ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

A Study on Bilingual Deep Learning PIS Neural Network Model Based on Graph-Text Modal Fusion

Yan Xie[1] and Jian Qu[2]

ABSTRACT

This research investigates a multilingual cross-modal pedestrian information search (PIS) technique based on graph-text modal fusion. Initially, we used a combination of replacement neural networks to improve the English Language-Based Pedestrian Information Search model with Graph-Text Modal Fusion (GTMFLPIS) performance. In addition, existing research lacks GTMFLPIS models for other languages. Therefore, we propose to train GTMFLPIS models for Chinese. The Chinese GTMFLPIS model was trained using our previously constructed Chinese CUHK-PEDES dataset. The Rank1 of the Chinese RN50_PMML12V2 model reached 0.5989. In addition, we found that a single model could not adapt to the limitations of multiple languages. Therefore, we propose a novel architecture to implement a single-model multilingual cross-modal GTMFLPIS model in this research. We propose RN50_DBMCV2 and ENB7_DBM-CV2, both of which have improved performance over the existing ones. We constructed a bilingual dataset using our Chinese CUHK-PEDES dataset and existing English CUHK-PEDES dataset to test our novel multilingual cross-modal GTMFLPIS model. In addition, we found that the loss function significantly impacts the model during our experiments. Therefore, we optimized the performance of the existing loss functions for cross-modal GTMFLPIS models. Our proposed CCMPM loss function improves the performance of the model by 2%. The experimental results of this research show that our proposed model has advantages in improving the accuracy of PIS.

Article information:

Keywords: GTMFLPIS (Language-Based Pedestrian Information Search model with Graph-Text Modal Fusion), Convolutional Neural Network, Pedestrian Information Search, Cross-Modal, Multi-Modal Fusion

Article history:

Received: April 18, 2024
Revised: August 8, 2024
Accepted: October 10, 2024
Published: November 16, 2024 (Online)

DOI: 10.37936/ecti-cit.2025191.256455

1. INTRODUCTION

The technique of quickly and accurately obtaining specific information about pedestrians from various sources is called Pedestrian Information Search (PIS).

PIS tasks can be mainly classified into language-based pedestrian information search (LPIS) and image-based pedestrian information search (IPIS). IPIS requires at least one target pedestrian image as a query basis. However, target pedestrian images are often difficult to obtain in real-world situations. In contrast to IPIS, LPIS does not have these limitations, and LPIS can be adapted to a broader range of scenarios. Therefore, our research proposes that LPIS is more meaningful.

Existing LPIS research has achieved good results. Yang *et al.* [1] proposed the APTM framework to implement text-based modal LPIS. Zhao *et al.* [2] proposed a deep, partially aware representation learning method for person retrieval implementing image-based modal LPIS. Kakouros *et al.* [3] proposed a novel attention-focused approach to speech-based modal LPIS based on the combination of correlation between representation coefficients and label smoothing. Zang *et al.* [4] proposed a multi-directional, multi-scale pyramid converter (PiT) to implement video-based LPIS. The above studies are deep neural network models for LPIS trained on individual modal information. However, Kaur *et al.* [5] pointed out that the model trained based on unimodal information cannot fully reflect the features and information of pedestrians.

Meanwhile, the accuracy of the unimodal deep neural network model is easily affected by the training samples, which leads to unstable results. Yuan

[1,2] The authors are with Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand 11120, Email: 6372100142@stu.pim.ac.th and jianqu@pim.ac.th
[2] Corresponding author: jianqu@pim.ac.th

et al. [6] showed that the unimodal deep neural network model needs to be stronger regarding robustness and model generalization ability. Therefore, the study of multimodal language-based pedestrian information search (MLPIS) is necessary. Alsaleh *et al.* [7] implemented MLPIS by mapping the relationship between spatial features and pedestrian cognitive abilities. Jiang *et al.* [8] achieved MLPIS by matching global images and texts through cross-modal implicit relational reasoning (IRR). Zhao *et al.* [9] pointed out that not all modalities are worth fusing, and not more information is better. There is a semantic gap between the data of different modalities, which makes it challenging to map and relate them directly. The focus of modal information fusion is dealing with the mapping relationship between different modal information. Simultaneously, the various characteristics and representations of different modal data will lead to more complex and challenging annotation of cross-modal data. Therefore, the selection of modal information objects is more important. Zhou *et al.* [10] showed that the modal fusion of text and images is a good choice. Images can provide intuitive visual information and be combined with text to present information more intuitively. Images are good at presenting visual features, and text is good at describing logic and abstract concepts, which can provide more comprehensive information after fusion.

Graphic and text modal fusion helps interpret and understand complex information better and improves the accuracy of the information. Therefore, our research proposes to train a multimodal Language-Based pedestrian information search model with Graph-Text Modal Fusion (GTMFL-PIS). The fusion of image-text modalities exhibits several advantages in PIS. Firstly, we integrate information from image and text modalities, providing more prosperous and intuitive results for PIS. Images can display detailed visual characteristics of pedestrians, while texts can provide descriptive feature information about pedestrians. These two modalities complement each other, searching results more comprehensive. Secondly, the fusion of image-text modalities makes personal information search more aligned with the cognitive habits of a user. Users can quickly retrieve pedestrians through intuitive images and detailed textual information. The fusion of image-text modalities enhances user experience and improves user satisfaction.

GTMFLPIS model mainly involves the selection of a deep image neural network model, a deep text neural network model, and loss function selection. In selecting image neural network models, existing neural networks perform well in single image feature extraction tasks. For example, Theckedath *et al.* [11] trained a ResNet50 model for image feature extraction. Majib *et al.* [12] trained a VGGNet model for image feature extraction. Atila *et al.* [13] trained an EfficientNet model for image feature extraction. In selecting text neural network models, existing neural networks perform well in single-text feature extraction tasks. Passaro *et al.* [14] trained a BERT-BASE-NLI-MEAN-TOKENS (BBNMT) model to achieve text feature extraction. Frick *et al.* [15] trained an ALL-BASE-V2 (AMBV2) model for text feature extraction, and Dang *et al.* [16] trained a BERT-BERT-BASE-UNCASED (BBU) model for text feature extraction. The above text models are all English text models. Moreover, there are some existing Chinese text models: Kapočiūtė *et al.* [17] trained a DISTILUSE-BASE-MULTILINGUAL-CASED-V2 (D-BMCV2) model for text feature extraction. Farray *et al.* [18] trained a PARAPHRASE-MULTILINGUAL-V2 (DBMCV2) model for text feature extraction. PARAPHR-ASE-MULTILINGUAL-MINILM-L12-V2 (PMMLV2) model for text feature extraction. Mishra *et al.* [19] trained a BERT-BASE-MULTILINGUAL-UNCASED (BBMU) model for text feature extraction. In terms of loss function selection, Chen *et al.* [20] proposed two loss functions: cross-modal projection matching (CMPM) and cross-modal projection classification (CMPC). Zhang *et al.* [21] proposed cross-modal cross entropy (CMCE) loss to extract cross-modal features. The diversity of choices of image neural network models, text neural network models, and loss functions illustrates the diversity of GTMFLPIS model combinations. Therefore, our research proposes to validate the feasibility of GTMFLPIS models under different combinations of text models, image models, and loss functions. In addition, our research proposes to validate the performance of trainable GTMFLPIS models in PIS tasks.

