Pattern Recognition 161 (2025) 111247

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Local-enhanced representation for text-based person search

Guoqing Zhang [a,b], Yuhao Chen [a], Yuhui Zheng [a], Gaven Martin [c], Ruili Wang [b,d,*]

[a] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China
[b] School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand
[c] Institute for Advanced Study, Massey University, Auckland, New Zealand
[d] School of Computer Science, University of Nottingham Ningbo China, Ningbo, China

ARTICLE INFO

Keywords:
Person re-identification
Cross-modal retrieval
Local representation

ABSTRACT

Text-based person search is a critical task in intelligent security, designed to locate a person of interest by text descriptions. The primary challenge in this task is to effectively bridge the significant gap between the text and image domains while simultaneously extracting the discriminative features that are crucial for the accurate identification of individuals. Existing methods have made some effective attempts by conducting cross-modal matching at the fine-grained representation level. However, these approaches frequently overlook two crucial factors: (i) the presence of noise in the local features during information fusion, and (ii) the lack of intra-modal matching when measuring feature similarity. To address the above issues, we propose a novel local-enhanced representation framework in this paper. Specifically, to restrain noises in local features, we design a Relation-based cross-modal local-enhanced fusion module, which can filter out weak related information by relation assessment. In addition, we explore an intra-cross modal projection strategy to overcome the limitations of existing cross-modal projection methods. This strategy jointly applies the intra-modal and cross-modal matching constrains in feature distribution. Finally, experiments on three mainstream datasets verify the performance superiority of our proposed method compared to existing state-of-the-art methods.

1. Introduction

Recently, computer vision technology has received significant attention and development in the field of public security, especially in tasks related to person. These tasks include but are not limited to pedestrian pose estimation [1], crowd counting [2] and person re-identification (ReID) [3]. In particular, person ReID, as a core component of intelligent video surveillance systems, has received widespread attention and has made remarkable achievements in technological progress, aiming at retrieving a specific target individual from a network of multiple surveillance cameras. Currently, most person ReID methods focus on obtaining visual information, such as person images and video sequences. However, these methods have certain limitations in specific criminal investigation situations. For example, a police sometimes need to find suspects based on the descriptions of witnesses rather than relying solely on image or video materials. Accordingly, the task of person ReID has been extended to the text-based person search field, intended to identify target pedestrians based on the appearance features of pedestrians described in natural language.

Different from traditional retrieval scenarios, text-based person search task faces a unique challenge: it must overcome the difficulty of image-based person ReID and accurate cross-modal matching. In complex video surveillance systems, a series of interference items need to be solved that can seriously distort the visual representation of a person, including but not limited to background noise [4], partial body occlusions [5], fluctuation in lighting conditions [6,7] and diversity of postures [8]. These small differences in conditions can significantly affect the reliability of visual cues. Besides, text descriptions are relatively coarse in capturing the details of person appearance, because similar clothes or stripes are normally indistinguishable in language descriptions, resulting in small intra-modality feature variances between different persons [9]. Furthermore, the difference between modalities brings huge domain gap in feature distribution, which causes serious confusion in distance measurement.

Many attempts have been made to overcome these difficulties and extract high-discriminative and well-matched features from both visual and textural modalities [10,11]. A significant portion of these efforts adopt local-matching strategies which can align visual and textual features at a fine-grained level. Compared with the global-matching strategy, these methods are able to mine key features of a person and retrieve them from a more detailed perspective and they can

* Corresponding author at: School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand.
E-mail addresses: guoqingzhang@nuist.edu.cn (G. Zhang), chinayhchen@gmail.com (Y. Chen), zheng_yuhui@nuist.edu.cn (Y. Zheng), G.J.Martin@massey.ac.nz (G. Martin), ruili.wang@massey.ac.nz (R. Wang).

https://doi.org/10.1016/j.patcog.2024.111247
Received 17 July 2024; Received in revised form 30 September 2024; Accepted 25 November 2024
Available online 5 December 2024

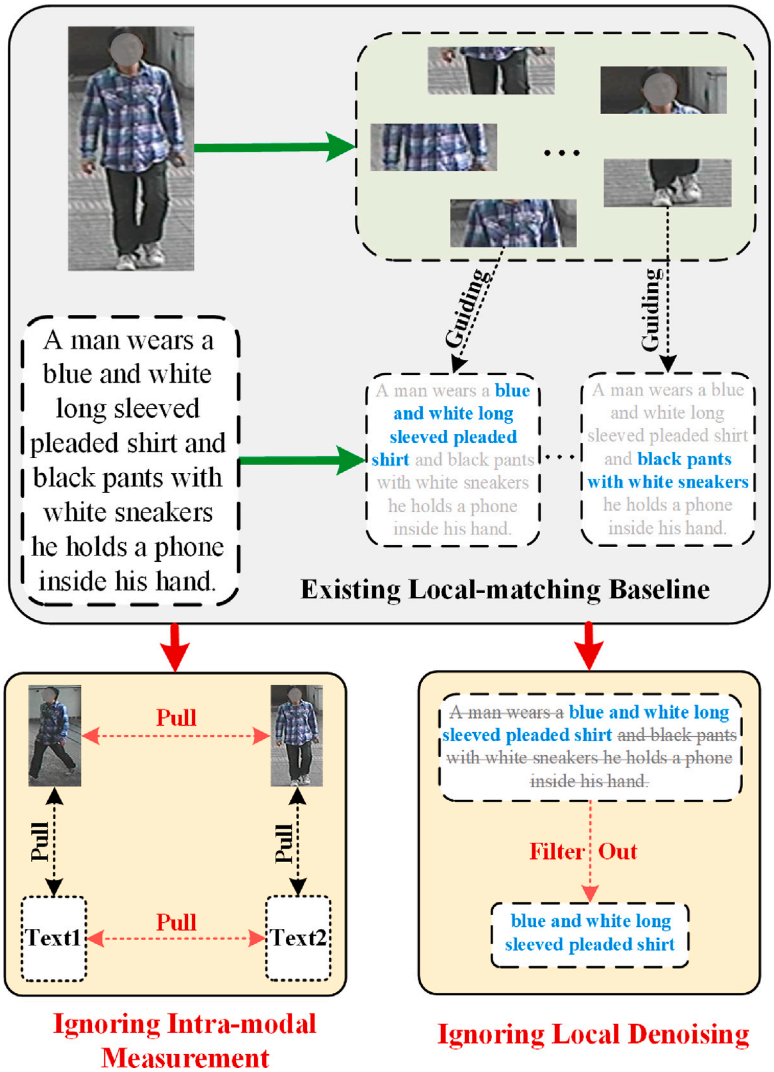

Fig. 1. The illustration of the local matching baseline of text-based person search and its limitations, including without considering intra-modal measurement and local denoising.

be mainly split into two categories: auxiliary-based and auxiliary-free local-matching methods. Specifically, the first type of methods uses additional models or multi-task learning in their frameworks, including attribute recognition [10], semantic segmentation [11] and pose estimation [12]. They can effectively conduct cross-modal local alignments but result in excessive model complexity and huge computational costs. The other methods [9,13,14] abandon these assistances and conduct local matching in a non-refined manner. For example, Chen et al. [9] proposed a simple but effective local-guiding baseline for this task (named TIPCB), which implements relevant feature filtering by narrowing the feature distribution between the two modalities.

