li, W. Yang, B. Zhang, and C. Sun, "Filtered shallow-deep feature channels for pedestrian detection," *Neurocomputing,* vol. 9, pp. 106-113, Aug. 2017.

[33] S. Zhang, R. Benenson, and B. Schiele, "Filtered feature channels for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1751-1760.

[34] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. 2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3361-3369.

[35] Z. Ma and P. P. Gao, "Research on the Cascade Pedestrian Detection Model Based on LDCF and CNN," in *Proc. 2018 IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, 2018, pp. 314-320.

[36] Y. Luo, C. Zhang, M. Zhao, H. Zhou, and J. Sun "Where, what, whether: Multi-modal learning meets pedestrian detection," *Computer Vision Foundation,* no. zrxiv2103.11599, pp. 14065, Dec. 2020.

[37] P. Dong and W. Wang, "Better region proposals for pedestrian detection with R-CNN," in *Proc. 2016 Visual Commun. Image Process. (VCIP)*, 2017, pp. 1-4.

[38] H. Zhang et al., "Pedestrian Detection Method Based on Faster R-CNN," in *Proc. 2017 13th Int. Conf. Comput. Intell. Security (CIS)*, 2018, pp. 427-430.

[39] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3908-3916.
[40] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *Proc. 2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2015, pp. 715-718.
[41] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2288-2295.
[42] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1239-1248.
[43] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037-3045, Oct. 2019.
[44] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 815-823.
[45] Y. Xu, X. Cao, and H. Qiao, "An efficient tree classifier ensemble-based approach for pedestrian detection," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 107-117, May 2011.
[46] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3415-3424.
[47] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492-3506, May 2017.
[48] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3346-3355.
[49] W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR 2011, 2011*, pp. 649-656.
[50] L. Zheng, L. Shen, L. Tian, J. Wang, and Q. Tian "Scalable person re-identification: A benchmark," in *Proc. 2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1116-1124.
[51] 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)," in *Proc. Eur. Conf. Comput. Vis.*, 2017, pp. 480-496.
[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84-90, Dec. 2012.
[53] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982-3992
[54] Q. Li and J. Qu, "A novel BNB-NO-BK method for detecting fraudulent crowdfunding projects," *SJST*, vol. 44, no. 5, pp. 1209-1219, Oct. 2022.
[55] W. Hou and J. Qu, "BM5-SP-SC: A dual model architecture for contradiction detection on crowdfunding projects," *CAST*, vol. 23, no. 6, pp. 1-29, Apr. 2023.
[56] W. Suo et al., "A simple and robust correlation filtering method for text-based person search," *Eur. Conf. Comput. Vis.*, vol. 13695, pp. 726-742, Nov. 2022.
[57] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," *Springer*, vol. 11205, pp. 686-701, Oct. 2018.
[58] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872-2893, Jan. 2021.
[59] T. Y. Lin et al., "Microsoft coco: Common objects in context," *Springer*, vol. 8693, pp. 740-755, Sep. 2014.
[60] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248-255.
[61] X. H. Phan, L. M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 91-100.
[62] Y. Zhang, R. Alturki, H. J. Alyamani, and M. Ikram, "Multilabel CNN-based hybrid learning metric for pedestrian reidentification," *Mobile Inf. Syst.*, vol. 2021, no. 7, pp. 1-7, Apr. 2021.
[63] J. Revaud, J. Almazán, R. S. Rezende, and C. R. de Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proc. 2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 5106-5115.
[64] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, and R. Valencia-García, "UMUTeam at SemEval-2023 Task 3: Multilingual transformer-based model for detecting the genre, the framing, and the persuasion techniques in online news," in *Proc. 17th Int. Workshop Semantic Eval.*, 2023, pp. 609-615.
[65] B. Bharathi and G. U. Samyuktha, "Machine learning based approach for sentiment analysis on multilingual code-mixing text," in *Proc. FIRE*, 2021. pp. T6-T18.
[66] K. Peyton and S. Unnikrishnan, "A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs," *Results Eng.*, vol. 17, p. 100856, Mar. 2023.
[67] G. Ashqar and A. Mutlu, "A Comparative assessment of various embeddings for keyword extraction," in Proc. *5th Int. Congr. Human-Computer Interaction,* 2023, pp. 1-6.
[68] R. Qin, "Bert-based feature extraction approach for software forum posts," *IEEE Access*, vol. 11, pp. 1-9, Jan. 2024, https://doi.org./10.1109/ACCESS.2024.3426976

Yan Xie is currently studying for the Master of Engineering Technology, Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand. She received B.B.A from Nanjing Tech University Pujiang Institute, China, in 2022. Her research interests are Research direction is artificial intelligence, image processing, and Natural Language Processing (NLP).

Jian Qu is an Assistant professor at the Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand He received a Ph.D. with an Outstanding Performance award from Japan Advanced Institute of Science and Technology, Japan, in 2013. He received B.B.A with Summa Cum Laude honors from Institute of International Studies of Ramkhamhaeng University, Thailand, in 2006, and M.S.I.T from Sirindhorn International Institute of Technology, Thammast University, Thailand, in 2010. He has been a house committee for Thai SuperAI since 2020. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval, and image processing.