At the same time, by reading existing journals and literature, we found that existing LPIS and MLPIS research can only achieve pedestrian information search in a single language (mainly English). There is almost a blank, especially in research on pedestrian information searches using the Chinese language. Therefore, we propose to train a Chinese graphic fusion model for pedestrian information search. Achieving pedestrian information search in Chinese means that we need Chinese language datasets that can be trained. However, no large-scale Chinese language pedestrian search dataset exists in existing research. Therefore, our research proposes to use Chinese to re-label the existing large-scale English pedestrian information search datasets. Shen *et al.* [22] summarized that there are many existing large-scale English pedestrian information search datasets, among which the most commonly used ones are the CUHK-PEDES dataset, the RSTPReid dataset, and the CUHK-SYSU dataset. Among these datasets, the CUHK-PEDES dataset is the first dataset dedicated to linguistic pedestrian search. The CUHK-PEDES dataset contains many images and text descriptions,

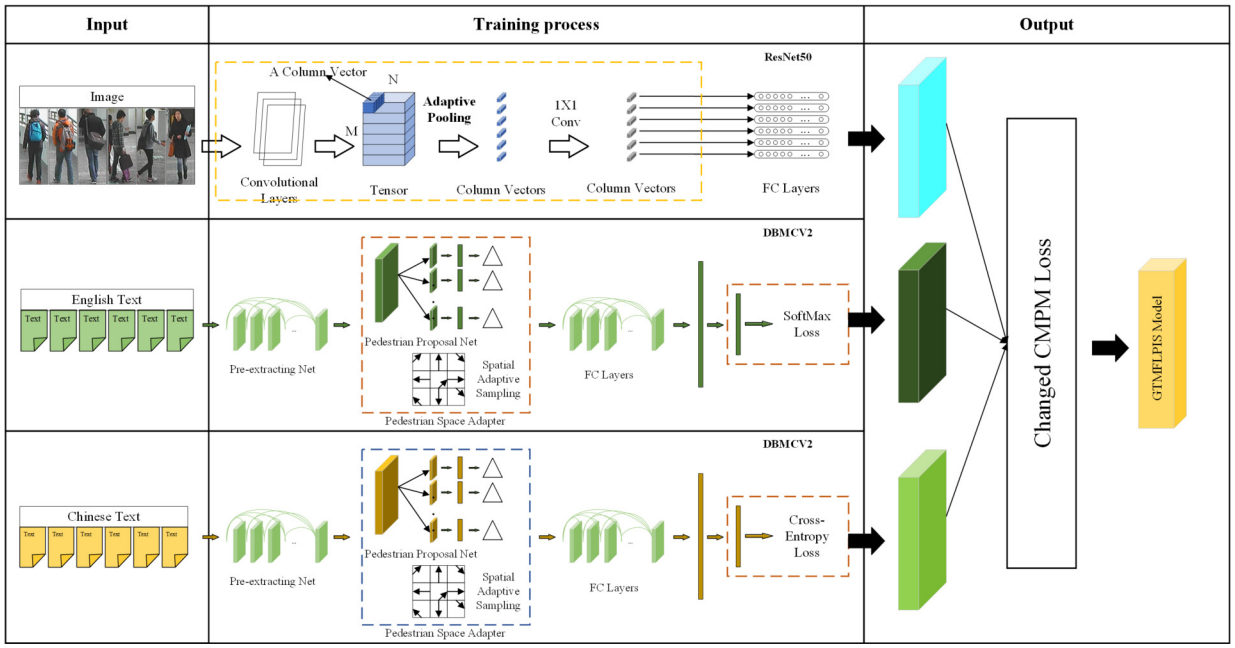

Fig.1: *Bilingual GTMFLPIS Model Framework.*

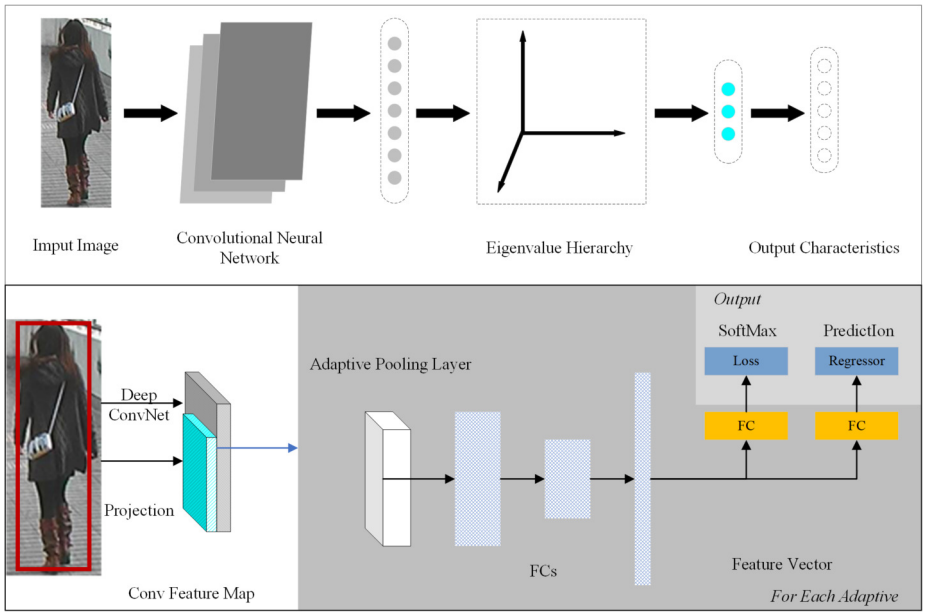

Fig.2: *The process of image feature extraction.*

divided in great detail. Therefore, our research proposes to annotate the CUHK-PEDES dataset using Chinese. The fusion of English and Chinese bilingualism exhibits several advantages in PIS. Firstly, a PIS system integrating English and Chinese textual information can cover various databases and information sources. Expanding the search scope increases the likelihood of finding relevant personal information. Secondly, a multilingual model implies support for users speaking multiple languages. With globalization, increasing numbers of users need to communicate in various languages. The fusion of English and Chinese bilingualism enhances the internationalization and inclusiveness of the PIS system.