Although this local matching strategy has a positive effect in improving retrieval performance, it does not fully consider the local denoising and intra-modal measurement, as displayed in Fig. 1. Specifically, when extracting local textual features, some irrelevant information is mixed into word embeddings, which cannot be filtered out through visual local-guiding and may reduce the discriminative ability of textual representation after information fusion. This problem is particularly evident when dealing with noisy or ambiguous textual inputs, because the model lacks the capacity to isolate meaningful features from distracting information. Besides, this strategy mainly focuses on reducing the gap between visual and textual features of the same identity, but fail to pay enough attention to sample clustering within these modalities.

To overcome the above limitations, we present a newly-designed **L**ocal-**E**nhanced **R**epresentation **F**ramework (**LERF**) in this paper. To fully leverage the efficiency and local feature extraction capabilities of the TIPCB framework without the need for additional models or intricate constraints, we adopt TIPCB as the baseline for extracting local matching representations. To reduce the irrelevant information mixed in these textual features, we explore a Relation-based Local-Enhanced Fusion (RLEF) Module instead of simple pooling fusion, which is capable to evaluate the correlation between word embeddings and textual features. In addition, during the distance measurement, we improve the original Cross-Modal Projection Matching (CMPM) Strategy [15] and propose a novel Intra-Cross Modal Projection (ICMP) Strategy. The former can only conduct feature matching of text-to-image pairs, while the latter is able to implement sample clustering in both cross-modal and intra-modal angles.

Our contributions are summarized as follows:

- A novel **L**ocal-**E**nhanced **R**epresentation **F**ramework (**LERF**) is designed to mine high-discriminative and well-matched features from person images and textual descriptions.
- A **R**elation-based **L**ocal-**E**nhanced **F**usion (**RIEF**) Module is inserted to filter out weakly correlated features such that the fused representations can obtain further information enhancement.

- An **I**ntra-**C**ross **M**odal **P**rojection (**ICMP**) Strategy is explored to refine the distance measurement, which conducts both cross-modal and intra-modal feature matching.
- Adequate qualitative and quantitative experiments are conducted on three mainstream datasets, including CUHK-PEDES [16], ICFG-PEDES [17] and RSTPReID [18], which fully demonstrate that our designed framework outperforms all the existing methods.

2. Related work

2.1. Image-text retrieval

Image-text retrieval is an important cross-modality task, which can be broadly divided into two categories according to its core ideas, including embedding visual/textual features into a shared feature space [19] and establishing accurate correspondences between visual regions and words [20,21]. Specifically, Gu et al. [19] presented to introduce generative processes into conventional cross-modality feature embedding through capturing detailed similarity between image and text modalities and learning concrete grounded representations. To mine fine-grained correspondence, Liu et al. [20] presented a novel Graph Structure Matching Network that can transform object, relation and attribute into a structured phrase. To avoid semantic confusion and loss of contextual understanding of similar objects during modal interaction, Yang et al. [21] proposed graph correlation inference and weighted adaptive filtering for local and global alignment between image text pairs.

As a sub-task of image-text retrieval, text-based person search faces more difficult challenges. For example, compared with broader text–image retrieval, the samples of text-based person search are mainly concentrated in a specific category (*i.e.*, pedestrians), and the emphasis of the text descriptions mainly focuses on individuals rather than the broader scenes. Therefore, simply adopting the aforementioned methods to the text-based person search task is not feasible for achieving satisfactory performance.

2.2. Text-based person search

Text-based person search is a cross-domain task that combines the characteristics of person ReID and image-text retrieval. The current methods can be classified into two types based on their alignment strategies: global-aligned methods [22,23], and local-aligned methods [11,24]. Global-aligned methods mainly uses different feature extraction networks, combined with different enhancement strategies, to extract global features. And then different distance metric functions are used to narrow the distance of features in the shared latent space. Li et al. [16] made the first attempt to address this challenge and introduced a gated neural attention mechanism into the recurrent neural network (RNN), which was able to perceive correlations between person images and corresponding texts. Zhang et al. [15] explored image-text representation from the angle of joint embedding learning, and designed cross-modal matching and cross-modal classification constrains. Sarafianos et al. [25] combined the cross-modal representation with the adversarial learning, which pushed the framework to generate modal-invariant features through an adversarial discriminator. Wang et al. [22] separated the color and structure information from two modalities, to balance all-round information and avoid over-reliance on color. However, these methods that only align them at the global level can easily lead to the loss of detailed information and the mixing of irrelevant noise.

To overcome these limitations, some local-aligned methods have achieved some progresses in recent years. Jing et al. [12] used pose estimation for exploring phrase-related regions in each person image. Wang et al. [11] introduced a human parsing model to catch persons' appearance components (such as heads, clothes and bags) and used K-reciprocal Sampling to align the corresponding textual phrase for each visual component. Aggarwal et al. [10] presented a multi-task learning framework to conduct cross-modal matching and attribute extracting simultaneously, so that semantics-related visual and textual representations can be mined. The auxiliary models used in the above methods bring the assistance for local-matching, but they have led to excessive model complexity and huge computational costs. Accordingly, Niu et al. [24] combined the attention mechanism with a multi-granularity strategy to conduct cross-modal alignment from global–global, global–local and local–local aspects.Considering the key cue role of mismatched region–word pairs and the problem of low similarity between matched region–word pairs, Shen et al. [26] modeled from the perspectives of negative similarity learning and positive similarity learning, effectively revealing the plausible and credible levels of contribution of pedestrian specific mismatch and matching region–word pairs towards overall similarity. In order to implement the granularity-unified representation, Shao et al. [27] designed a modal-shared dictionary to remodel modal features and projected these features into a unified format by sharing learnable prototypes. Furthermore, Chen et al. [9] proposed a local-guiding cross-modal matching strategy (TIPCB) that uses segmented person images to guide the corresponding textual information filtering. We notice that this method can achieve local matching simply and effectively to some extent, but without considering local denoising in information fusion and intra-modal matching in distance measurement, which may limit its performance.

2.3. Text–image part-based convolutional baseline

In this subsection, we provide a concise overview of the fundamental principles underlying TIPCB [9], a simple and effective dual stream framework designed for text-based image retrieval, which includes both visual and textual branches. TIPCB stands out for its lack of necessity for additional models or complex constraints, and it boasts robust capabilities for capturing local features, making it more practical.Specifically, suppose that there are N pairs of image-text pairs in the training set, denoted as $D = \left\{I_i, T_i\right\}_{i=1}^{N}$, where I_i and T_i are the ith image and text pair, respectively.

In the visual stream of the framework, the outputs generated by the third and fourth residual blocks of ResNet50 are defined as the low-stage map $f_l^v \in R^{H\times W\times C_l}$ and high-stage map $f_h^v \in R^{H\times W\times C_h}$. Here H and W are the height and width of the feature maps, while C_l and C_h represent the number of channels for the low-stage and high-stage, respectively. Subsequently, the PCB method is applied to the high-stage map to obtain visual local regions, which are denoted as $\left\{f_i^V\right\}_{i=1}^{K}$, where K is the number of regions and $f_i^V \in R^{\frac{H}{K}\times W\times C_h}$. Finally, the low-stage feature map and all regions are input into a global max-pooling layer to obtain the visual low-stage feature $v_l \in R^{C_l}$ and local features $\left\{v_i\right\}_{i=1}^{K} \left(v_i \in R^{C_h}\right)$:

$$v_l = GMP\left(f_l^v\right) \tag{1}$$

$$v_p = GMP\left(f_1^V, f_2^V, \dots, f_K^V\right) \tag{2}$$

where GMP represents a global max-pooling layer. The visual global feature $v_g \in \mathbb{R}^{C_h}$ is obtained by selecting the maximum value of each element in the channel dimension:

$$v_g = Max\left(v_p\right) \tag{3}$$

where Max represents a channel-wise max-pooling layer. Therefore, the obtained visual feature set $V = \{v_l, v_1, v_2, \dots, v_K, v_g\}$ contains low-level, local-level, and global-level representations.

In the text branch, after the text is divided into a word list and tokenized, the output is sent into the BERT model to get the preliminary word embeddings $w \in \mathbb{R}^{1\times L\times C_w}$, where L is the fixed text length and C_w is the dimension of each word embedding. Next, a single convolutional

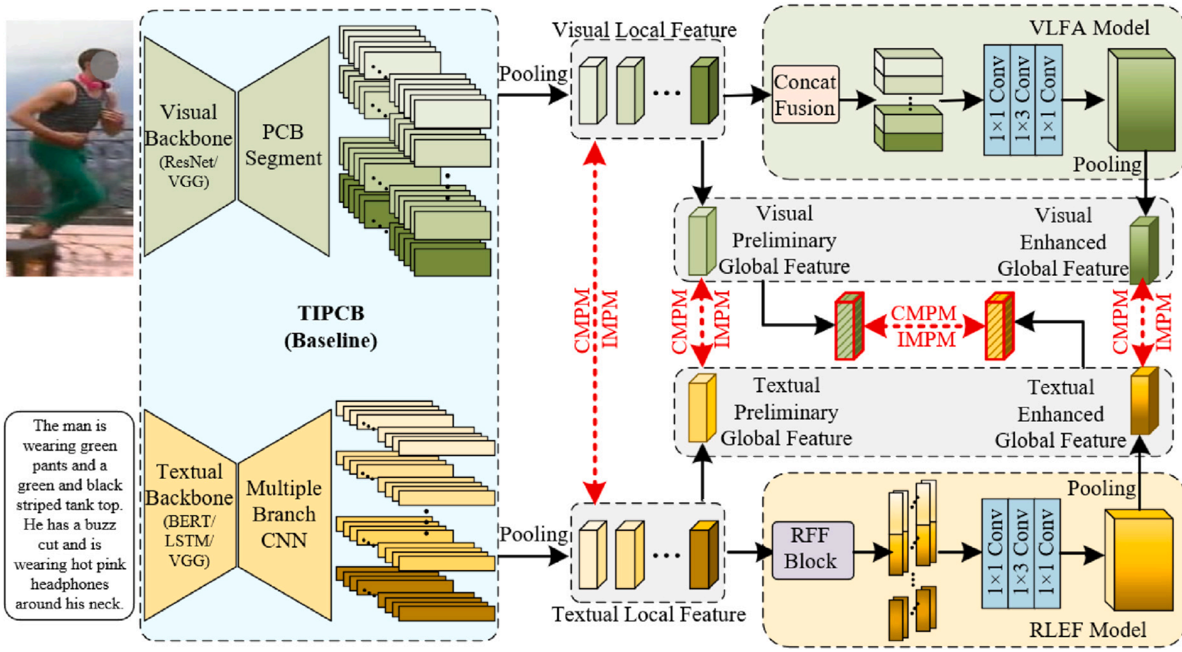

Fig. 2. The overview architecture of our proposed LERF. We use TIPBC [9] as a baseline for the text-to-image local representation. Then the visual features are sent to the Visual Local Feature Aggregation (VLFA) Module and the textual features are sent to the Relation-based Local-enhance Fusion (RLEF) Module. Finally, the Intra-cross Modal Projection Strategy is adopted to jointly conduct Cross-modal Projection Matching (CMPM) and Intra-modal Projection Matching (IMPM) constrains.

layer is used to adjust the embedding's dimension, to obtain a low-stage textual map $f_l^T \in \mathbb{R}^{1\times L\times C_h}$, which keeps the same channel dimension as the low-stage visual map. Instead of segmenting texts, the baseline introduces a K-branch convolutional network (CNN) to extract textual local features, where each branch corresponds to one region in person images. Each branch can be trained to filter the local relevant information from the word embeddings by forcibly narrowing the spatial distributions between local images and global texts. The textual branch then outputs the textual local feature maps $\{f_i^T\}_{i=1}^K$, where $f_i^T \in \mathbb{R}^{1\times L\times C_h}$. Similar to the visual branch, we can obtain the textual low-stage feature $t_l \in \mathbb{R}^{C_l}$, local features $\{t_i\}_{i=1}^K (t_i \in \mathbb{R}^{C_h})$, and global feature $t_g \in \mathbb{R}^{C_h}$. Then, the textual feature set $T = \{t_l, t_1, t_2, \dots, t_K, t_g\}$ contains low-level, local-level, and global-level representations.