In addition, after we completed the Chinese GTMFLPIS training and the application, we perceived a less convenient problem with this approach. The Chinese GTMFLPIS model can only be searched in Chinese, and the English GTMFLPIS model can only be searched in English. Therefore, our study proposes to train a GTMFLPIS model to achieve bilingual (Chinese and English) pedestrian information searches. The integration of image-text modality and English-Chinese bilingualism can further amplify their respective advantages in PIS, bringing comprehensive improvements to the system. Firstly, the fusion of ima-

Table 1: *Overview of the CCMPM Loss Function.*

Algorithm 1 Changed Cross-modal projection matching(CCMPM)	
Input: image_embeddings(*IE*), text_embeddings_EN(*TEE*), text_embeddings_CHS(*TEC*), labels(*Ls*)	
Output: ccmpm_loss(*Loss*)	
1:	Initialization function, receiving parameters(Epsilon(*E*); Weight parameter(*W*);)
2:	Using *Xavier* to homogenize data for initialization of weights
3:	*Batch_size(BS)* = *IM*.shape[0]; //Get Batch Size
4:	Labels_reshape(*IR*) = torch.reshape(*Ls*, (*BS*, 1)); //Reinventing the Label
5:	Labels_dist(*LD*) = *IR* - *IR*.t(); //Calculate label distance
6:	Labels_mask(*LM*) = (*LD* == 0); //Tagging Label Mask
7:	Image_norm(*IN*) = *IE* / *IE*.norm(dim=1, keepdim=True); //Normalize the image embedding
8:	Normalization of English text embedding -> *TNE*
9:	Image to Engilsh text projection -> *IPTE*
10:	Engilsh text toImage projection -> *TPIE*
11:	Normalization of Chinese text embedding -> *TNC*
12:	Image to Chinese text projection -> *IPTC*
13:	Chinese text to Image projection -> *TPIC*
14:	Normalized true matching distribution -> *LMN*
15:	Calculate image-to-text predictions(Engilsh)
16:	Calculate image-to-text predictions(Chinese)
17:	Adding a regularisation term -> λ*Reg*
18:	ccmpm_loss += λ*Reg
19:	Forward Propagation Calculated Loss
20:	Dynamic adjustment of λ based on model performance on validation sets
21:	**return** ccmpm_loss

ge-text modality and English-Chinese bilingual textual information enables the PIS system to achieve multi-dimensional information integration and retrieval. Secondly, PIS may involve multiple languages and visual features in practical applications. A PIS system incorporating image-text modality and English-Chinese bilingual textual information can better adapt to this complexity. Therefore, our research proposes to train a GTMFLPIS model to achieve bilingual (Chinese and English) pedestrian information search.

In summary, this research uses the Chinese language to re-label the existing English CUHK-PEDES dataset. Our research trained the Chinese GTMFLPIS model using our Chinese CUHK-PEDES dataset. In addition, we proposed a novel bilingual GTMFLPIS model architecture. Our bilingual GTMFLPIS model achieved the single-model multilingual pedestrian information search task well.

2. MATERIALS AND METHODS

In this chapter, we describe the design of the GTMFLPIS model framework, the dataset, the experimental setup, and the metrics for evaluating the experimental results of this research.

2.1 The design of the GTMFLPIS Model

2.1.1 Bilingual GTMFLPIS Model Framework

The input, output, and training process for training image and language models are shown in Figure 1. Details of English and Chinese language models are shown in Figure 1. The overall view of Figure 1 is the novel Bilingual GTMFLPIS Model proposed in this research. The training of the Bilingual GTMFLPIS Model consists of three inputs: images, English language labelled text dataset, and Chinese language labelled text dataset. Different pre-training models were used for feature extraction for different datasets. After the features are extracted from the image neural network model and the text neural network model, the features are balanced by a modified loss function to balance the mapping relationship between the individual modal information.

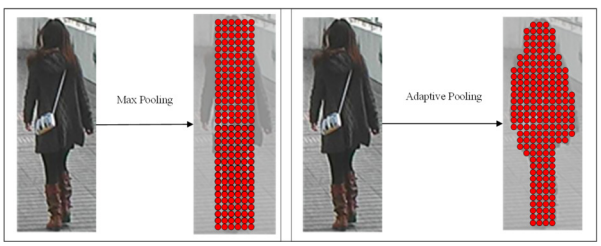

Fig.3: *Different Pooling.*

Table 2: *Overview of the datasets.*

	Language	Part	Number of images	Number of texts	Number of pedestrians
CUHK-PEDES Dataset	English	Training set	34054	68108	11003
		Validation set	3078	6158	1000
		Test set	3074	6156	1000
		Total	40206	80412	13003
	Chinese	Training set	34054	102162	11003
		Validation set	3078	9234	1000
		Test set	3074	9222	1000
		Total	40206	120618	13003

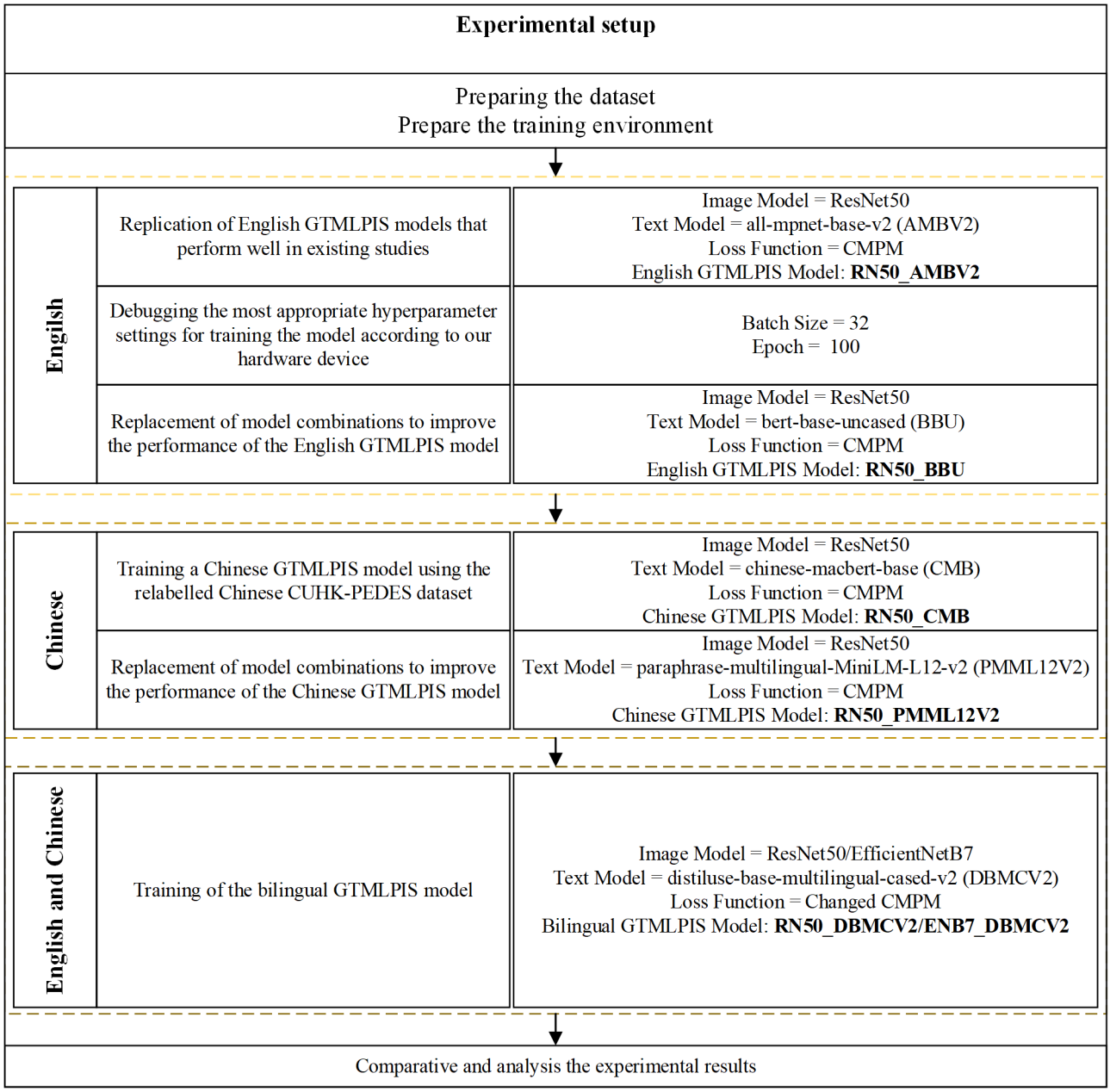

Fig.4: *The experimental configuration for training the GTMFLPIS model.*

Table 3: *Model and loss function candidates of the combination for model training.*