Finally, the CMPM loss is used to eliminate the modality gap at the three feature levels respectively. Compared to other image and text encoding methods, TIPCB introduces an adaptive $K-1$ branch structure in the text encoding process. Unlike text deep convolution, this branch stacks some bottlenecks to avoid information loss caused by deep convolution. However, TIPCB did not consider denoising and intra-modal distance measurement in the process of local feature extraction.

3. Methodology

3.1. Framework overview

The overview of our proposed Local-Enhanced Representation Framework (LERF) is displayed in Fig. 2, which mainly consists of three components, including a local representation baseline (TIPCB [9]), a Visual Local Feature Aggregation (VLFA) Module, our designed Relation-based Local-Enhance Fusion Module (RLEF) and our designed Intra-Cross Modal Projection (ICMP) Strategy. The baseline is capable to extract preliminary visual and textual local features through a local-guiding strategy and then sent the text and visual features to the RLEF and VLFA modules, respectively. In the RLEF module, the textual local features are first enhanced by the Relation-based Feature Filter (RFF) module and then concatenated to obtain global features, which are enhanced through a bottleneck structure. Different from textual features, in the VLFA module, the visual local features are directly concatenated to obtain global features, which are also enhanced through a bottleneck structure. During the distance measurement, we adopt the multi-stage matching strategy [12], to jointly conduct the cross-modal and intra-modal matching, such that the extracted features have strong discrimination and modal compatibility.

Existing methods usually use natural language tools to extract textual local features based on attributes such as pedestrian gender, attire, accessories, etc. The number of local feature parts extracted by such methods is usually not fixed and there is no unified standard. In addition, our method fixes the number of visual local features by averagely segmenting the image. Considering that TIPCB does not require additional models or complex constraints and has a strong ability to capture local features, we build upon TIPCB to generate low-level and local features for both image and text modalities.

3.2. Data pre-processing

Following TIPCB, we can obtain the visual low-stage feature $v_l \in R^{C_l}$, visual local feature maps $\{f_i^V\}_{i=1}^K (f_i^V \in R^{\frac{H}{K}\times W\times C_h})$, visual local features $\{v_i\}_{i=1}^K (v_i \in R^{C_h})$, textual low-stage feature $t_l \in R^{C_l}$, textual local feature maps $\{f_i^T\}_{i=1}^K (f_i^T \in R^{1\times L\times C_h})$, and textual local features $\{t_i\}_{i=1}^K (t_i \in R^{C_h})$, where H, W, L, K, C_l and C_h are height, width, text length, the number of local features, low-stage channel and high-stage channel dimensions.

It is worth noting that TIPCB uses ResNet-50 to extract image features and BERT to extract text features, while our model allows the selection of image encoders (e.g., ResNet-50 [28], VGG-16 [29]) and text encoders (e.g., Bi-GRU [30], Bi-LSTM [31], BERT [32]). Therefore, the proposed method can achieve stable search results with 6 encoder combinations. This flexibility enhances adaptability across different datasets, leading to a more robust feature extraction process that better captures multi-granular representations from both modalities.

3.3. Relation-based local-enhanced fusion

In the previous stage, we extract the preliminary local features from visual and textual modalities. However, during the textual representation learning, the baseline still keeps the length of word embeddings

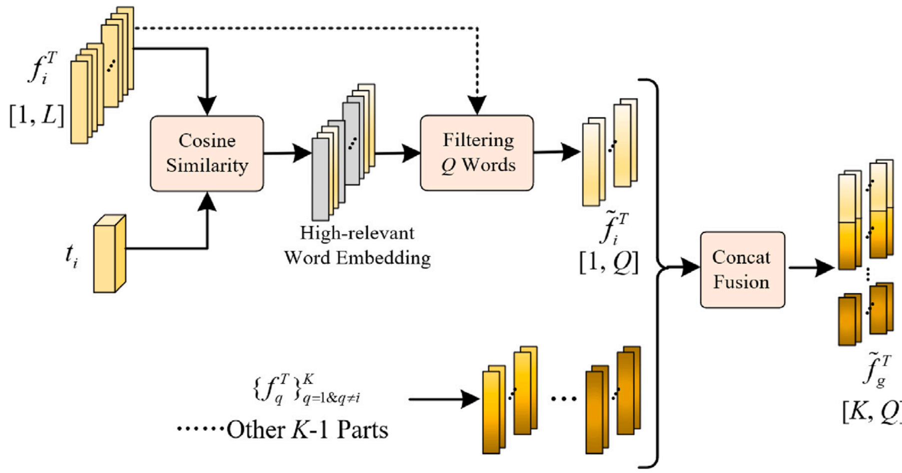

Fig. 3. The details of our proposed Relation-based Feature Filter (RFF).

instead of compressing them, resulting in the mixture of unrelated information. In order to mine features with strong relevance and improve the discrimination of the fused representation, we design a RLEF Module after the baseline.

In the textual branch, we feed the textual local features into the RFF for local enhancement, as shown in Fig. 3. For the K text local feature maps obtained from data pre-processing, we take the ith one as an example, which represented as $f_i^T \in \mathbb{R}^{1\times L\times C_h}$, and calculate word-level cosine similarity with the ith extracted local feature vector $t_i \in \mathbb{R}^{C_h}$. Then we sort the word-level features of the feature map according to the obtained similarity order. Only the first Q word-level features are filtered as the ith high-relevant feature map, which is denoted as $\tilde{f}_i^T \in \mathbb{R}^{1\times Q\times C_h}$, and the other $K-1$ parts can be calculated similarly. By removing the noise word vectors with weak relevance, the new textual feature map can achieve higher local concentration and stronger discriminative ability. Next, we fuse these high-relevant maps $\{\tilde{f}_i^T\}_{i=1}^K$ by concatenation and then get the enhanced global feature map $\tilde{f}_g^T \in \mathbb{R}^{K\times Q\times C_h}$.

Different from the textual branch, we simply concatenate the local features $\{v_i\}_{i=1}^K$ in the visual branch, because the visual regions themselves have little irrelevant information from other human parts through segmentation. Then we can also get the visual global feature map $\tilde{f}_g^V \in \mathbb{R}^{K\times 1\times C_h}$.

For each branch, we adopt a bottleneck structure to process its global feature map, which is mainly made of three convolutional layers. The first and last layers are both 1×1 convolutional layers to conduct dimension reduction and upgrading, respectively, thereby reducing the parameters of the structure. The middle layer is a 1×3 convolutional layer, which is able to further mine the global-level feature.