Model Item		Item Name
Image neural network model		ResNet50
		ResNet18
		ResNet34
		VGG16Net
		EfficientNetV2
		EfficientNetB7
Text neural network model	English and Chinese	paraphrase-multilingual-MiniLM-L12-v2 (PMML12V2)
		distiluse-base-multilingual-cased-v2 (DBMCV2)
	English	Bert-base-uncased (BBU)
		Bert-base-nli-mean-tokens (BBNMT)
		all-mpnet-base-v2 (AMBV2)
		MiniLM-L12-H384-uncased (MLHU)
	Chinese	Chinese-macBert-base (CMB)
Loss Function		Cross-Modal Projection Matching (CMPM)
		Cross-Modal Projection Classification loss (CMPC)
		Cross-Modal Cross-Entropy loss function (CMCE)
		Changed Cross-Modal Projection Matching (CCMPM)

Table 4: *Software and hardware parameters.*

Hardware	CPU	Intel(R)Core(TM)i7-10710U
	Random Access Memory (RAM)	32 Gigabytes
	GPU	NVIDIA GeForce RTX 3080 Ti
	Display memory	12 Gigabytes GDDR6X Memory
	Hard Drive Size	2 Terabytes
Software	Deep Learning Frameworks	PyTorch 1.8.1
	Programming Language	Python 3.8
	Torch	1.8.1
	TorchVison	0.12.0
	CUDA	12
	Transformers	4.30.2

2.1.2 Feature Extraction of Bilingual GTMFLPIS Model

Figure 2 shows the feature extraction process of the image model. Before proceeding with the feature extraction process of the image model, we pre-process the image. Our pre-processing takes the image for enhancement and de-noising. Image pre-processing enhances the feature extraction. Then, we feed the processed image into a convolutional neural network. A neural network is equivalent to a specific algorithm to extract features from an image. Therefore, we focused on the layers in the neural network which affected the feature extraction performance. We experimentally found that the impact of a suitable pooling strategy on feature extraction performance is more prominent. Therefore, we experimented with the impact of different pooling layers. Figure 3 shows that the performance of the adaptive pooling layer is better than that of the maximum pooling layer.

2.1.3 Loss Functions

Loss functions provide a quantitative way to measure model performance on training data. A loss function that accurately quantifies the performance of a model can reduce the time spent on model validation and determine the direction in which the model should be optimized. As shown in Table 1, this research modified the CMPM loss function. The Changed CMPM (CCMPM) loss function inputs are image embeddings, English text embeddings, Chinese text embeddings, and labels. The input part of the model increases the possibility of the model being overfitted. Therefore, we added the regularization term in the loss function. Meanwhile, we found that the weight of the regularization term significantly impacts the performance of the deep learning model. Therefore, to obtain the optimal weights, we propose to dynamically adjust the regularization weights according to the performance of the model on the validation set. As shown in Table 1, λ is the adaptive regularization weight.

2.2 Datasets and experimental environments

able 2 shows the parameters of the English CUHK-PEDES dataset and our Chinese CUHK-PEDES dataset re-labeled in Chinese. For the bilingual study, we invited twenty students from the College of Arts and Letters of the Nanjing University of Technology to help with the data alignment and labelling. At the same time, we chose to annotate a picture with three different Chinese language expressions. The specific parameters of the dataset are given in Table 2. Table 4 shows the software and hardware configurations we used to train the GTMFLPIS model.

2.3 Model Training Setups

Figure 3 shows the whole experimental process and experimental setup of our research. This re-