During the feature fusion, we consider the comprehensive utilization of two-stage information. For each modality, we first fuse the preliminary local features $\{v_i\}_{i=1}^K/\{t_i\}_{i=1}^K$ as $v_g^1/t_g^1 \in \mathbb{R}^{C_h}$ by selecting their maximum in the channel dimension. Then a global maxpooling layer is used in its global feature map to obtain the enhanced feature $v_g^2/t_g^2 \in \mathbb{R}^{C_h}$. Ultimately, the final fused feature $\tilde{v}_g/\tilde{t}_g \in \mathbb{R}^{C_h}$ can be calculated by selecting the channel maximum again, aiming to mine the significant information from features of two stages.

3.4. Intra-cross modal projection

In our designed framework, we can obtain cross-modal features of five stages, including low-stage features v_l/t_l, local features $\{v_i\}_{i=1}^K/\{t_i\}_{i=1}^K$, preliminary global features v_g^1/t_g^1, enhanced global features v_g^2/t_g^2, and fused features $\tilde{v}_g/\tilde{t}_g$. We conduct the Intra-Cross Modal Projection (ICMP) Strategy for all stages in the training phase, and only use fused features for retrieval in the testing phase.

Compared with the single cross-modal distance measurement in TIPCB [9], our LERF adopts the cross-modal and intra-modal joint measurement. As displayed in Fig. 4, our ICMP strategy mainly includes two constrains, namely the original CMPM loss [15] and the newly-designed IMPM loss. The former can reduce the domain gap between visual and textual modalities, and the latter is capable to improve the clustering effect of the same identities within each modality. It is worth noting that even though CMPM and IMPM are combined into a single loss function, they are indeed computed and monitored separately throughout the training process. This separation ensures that the intra-modal relationships governed by IMPM can be tracked independently from the cross-modal alignment of CMPM, so that the contribution of each loss can be adjusted more precisely. During the training process, the IMPM loss is applied early to stabilize intra-modal relationships, which in turn helps CMPM to more effectively align features across modalities in the later stage of training.

To illustrate these loss calculations, we simplify the marks of the five stages of features. Specifically, we denote the extract visual and textual features as $\{v_i\}_{i=1}^N$ and $\{t_i\}_{i=1}^N$. Besides, $y_{i,j}^{T2V}$, $y_{i,j}^{V2V}$ and $y_{i,j}^{T2T}$ represent the ground truth of cross-modal, visual intra-modal and textual intra-modal matching between the ith sample and the jth sample, respectively. $y_{i,j}^{T2V}/y_{i,j}^{V2V}/y_{i,j}^{T2T} = 1$ means that both samples have the same identity, while $y_{i,j}^{T2V}/y_{i,j}^{V2V}/y_{i,j}^{T2T} = 0$ indicates that they are not a matched pair.

The cross-modal matching probability between the ith text and the jth image is denoted as $p_{i,j}^{T2V} \in [0,1]$, which is calculated by:

$$p_{i,j}^{T2V} = \frac{\exp\left(\mathrm{dis}\left(t_i, v_j\right)\right)}{\sum_{k=1}^N \exp\left(\mathrm{dis}\left(t_i, v_k\right)\right)} \tag{4}$$

where dis represents the distance projection between features. Similarly, the visual intra-modal matching probability between the ith and jth images is denoted as $p_{i,j}^{V2V} \in [0,1]$, which is calculated as follows:

$$p_{i,j}^{V2V} = \frac{\exp\left(\mathrm{dis}\left(v_i, v_j\right)\right)}{\sum_{k=1}^N \exp\left(\mathrm{dis}\left(v_i, v_k\right)\right)} \tag{5}$$

The same procedure can be easily adapted to textual intra-modal matching to obtain $p_{i,j}^{T2T}$.

We use the cosine distance measurement in feature projection, and define the projection of feature a onto feature b as:

$$\mathrm{dis}(a,b) = a^\top \bar{b} \tag{6}$$

where $\bar{b}$ is the normalized vector by $\bar{b} = \frac{b}{\|b\|}$.

The calculation of feature projection is bidirectional. Therefore, we denote the probability $p_{i,j}^{T2V}$ of projecting the ith text onto the jth image as $\vec{p}{}_{i,j}^{T2V}$ and denote the probability $p_{i,j}^{T2V}$ of projecting the jth image

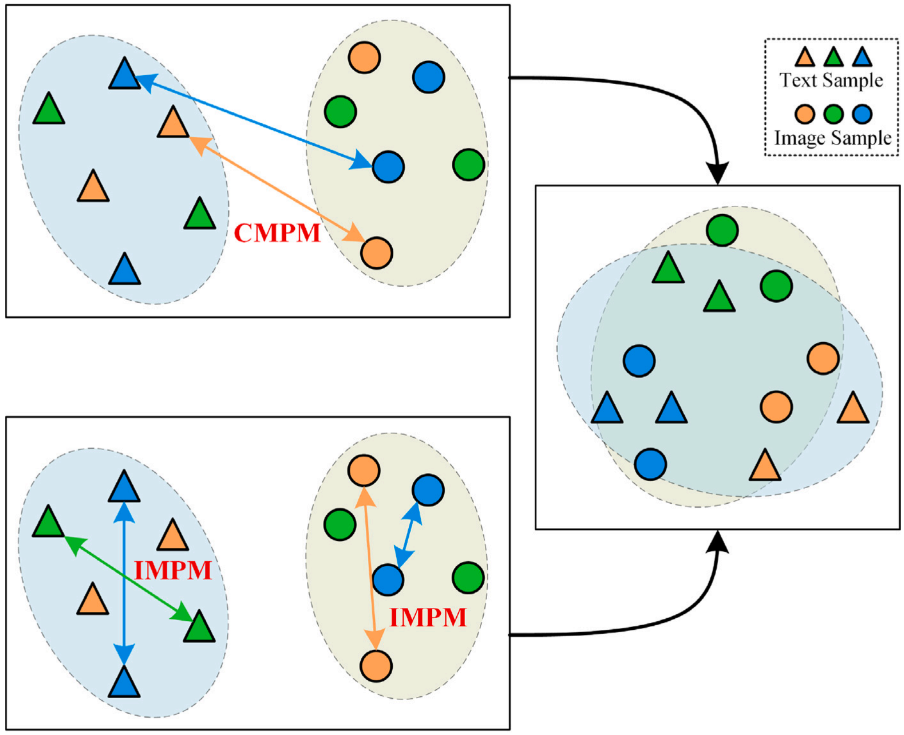

Fig. 4. The illumination of our proposed Intra-Cross Modal Projection (ICMP) Strategy. The colors of samples represent their identities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

onto the ith text as $\overleftarrow{p}_{i,j}^{T2V}$. Similarly, we can also obtain the intra-modal matching probabilities $\overrightarrow{p}_{i,j}^{V2V}$, $\overleftarrow{p}_{i,j}^{V2V}$, $\overrightarrow{p}_{i,j}^{T2T}$ and $\overleftarrow{p}_{i,j}^{T2T}$.