Table 5: *English GTMFLPIS model training results.*

Image Model	Text Model	Loss Function	Rank1	Rank5	Rank10	mAP	Image Model	Text Model	Loss Function	Rank1	Rank5	Rank10	mAP
ResNet50	AMBV2	CMPM	0.5231	0.7137	0.752	0.4755	VGG16Net	AMBV2	CMPM	0.5152	0.7378	0.7535	0.467
		CMPC	0.4885	0.6954	0.747	0.4469			CMPC	0.4779	0.6907	0.7343	0.4431
		CMCE	0.4834	0.7024	0.7459	0.455			CMCE	0.5108	0.6868	0.7308	0.452
	PMML12V2	CMPM	0.5249	0.7119	0.7604	0.4622		PMML12V2	CMPM	0.5203	0.7117	0.7571	0.4696
		CMPC	0.4767	0.6992	0.7066	0.4462			CMPC	0.4821	0.7015	0.7433	0.4514
		CMCE	0.4666	0.6988	0.7289	0.4552			CMCE	0.4681	0.6882	0.734	0.445
	DBMCV2	CMPM	0.5331	0.7426	0.7515	0.4632		DBMCV2	CMPM	0.5386	0.7315	0.7625	0.478
		CMPC	0.4887	0.7014	0.7143	0.4416			CMPC	0.5039	0.6828	0.738	0.452
		CMCE	0.4697	0.7049	0.7103	0.4538			CMCE	0.4675	0.6994	0.714	0.442
	BBU	**CMPM**	**0.5942**	**0.7992**	**0.8017**	**0.5099**		BBU	CMPM	0.5304	0.7196	0.765	0.4793
		CMPC	0.4874	0.6867	0.7109	0.4599			CMPC	0.4864	0.6905	0.7461	0.448
		CMCE	0.4624	0.6895	0.7273	0.4457			CMCE	0.4955	0.6826	0.724	0.4525
	BBNMT	CMPM	0.5267	0.7237	0.7645	0.4696		BBNMT	CMPM	0.5213	0.7344	0.7693	0.4621
		CMPC	0.4989	0.6878	0.7431	0.4531			CMPC	0.4767	0.6856	0.7179	0.4516
		CMCE	0.5064	0.6839	0.7242	0.4435			CMCE	0.5005	0.6881	0.7089	0.45
	MLHU	CMPM	0.5183	0.7417	0.7653	0.4791		MLHU	CMPM	0.5197	0.7206	0.7679	0.4624
		CMPC	0.4739	0.684	0.7398	0.4471			CMPC	0.492	0.6814	0.711	0.46
		CMCE	0.5078	0.6988	0.7081	0.4588			CMCE	0.4847	0.7013	0.7403	0.4462
ResNet18	AMBV2	CMPM	0.5366	0.7478	0.7607	0.4631	EfficientNetV2	AMBV2	CMPM	0.5347	0.7175	0.7599	0.475
		CMPC	0.493	0.6871	0.7154	0.4445			CMPC	0.4716	0.7013	0.7082	0.4459
		CMCE	0.4803	0.7034	0.7302	0.4504			CMCE	0.4732	0.6866	0.7147	0.4406
	PMML12V2	CMPM	0.5196	0.7304	0.7628	0.4654		PMML12V2	CMPM	0.52	0.7081	0.7692	0.4777
		CMPC	0.4946	0.6888	0.7199	0.4491			CMPC	0.4865	0.6968	0.7208	0.4525
		CMCE	0.4907	0.6963	0.7106	0.4413			CMCE	0.4899	0.6971	0.7289	0.4407
	DBMCV2	CMPM	0.5355	0.7143	0.7517	0.4684		DBMCV2	CMPM	0.53	0.7273	0.7698	0.4701
		CMPC	0.4953	0.6955	0.7317	0.4582			CMPC	0.5032	0.7008	0.7341	0.4552
		CMCE	0.5053	0.6855	0.7206	0.445			CMCE	0.4928	0.6928	0.732	0.444
	BBU	CMPM	0.5163	0.7248	0.7629	0.4691		BBU	CMPM	0.539	0.7361	0.7614	0.4728
		CMPC	0.5061	0.6864	0.7263	0.4501			CMPC	0.4719	0.6928	0.7342	0.4487
		CMCE	0.4845	0.6849	0.7351	0.4495			CMCE	0.5049	0.6832	0.7475	0.4554
	BBNMT	CMPM	0.5235	0.7307	0.7544	0.4738		BBNMT	CMPM	0.5372	0.7483	0.7579	0.4668
		CMPC	0.4658	0.7048	0.71	0.4527			CMPC	0.4661	0.6963	0.7203	0.453
		CMCE	0.4954	0.685	0.7418	0.457			CMCE	0.5059	0.7018	0.7121	0.4542
	MLHU	CMPM	0.528	0.7433	0.7504	0.4718		MLHU	CMPM	0.5179	0.7382	0.7538	0.4748
		CMPC	0.511	0.6861	0.7487	0.4427			CMPC	0.467	0.6904	0.7472	0.4437
		CMCE	0.4711	0.7021	0.7375	0.4466			CMCE	0.4763	0.695	0.7157	0.4417
ResNet34	AMBV2	CMPM	0.5343	0.7146	0.7563	0.462	EfficientNetB7	AMBV2	CMPM	0.5306	0.7316	0.7532	0.4667
		CMPC	0.4707	0.6952	0.7284	0.4426			CMPC	0.4775	0.6987	0.709	0.4577
		CMCE	0.5114	0.6889	0.7418	0.4487			CMCE	0.4653	0.6977	0.7362	0.4546
	PMML12V2	CMPM	0.5291	0.7386	0.7624	0.477		PMML12V2	CMPM	0.5199	0.7079	0.7522	0.472
		CMPC	0.4891	0.6828	0.7311	0.4595			CMPC	0.4772	0.6955	0.7415	0.4572
		CMCE	0.4653	0.6818	0.7175	0.4476			CMCE	0.4958	0.6807	0.7407	0.4547
	DBMCV2	CMPM	0.5253	0.7261	0.7615	0.4652		DBMCV2	CMPM	0.5266	0.7357	0.7581	0.46
		CMPC	0.5136	0.6992	0.716	0.4578			CMPC	0.5117	0.6813	0.7408	0.4417
		CMCE	0.4842	0.7006	0.7447	0.4422			CMCE	0.5017	0.698	0.7231	0.4493
	BBU	CMPM	0.535	0.7184	0.7572	0.4694		BBU	CMPM	0.538	0.7154	0.7604	0.4702
		CMPC	0.4667	0.6954	0.7167	0.4511			CMPC	0.4839	0.7022	0.7078	0.4549
		CMCE	0.5033	0.7013	0.7084	0.4528			CMCE	0.4899	0.6925	0.7403	0.4469
	BBNMT	CMPM	0.5308	0.7263	0.7626	0.4702		BBNMT	CMPM	0.529	0.749	0.7523	0.4715
		CMPC	0.5105	0.7045	0.7291	0.4464			CMPC	0.4736	0.7037	0.7181	0.4585
		CMCE	0.4839	0.7024	0.7058	0.4516			CMCE	0.5014	0.6832	0.7465	0.452
	MLHU	CMPM	0.5355	0.7058	0.7675	0.4755		MLHU	CMPM	0.5356	0.7154	0.7693	0.4772
		CMPC	0.4921	0.6859	0.7071	0.4479			CMPC	0.514	0.6931	0.7072	0.4545
		CMCE	0.4948	0.6885	0.7165	0.458			CMCE	0.508	0.7006	0.7096	0.4483

search on GTMFLPIS can be divided into six main parts: pre-preparation of the experiment, reproduction, optimization of the English GTMFLPIS model, implementation and optimization of the Chinese GTMFLPIS model, implementation of the bilingual GTMFLPIS model architecture, and analysis of the experimental results. The pre-preparation for the experiment consists of labelling the dataset and building the computer environment for training the model. In the reproduction and optimization section of the English GTMFLPIS model, this research reproduces the RN50_AMBV2 model and obtains the RN50_BBU model, which yields better performance by substituting different model combinations. In the implementation and optimization part of the Chinese GTMFLPIS model, this research achieved the RN50_CMB model and obtained the RN50_BBU model and RN50_PMML12V2 model, which yield better performance by substituting different model combinations. In the implementation part of the bilingual GTMFLPIS model architecture, the linguistic dataset of this research was trained in Chinese and English.

As shown in Figure 4, we replace the pooling layer with an adaptive pooling layer and change the loss function to improve the performance of the bilingual GTMFLPIS model. Finally, we will compare and analyze the experimental results. Table 3 shows the models and loss functions used in the GTMFLPIS model. Detailed cross-experiments will be conducted using models and loss functions listed in Table 3. Table 3 shows four different loss functions: CMPM, CMPC [20], and CMCE [21] are mentioned in existing research. CCPPM is the loss function that we