The CMPM loss can be calculated as:

$$L_{CMPM} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\overrightarrow{p}_{i,j}^{T2V}\log\left(\frac{\overrightarrow{p}_{i,j}^{T2V}}{\overrightarrow{y}_{i,j}^{T2V}+\varepsilon}\right)+\overleftarrow{p}_{i,j}^{T2V}\log\left(\frac{\overleftarrow{p}_{i,j}^{T2V}}{\overleftarrow{y}_{i,j}^{T2V}+\varepsilon}\right)\right) \quad (7)$$

where ε is a extremely small constant. The ground truth $y_{i,j}^{T2V}$ is normalized to $\overrightarrow{y}_{i,j}^{T2V} = \frac{y_{i,j}^{T2V}}{\sum_{k=1}^{N} y_{i,k}^{T2V}}$ when conducting the $i \rightarrow j$ projection, and $\overleftarrow{y}_{i,j}^{T2V}$ can be obtained in an opposite manner.

Similarly, we can compute the IMPM loss in textual modality by:

$$L_{IMPM}^{T2T} = \frac{2}{N}\sum_{i=1}^{N/2}\sum_{j=N/2+1}^{N}\left(\overrightarrow{p}_{i,j}^{T2T}\log\left(\frac{\overrightarrow{p}_{i,j}^{T2T}}{\overrightarrow{y}_{i,j}^{T2T}+\varepsilon}\right)+\overleftarrow{p}_{i,j}^{T2T}\log\left(\frac{\overleftarrow{p}_{i,j}^{T2T}}{\overleftarrow{y}_{i,j}^{T2T}+\varepsilon}\right)\right) \quad (8)$$

where the input text samples are guaranteed to be are paired, and the first half and the remaining half of the samples have the same labels. The same procedure can be easily conducted on the calculation of visual IMPM loss to obtain L_{IMPM}^{V2V}. Then the holistic IMPM loss can be obtained as:

$$L_{IMPM} = L_{IMPM}^{T2T} + L_{IMPM}^{V2V} \quad (9)$$

Finally, we can get our training objective function by:

$$L = L_{CMPM} + \lambda L_{IMPM} \quad (10)$$

where λ is a hyper-parameter that controls the training weight of the IMPM loss. It is worth noting that we apply the ICMP strategy at all feature stages and combine these losses by adding them together.

3.5. Model training

According to the combination of visual and textual backbones, the architecture of our method can be classified into VGG-LSTM, VGG-GRU, VGG-BERT, ResNet-LSTM, ResNet-GRU and ResNet-BERT. Due to the strong representation ability of the pretrained BERT model itself, we freeze the weight parameters of the BERT model and only train other layers for the VGG-BERT and ResNet-BERT architecture. In addition, loading the pretrained model can further accelerate the convergence of the network. Since there are no pretrained models for LSTM and GRU that can be used directly, we conduct end-to-end training on these architectures.

4. Experiments

4.1. Datasets

Adequate experiments are performed and analyzed on three relevant mainstream datasets, including CUHK-PEDES [16], ICFG-PEDES [17] and RSTPReID [18], as shown in Fig. 5. The Rank-1, Rank-5 and Rank-10 accuracies are reported to verify the performance of our model.

The CUHK-PEDES dataset is derived from five traditional image-based ReID datasets, including CUHK01, CUHK03, Market-1501, SSM and VIPER. It has totally 40,206 images of 13,003 identities, and each image is manually annotated with two corresponding texts. All these texts cover more than 9000 words, and each text averagely contains about 24 words. Among them, 11,003 identities are used for training; 1000 identities are used for validating, and the rest 1000 identities are used for testing.

The ICFG-PEDES dataset is derived from a large image-based ReID dataset, MSMT17, which contains totally 54,522 images of 4102 identities. Each image corresponds to only one text with about 37 words on average, which is more identity-centric and fine-grained than the texts in CUHK-PEDES. All these texts cover 5554 words. The ICFG-PEDES

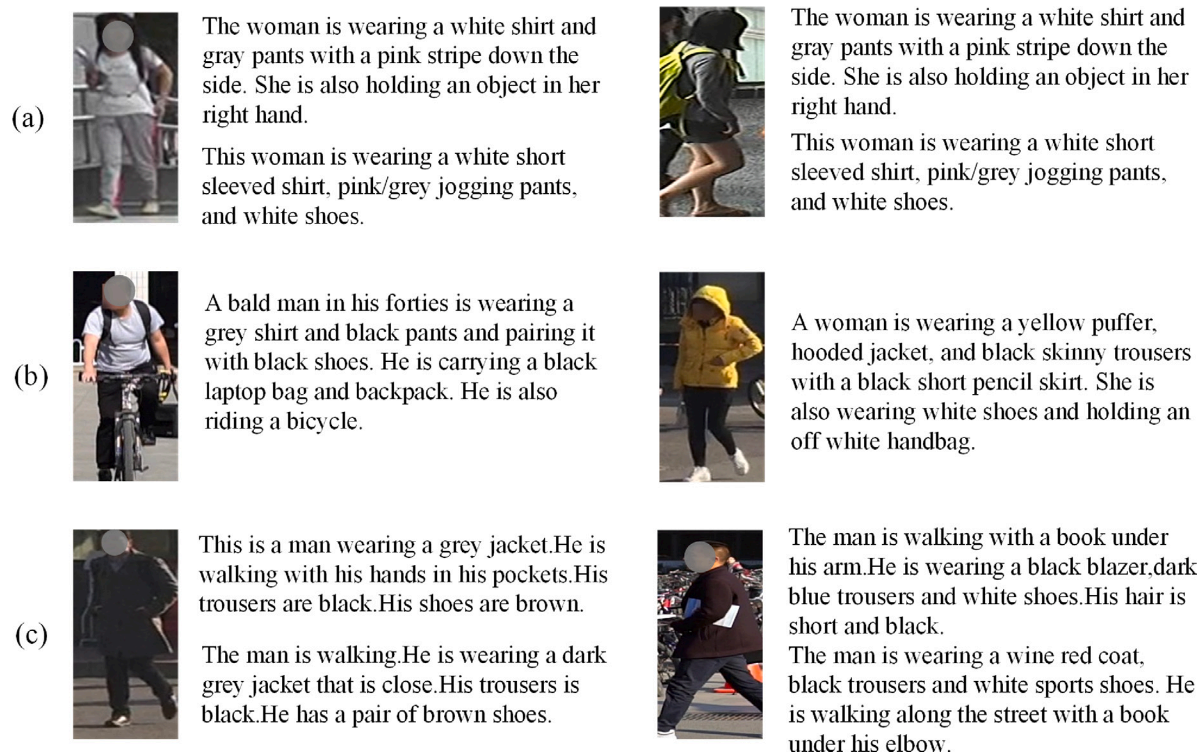

Fig. 5. Samples in the CUHK-PEDES (a), ICFG-PEDES (b) and RSTPReID (c) datasets.