Table 6: *Chinese GTMFLPIS model training results.*

Image Model	Text Model	Loss Function	Rank1	Rank5	Rank10	mAP	Image Model	Text Model	Loss Function	Rank1	Rank5	Rank10	mAP
ResNet50	CMB	CCMPM	0.5519	0.7538	0.8052	0.493	VGG16Net	CMB	CCMPM	0.5566	0.7635	0.8279	0.5048
		CMPM	0.5383	0.7121	0.7581	0.4762			CMPM	0.5398	0.7428	0.7575	0.4756
		CMPC	0.4939	0.6913	0.7098	0.4546			CMPC	0.4943	0.6801	0.7285	0.4401
		CMCE	0.4929	0.6913	0.7153	0.4491			CMCE	0.4831	0.6908	0.7211	0.4583
	PMML12V2	**CCMPM**	**0.6055**	**0.7998**	**0.8599**	**0.5246**		PMML12V2	CCMPM	0.5415	0.7658	0.8529	0.4877
		CMPM	0.5249	0.7119	0.7604	0.4622			CMPM	0.5203	0.7117	0.7571	0.4696
		CMPC	0.4767	0.6992	0.7066	0.4462			CMPC	0.4821	0.7015	0.7433	0.4514
		CMCE	0.4666	0.6988	0.7289	0.4552			CMCE	0.4681	0.6882	0.734	0.445
	DBMCV2	CCMPM	0.569	0.7692	0.8478	0.4944		DBMCV2	CCMPM	0.5439	0.7583	0.7735	0.4801
		CMPM	0.5331	0.7426	0.7515	0.4632			CMPM	0.5386	0.7315	0.7625	0.478
		CMPC	0.4887	0.7014	0.7143	0.4416			CMPC	0.5039	0.6828	0.738	0.452
		CMCE	0.4697	0.7049	0.7103	0.4538			CMCE	0.4675	0.6994	0.714	0.442
ResNet18	CMB	CCMPM	0.5539	0.7629	0.8396	0.5041	EfficientNetV2	CMB	CCMPM	0.5409	0.7507	0.8438	0.4881
		CMPM	0.5392	0.7076	0.7525	0.4697			CMPM	0.5306	0.708	0.7521	0.4697
		CMPC	0.5146	0.687	0.7313	0.4486			CMPC	0.512	0.698	0.721	0.4422
		CMCE	0.4819	0.6879	0.7159	0.4516			CMCE	0.4997	0.6945	0.737	0.4436
	PMML12V2	CCMPM	0.5516	0.7638	0.7889	0.4955		PMML12V2	CCMPM	0.5478	0.757	0.7761	0.4935
		CMPM	0.5196	0.7304	0.7628	0.4654			CMPM	0.52	0.7081	0.7692	0.4777
		CMPC	0.4946	0.6888	0.7199	0.4491			CMPC	0.4865	0.6968	0.7208	0.4525
		CMCE	0.4907	0.6963	0.7106	0.4413			CMCE	0.4899	0.6971	0.7289	0.4407
	DBMCV2	CCMPM	0.5404	0.7592	0.8131	0.5025		DBMCV2	CCMPM	0.5569	0.7662	0.7833	0.505
		CMPM	0.5355	0.7143	0.7517	0.4684			CMPM	0.53	0.7273	0.7698	0.4701
		CMPC	0.4953	0.6955	0.7317	0.4582			CMPC	0.5032	0.7008	0.7341	0.4552
		CMCE	0.5053	0.6855	0.7206	0.445			CMCE	0.4928	0.6928	0.732	0.444
ResNet34	CMB	CCMPM	0.5632	0.77	0.8165	0.4977	EfficientNetB7	CMB	CCMPM	0.5573	0.7622	0.8599	0.4948
		CMPM	0.5326	0.7132	0.7634	0.4712			CMPM	0.5388	0.7481	0.764	0.4647
		CMPC	0.5127	0.6832	0.7299	0.4442			CMPC	0.5025	0.6892	0.7432	0.4425
		CMCE	0.4921	0.7016	0.705	0.4448			CMCE	0.4615	0.692	0.7268	0.4448
	PMML12V2	CCMPM	0.5659	0.7669	0.8591	0.4925		PMML12V2	CCMPM	0.5673	0.7553	0.7946	0.5095
		CMPM	0.5291	0.7386	0.7624	0.477			CMPM	0.5199	0.7079	0.7522	0.472
		CMPC	0.4891	0.6828	0.7311	0.4595			CMPC	0.4772	0.6955	0.7415	0.4572
		CMCE	0.4653	0.6818	0.7175	0.4476			CMCE	0.4958	0.6807	0.7407	0.4547
	DBMCV2	CCMPM	0.5665	0.7634	0.8212	0.4954		DBMCV2	CCMPM	0.5505	0.7671	0.8126	0.4837
		CMPM	0.5253	0.7261	0.7615	0.4652			CMPM	0.5266	0.7357	0.7581	0.46
		CMPC	0.5136	0.6992	0.716	0.4578			CMPC	0.5117	0.6813	0.7408	0.4417
		CMCE	0.4842	0.7006	0.7447	0.4422			CMCE	0.5017	0.698	0.7231	0.4493

Table 7: *Bilingual GTMFLPIS model training results.*

Image Model	Text Model	Loss Function	Rank1	Rank5	Rank10	mAP	Image Model	Text Model	Loss Function	Rank1	Rank5	Rank10	mAP
ResNet50	PMML12V2	CCMPM	0.5436	0.7554	0.8124	0.4881	VGG16Net	PMML12V2	CCMPM	0.5421	0.7538	0.8494	0.4952
		CMPM	0.5197	0.7132	0.7516	0.4632			CMPM	0.5303	0.7292	0.7657	0.4787
		CMPC	0.5042	0.6838	0.7208	0.443			CMPC	0.4633	0.685	0.732	0.4597
		CMCE	0.4993	0.6825	0.7375	0.4568			CMCE	0.4684	0.6955	0.7075	0.4571
	DBMCV2	CCMPM	**0.6264**	**0.7787**	**0.8331**	**0.5595**		DBMCV2	CCMPM	0.5558	0.7619	0.7708	0.487
		CMPM	0.522	0.7391	0.7544	0.4751			CMPM	0.5158	0.7452	0.7683	0.4795
		CMPC	0.462	0.6926	0.7137	0.4524			CMPC	0.4966	0.7	0.7162	0.4539
		CMCE	0.4912	0.6929	0.7225	0.4432			CMCE	0.5089	0.6896	0.7221	0.4491
ResNet18	PMML12V2	CCMPM	0.5484	0.761	0.8527	0.5003	EfficientNetV2	PMML12V2	CCMPM	0.5438	0.7675	0.8252	0.5094
		CMPM	0.5323	0.7179	0.7612	0.4643			CMPM	0.5199	0.7121	0.7598	0.4678
		CMPC	0.4675	0.6891	0.7137	0.4443			CMPC	0.5064	0.6819	0.7422	0.4468
		CMCE	0.4972	0.7011	0.7278	0.457			CMCE	0.4718	0.6912	0.7185	0.4533
	DBMCV2	CCMPM	0.5619	0.7607	0.7868	0.4868		DBMCV2	CCMPM	**0.6229**	**0.7759**	**0.8322**	**0.5543**
		CMPM	0.5166	0.7377	0.7637	0.4744			CMPM	0.5197	0.732	0.769	0.4755
		CMPC	0.4731	0.6997	0.7257	0.4547			CMPC	0.4611	0.6835	0.7418	0.4433
		CMCE	0.4899	0.6949	0.7277	0.4495			CMCE	0.4655	0.7023	0.7495	0.4409
ResNet34	PMML12V2	CCMPM	0.5422	0.7652	0.8175	0.5032	EfficientNetB7	PMML12V2	CCMPM	0.5401	0.7655	0.7717	0.4885
		CMPM	0.5285	0.7365	0.7636	0.4605			CMPM	0.5353	0.7322	0.7563	0.4734
		CMPC	0.4608	0.6933	0.7242	0.4421			CMPC	0.4904	0.6965	0.7088	0.4446
		CMCE	0.4924	0.6876	0.7326	0.4568			CMCE	0.5013	0.6822	0.7115	0.4529
	DBMCV2	CCMPM	0.5626	0.7548	0.8267	0.5039		DBMCV2	CCMPM	0.5502	0.7622	0.8551	0.4885
		CMPM	0.5239	0.726	0.7678	0.4668			CMPM	0.517	0.7376	0.7513	0.4794
		CMPC	0.4991	0.699	0.7089	0.4495			CMPC	0.4979	0.6806	0.7171	0.4509
		CMCE	0.4854	0.6896	0.711	0.4402			CMCE	0.4938	0.6863	0.7071	0.4493

have optimized and designed.

2.4 Metrics for Model Evaluation

This research can be divided into model training and model retrieval validation. In the model training part, we use Rank-K as the evaluation metric, one of the standard metrics for evaluating the performance of sorting algorithms [23]. Rank-K shows the accuracy of the first K retrieved targets in the sorted list returned by the algorithm. Meanwhile, we calculate the Precision and Recall of the retrieval targets and combine them with the Rank value to calculate an Average Precision (AP) value [24]. Finally, we calculate the mAP value by averaging all categories [24]. The mAP measures the accuracy of the model over all categories. In addition, we conducted actual PIS retrieval using different pedestrian information retrieval models.

We selected a fixed test dataset [25] and designed a description for the retrieval task. Our retrieval task will perform descriptive accuracy statistics on the colour and length of hair, tops, and pants to calculate the PIS accuracy(PISA). In the model retrieval validation section, we record the attainment rate for the same retrieval task using different models.

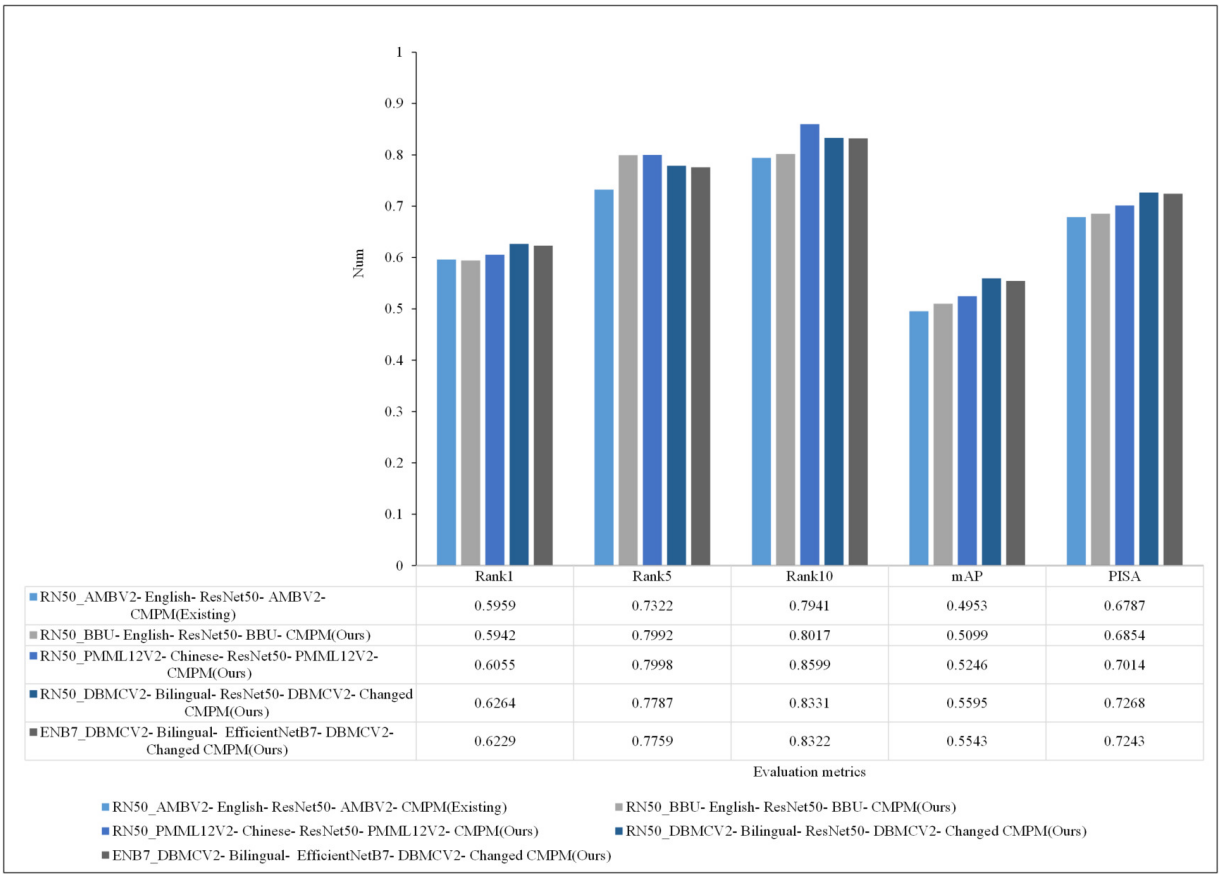

Fig.5: *Verification results of actual pedestrian search tasks.*

3. RESULTS AND DISCUSSION

3.1 Analysis of GTMFLPIS Model Training Results

3.1.1 Analysis of the training results of the English GTMFLPIS model

The training of the English GTMFLPIS model consists of an experimental replication of the existing English GTMFLPIS model and an English GTMFLPIS model optimization process. The optimization approach used in this research is to use different combinations of neural networks and loss functions to improve the performance of the model. We found that during the experimental process, there may be fluctuations in the results of individual experiments. We analyzed the experimental results and found that the changes were within a certain range. Therefore, we have used the average of five experimental results to evaluate the model.

The combination of English GTMFLPIS models trained for this research is shown in Table 5. Our proposed optimization approach is practical. One of our GTMFLPIS models outperforming existing research is the RN50_BBU model, which consists of ResNet50, BBU, and CCMPM loss function. Based on the overall training results, the CMPM loss function shows better training results than the other two loss functions. We created CCMPM based on CMPM to improve the training outcomes of the model.

3.1.2 Analysis of the training results of the Chinese GTMFLPIS model

In this research, we trained the Chinese GTMFLPIS model by replacing the text pre-training model and the training dataset according to the training architecture of the English GTMFLPIS model. After achieving the training of the Chinese GTMFLPIS model, we optimized the Chinese GTMFLPIS model according to the optimization of the English GTMFLPIS model. At the same time, we have readjusted and optimized the CMPM loss function that performed well in English training from section 3.1.1. The results of the GTMFLPIS models trained under different combinations are shown in Table 6.

As shown in Table 6, the Chinese GTMFLPIS models that perform better in our trained GTMFLPIS models are ResNet50, PMML12V2, and CCMPM loss functions.

3.1.3 Analysis of the training results of the bilingual GTMFLPIS model

In this research, we propose a bilingual GTMFLPIS model training architecture to address the situation where a single model cannot handle multiple

languages. The model training results of our trained bilingual GTMFLPIS model are shown in Table 7.

As shown in Table 7, the GTMFLPIS model that performs better in the results of our trained GTMFLPIS model is the RN50_DBMCV2 model with the combination of ResNet50, DBMCV2, and CMPM loss functions. Meanwhile, ENB7_DBMCV2, combined with EfficientNetB7, DBMCV2, and CMPM loss function, is also a good performing model. The performance of the ENB7_DBMCV2 model could be better than the RN50_DBMCV2 model if train efficiency is considered.

In addition, we adjusted the loss function to fit the three-input bilingual GTMFLPIS model better. The performance of the RN50_DBMCV2 model and the ENB7_DBMCV2 model with modified loss functions is slightly improved. The RN50_DBMCV2 model and ENB7_DBMCV2 model proposed in this study have their advantages regarding retrieval performance and efficiency, and they are the first available methods to handle the problem of multilingual cross-modal pedestrian information search.

Experimental training results show that using different combinations of image models, text models, and loss functions can improve the training performance of the model. Meanwhile, we can observe that the optimized loss function can steadily improve the training performance of the model. The CCMPM loss function increases the robustness of the model.