Table 1
Parameter settings in different backbones.

Backbone structure	Total epochs	IMPM training epochs	LR decay epoch	C_w	C_l	C_h
VGG16 + Bi-GRU	50	12	25	1024	512	512
VGG16 + Bi-LSTM	50	12	25	512	512	512
VGG16 + BERT	60	15	35	768	512	512
ResNet50 + Bi-GRU	50	12	25	1024	1024	2048
ResNet50 + Bi-LSTM	50	12	25	512	1024	2048
ResNet50 + BERT	60	15	35	768	1024	2048

dataset is divided into a training set and testing set by the identity ratio of 3102:1000.

The RSTPReID dataset is also collected 20,505 images of 4101 identities from MSMT17. Each identity has 5 images with 2 text annotations. Each text is no shorter than 23 words, and all these texts have 2204 words which appear more than once. The RSTPReID dataset is divided into a training, validation and testing sets by 3701: 200: 200.

4.2. Implementation details

For a fair and comprehensive comparison with existing methods, we conduct experiments on most existing visual and textual backbones, including: VGG-16/ResNet-50 and BERT/Bi-LSTM/Bi-GRU. Both ResNet-50 and VGG-16 are pretrained on ImageNet, and BERT is pretrained on the Toronto Book Corpus and Wikipedia.

All input images are reshaped to 384 × 128, and all input texts are unified to L=64. The region number is set to K=6. In the training phase, the horizontally flipping with 50% probability is applied, and Adam with weight decay 4×10^{-5} is selected as optimizer. Each batch of data consists of 64 sets of samples, and each set contains 2 text–image pairs when the CMPM and IMPM losses are applied, while contains one text–image pair when just CMPM loss is applied. The learning rate is initially set to 3×10^{-3}, and then decreased by 10% after some epochs. The hyper-parameter λ in the final loss is set to 0.1. Considering that the representation or distribution of information in the same modality is more consistent than that between different modalities, and that inter-modal optimization requires balancing the learning between the two modalities, unimodal representation learning is relatively simple. This leads to the speed of convergence within modalities is significantly faster than that between modalities. Therefore, in our method, the IMPM loss is only applied for a few epochs and the weight of IMPM loss is smaller than CMPM loss, because the speed of convergence within modalities is significantly faster than that between modalities. The IMPM loss can be stabilized with a small amount of training. In addition, we list the other settings according to their backbones, as shown in Table 1. The cosine distance measurement is adopted during the testing phase. The Rank-1 and mAP curves during the training process are displayed in Fig. 6.

4.3. Comparison with state-of-the-art methods

We compare our method with a series of existing state-of-the-art methods. On the CUHK-PEDES dataset, we divide these methods into 6 groups according to their visual and textual backbones to ensure the fairness, as reported in Table 2. We can obviously find that our LERF obtains the best performance on all these groups. Especially on the VGG + GRU and ResNet + BERT structures, our method achieves 1.95% and 1.40% Rank-1 improvements. Besides, compared with the baseline TIPCB [9], our method obtains 3.10% and 1.58% Rank-1 improvements. The outstanding accuracy of our method is mainly owes to the full mining and utilization of local information, as well as the joint measurement between and within modalities.

Compared to the CUHK-PEDES dataset, the ICFG-PEDES and RSTPReID datasets are newly released and only a few methods have been

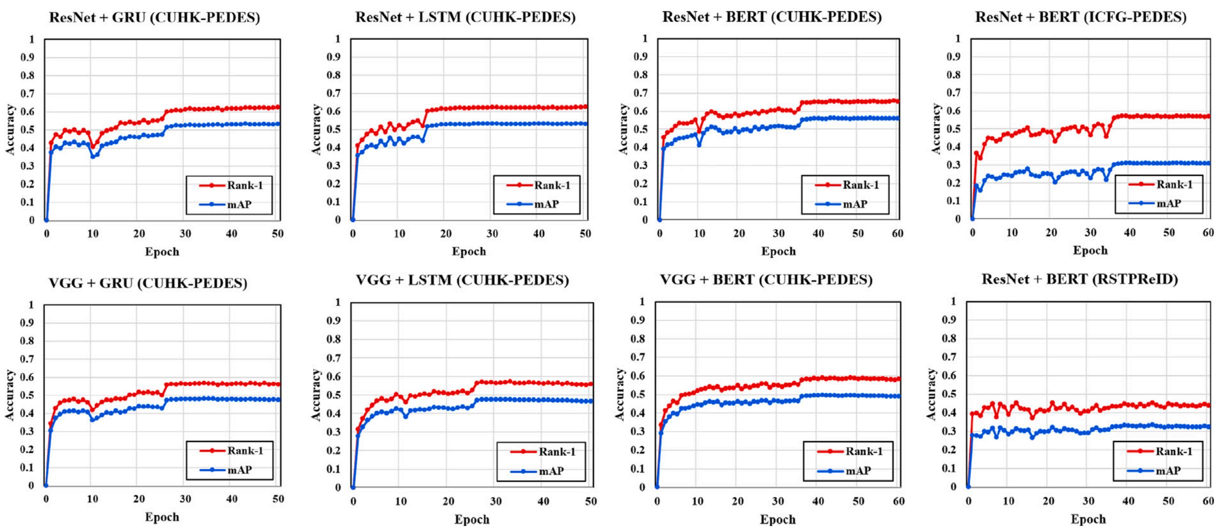

Fig. 6. The curves of Rank-1 and mAP during the training process until convergence.

Table 2
Performance comparison with existing methods on the CUHK-PEDES dataset.