3.2 Analysis of the Results of Validation of GTMFLPIS Model for the Pedestrian Information Search Task

The way to validate the model retrieval performance is to judge it by the search results of actual pedestrian information search tasks. We created two hundred different pedestrian information search tasks. We used different models to perform the pedestrian search and recorded the number of times the retrieval results of each model matched. In this research, we have selected only a few models that performed well in the training results for this actual pedestrian search experiments. The accuracy of pedestrian information search for different models is shown in Figure 5.

As shown in Figure 5, the pedestrian information search accuracy of several novel models proposed in this study is slightly improved over the accuracy of the existing models.

4. CONCLUSIONS

This research is a study to enhance the performance and application scope of Pedestrian Information Retrieval (PIS) models. The main contributions of our research are optimizing the existing English PIS models, achieving the training of PIS models in Chinese, and implementing the training of PIS models in Chinese and English. At the same time, we labelled the Chinese PIS dataset and designed and optimized the loss function in the model tuning process. This research proposes a cross-modal pedestrian information search method based on a language-based graph-text fusion model.

We replicated existing research and used the combinatorial substitution method to find possible model combinations which yield better performance. One hundred and eight combinations of English cross-modal pedestrian information search methods are trained and tested. The English RN50_BBU model with improved performance is proposed in this research. In addition, the Chinese GTMF-LPIS_RN50_PMML12V2 model is proposed utilizing our previously constructed Chinese CUHK-PEDES dataset. Seventy-two combinations for the Chinese cross-modal pedestrian information search methods are trained and tested. The Chinese RN50_PMML12V2 model performs better than the existing English model.

In the meantime, to solve the limitation of a single model for handling multiple languages, we propose a training architecture for multilingual GTMFLPIS models. We constructed a bilingual dataset using our Chinese CUHK-PEDES dataset and the existing English CUHK-PEDES dataset, which contains 201,030 strings and 40,206 images. Forty-eight combinations for multilingual cross-modal pedestrian information search methods are trained and tested. We found that the loss function significantly impacts the model performance during the training process of the multilingual GTMFLPIS model. Therefore, we adjusted the loss function to improve the performance of the model. The performance and robustness of the multilingual GTMFLPIS model proposed in this study have been improved.

Future research can explore the language specification in more depth to improve the performance of pedestrian information retrieval.

ACKNOWLEDGEMENT

The first author received full scholarship from CPALL while conducting this research in PIM.

The first and second authors each contributed 50% equally to this work. The second author is the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.X. and J. Q.; methodology, Y.X. and J. Q.; software, Y.X. and J. Q.; validation, Y.X. and J. Q.; formal analysis, Y.X. and J. Q.; investigation, Y.X. and J. Q.; data curation, Y.X. and J. Q.; writing—original draft preparation, Y.X. and J. Q.; writing—review and editing, Y.X. and J. Q.; visualization, Y.X. and J. Q.; supervision, J. Q.; All authors have read and agreed to the published version of the manuscript.

References

[1] S. Yang *et al.*, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4492-4501, 2023.

[2] Y. Zhao *et al.*, "Learning deep part-aware embedding for person retrieval," *Pattern Recognition*, vol. 116, no. 107938, 2021.

[3] S. Kakouros, T. Stafylakis, L. Mošsner and L. Burget, "Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing," *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[4] X. Zang, G. Li and W. Gao, "Multidirection and Multiscale Pyramid in Transformer for Video-Based Pedestrian Retrieval," in *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 8776-8785, Dec. 2022.

[5] P. Kaur, H. S. Pannu and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Computer Science Review*, vol. 39, no. 100336, 2021.

[6] Z. Yuan *et al.*, "Remote sensing cross-modal text-image retrieval based on global and local information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-16, 2022.

[7] R. Alsaleh, T. Sayed and M. H. Zaki, "Assessing the effect of pedestrians'use of cell phones on their walking behavior: A study based on automated video analysis," *Transportation research record*, vol. 2672, no. 35, pp. 46-57, 2018.

[8] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2787-2797, 2023.

[9] Y. Zhao, H. Cheng and C. Huang, "Cross-modal pedestrian re-recognition based on attention mechanism," *The Visual Computer*, vol. 40, pp. 2405-2418, 2023.

[10] T. Zhou, S. Ruan and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3-4, no. 100004, 2019.

[11] D. Theckedath and R. R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *SN Computer Science*, vol. 1, no. 79, 2020.

[12] M. S. Majib, M. M. Rahman, T. M. S. Sazzad, N. I. Khan and S. K. Dey, "VGG-SCNet: A VGG Net-Based Deep Learning Framework for Brain Tumor Detection on MRI Images," in *IEEE Access*, vol. 9, pp. 116942-116952, 2021.

[13] Ü. Atila, M. Uçar, K. Akyol and E. Uçar, "Plant leaf disease classification using EfficientNet deep learning model," *Ecological Informatics*, vol. 61, no. 101182, 2021.

[14] L. C. Passaro *et al.*, "UNIPI-NLE at CheckThat! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers," *CEUR Workshop Proceedings*, vol. 2696, 2020.

[15] R. A. Frick and I. Vogel, "Fraunhofer SIT at CheckThat!-2022: Ensemble Similarity Estimation for Finding Previously Fact-Checked Claims," *CEUR Workshop Proceedings*, vol. 3180, 2022.

[16] H. Dang, K. Lee, S. Henry and Ö Uzuner, "Ensemble BERT for classifying medication-mentioning tweets," *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pp. 34-41, 2020.

[17] J. Kapočiūtė-Dzikienė, A. Salimbajevs and R. Skadiņš, "Monolingual and cross-lingual intent detection without training data in target languages," *Electronics*, vol. 10, no.12:1412, 2021.

[18] J. F. Rodríguez, A. J. Fernández-García and E Verdú, "Semantic Similarity Between Medium-Sized Texts," *Management of Digital EcoSystems*, pp. 361-373, 2024.

[19] S. Mishra and S. Mishra, "3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages," *CEUR Workshop Proceedings (FIRE 2019)*, vol. 2517, 2019.

[20] Y. Chen *et al.*, "Tipcb: A simple but effective part-based convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171-181, 2022.

[21] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," *Proceedings of the European conference on computer vision (ECCV)*, pp. 1-16, 2018.

[22] F. Shen *et al.*, "Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval," *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8922-8931, 2023.

[23] Z. LIU and P. WAN, "Pedestrian re-identification feature extraction method based on attention mechanism," *Journal of Computer Applications*, vol. 40, no. 3, p. 672, 2023.

[24] L. Zhang *et al.*, "Cross-modality interactive attention network for multispectral pedestrian detection," *Information Fusion*, vol. 50, pp. 20-29, 2019.

[25] Xie and Qu, "A Study on Chinese Language Cross-Modal Pedestrian Image Information Retrieval," *Songklanakarin Journal of Science and Technology (SJST)*, 2024.

Yan Xie is currently studying for the Master of Engineering Technology, Faculty of Engineering and Technolog-y, 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). She is awarded with a full scholarship from CPALL while conducting her research in PIM.

Jian Qu is an Assistant professor at the Faculty of Engineering and Technology, Panyapiwat Institute of Management. He received Ph.D. with 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.