Method	Ref	Image	Text	Rank-1	Rank-5	Rank-10
GNA-RNN [16]	CVPR17	VGG	LSTM	19.05	–	53.64
PMA [12]	AAAI20	VGG	LSTM	47.02	68.54	78.06
SSAN [17]	arXiv21	VGG	LSTM	55.52	76.17	83.31
TIPCB [9]	NEUCOM22	VGG	LSTM	56.86	77.89	**85.36**
Ours	–	VGG	LSTM	**57.20**	**78.23**	85.10
MIA [24]	TIP20	VGG	GRU	48.00	70.70	79.30
TIPCB [9]	NEUCOM22	VGG	GRU	54.81	76.13	83.50
Ours	–	VGG	GRU	**56.76**	**77.81**	**84.75**
TIPCB [9]	NEUCOM22	VGG	BERT	58.66	**80.04**	**86.81**
Ours	–	VGG	BERT	**59.00**	79.71	86.65
CMPM+CMPC [15]	ECCV18	ResNet	LSTM	49.37	–	79.27
PMA [12]	AAAI20	ResNet	LSTM	53.81	73.54	81.23
ViTAA [11]	ECCV20	ResNet	LSTM	55.97	75.84	83.52
MGEL [33]	IJCAI21	ResNet	LSTM	60.27	80.01	86.74
SSAN [17]	arXiv21	ResNet	LSTM	61.37	80.15	86.73
TIPCB [9]	NEUCOM22	ResNet	LSTM	62.33	81.32	87.35
Ours	–	ResNet	LSTM	**62.48**	**82.15**	**88.26**
MIA [24]	TIP20	ResNet	GRU	53.10	75.00	82.90
DSSL [18]	ACMMM21	ResNet	GRU	59.98	80.41	87.56
TIPCB [9]	NEUCOM22	ResNet	GRU	59.41	80.04	86.37
LBUL [34]	ACMMM22	ResNet	GRU	61.95	81.16	87.19
Ours	–	ResNet	GRU	**62.51**	**82.41**	**88.52**
TIMAM [25]	ICCV19	ResNet	BERT	54.51	77.56	84.78
TIPCB [9]	NEUCOM22	ResNet	BERT	64.26	83.19	89.10
LBUL [34]	ACMMM22	ResNet	BERT	64.04	82.66	87.22
CAIBC [22]	ACMMM22	ResNet	BERT	64.43	82.87	88.37
TGDA [35]	TCSVT23	ResNet	BERT	64.64	83.38	89.34
CAPL [36]	TIP24	ResNet	BERT	65.63	**84.81**	90.21
Ours	–	ResNet	BERT	**65.84**	84.24	**90.22**
Dual Path [37]	TOMM20	Other		44.40	66.26	75.07
CMAAM [10]	WACV20	Other		56.68	77.18	84.86
MANet [38]	TNNLS23	Other		65.64	83.01	88.78

Table 3
Performance comparison with existing methods on the ICFG-PEDES dataset.

Method	Ref	Rank-1	Rank-5	Rank-10
Dual Path [37]	TOMM20	38.99	59.44	68.41
CMPM+CMPC [15]	ECCV18	43.51	65.44	74.26
MIA [24]	TIP20	46.49	67.14	75.18
ViTAA [11]	ECCV20	50.98	68.79	75.78
SSAN [17]	arXiv21	54.23	72.63	79.53
TIPCB [9]	NEUCOM22	54.96	74.72	81.89
IVT [23]	ECCVW22	56.04	73.60	80.22
ASAMN [39]	TIP23	57.09	76.33	82.84
TGDA [35]	TCSVT23	**57.27**	75.19	81.80
Ours	–	57.23	**76.64**	**83.11**

Table 4
Performance comparison with existing methods on the RSTPReID dataset.

Method	Ref	Rank-1	Rank-5	Rank-10
AMEN [40]	PRCV21	38.45	62.40	73.80
DSSL [18]	ACMMM21	39.05	62.60	73.95
SSAN [17]	arXiv21	43.50	67.80	77.15
LBUL [34]	ACMMM22	43.35	66.85	76.50
Ours	–	**46.75**	**71.30**	**81.60**

Table 5
Performance comparison of different component combinations in text-to-image task.

BL	ICMP	RLEF	CUHK-PEDES			ICFG-PEDES		
			Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
✓			63.34	83.06	89.39	55.85	75.33	82.16
✓	✓		63.61	83.09	89.00	56.16	75.50	82.16
✓		✓	64.65	**84.28**	89.78	56.50	75.69	82.61
✓	✓	✓	**65.84**	84.24	**90.22**	**57.23**	**76.64**	**83.11**

conducted experiments on them. Therefore, we only report their best results regardless of backbones. The comparable results are listed in Tables 3 and 4. On the ICFG-PEDES dataset, our method achieves 2.27%, 1.92% and 1.22% improvements over TIPCB [9] on Rank-1, Rank-5 and Rank-10, respectively. On the RSTPReID dataset, our method surpasses LBUL [34] by 3.40%, 4.45% and 5.10% on Rank-1, Rank-5 and Rank-10, respectively. These significant performance improvements further prove the effectiveness and progressiveness of our approach.

4.4. Ablation study

4.4.1. Effectiveness of each component

To verify the effectiveness of our designed RLEF module and ICMP strategy, we conduct the following ablation studies as reported in Tables 5 and 6. Among them, BL (BaseLine) means the simplified TIPCB which only uses 1 bottleneck in textual convolutional branches. In the text-to-image retrieval scenario, we can find that the RLEF module obtains 1.31% and 0.65% Rank-1 improvements for the baseline on CUHK-PEDES and ICFG-PEDES, which reflects that the RLEF module

Table 6
Performance comparison of different component combinations in image-to-text task.

BL	ICMP	RLEF	CUHK-PEDES			ICFG-PEDES		
			Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
✓			74.40	92.29	96.00	57.13	79.20	85.90
✓	✓		75.57	92.39	96.42	58.81	80.41	86.82
✓		✓	75.15	**93.59**	96.84	58.17	79.83	86.53
✓	✓	✓	**76.74**	93.07	**97.01**	**60.48**	**81.30**	**87.45**

Table 7
Performance of different backbone combinations on the CUHK-PEDES dataset under different parameter settings.

Backbone structure	Total epochs	IMPM training epochs	Rank-1	Rank-5	Rank-10
ResNet50 + Bi-LSTM	50	0	56.42	78.63	84.65
ResNet50 + Bi-LSTM	50	50	50.53	73.92	81.28
ResNet50 + Bi-LSTM	50	12	62.48	82.15	88.26
ResNet50 + BERT	60	0	65.41	84.54	90.14
ResNet50 + BERT	60	60	64.14	82.83	88.98
ResNet50 + BERT	60	15	65.84	84.24	90.22

is able to enhance and fuse local features more efficiently. Then the utilization of the ICMP strategy can further bring 1.19% and 0.73% Rank-1 improvements on the two datasets. That is because our ICMP strategy reduces the image-text domain gap and simultaneously improves sample clustering within each modality, while the baseline only takes the former into consideration. In the image-to-text retrieval scenario, we can draw a similar conclusion that both RLEF and ICMP have significant effects on improving retrieval accuracy.

In addition, in order to verify the contribution of the two components of ICMP loss (*i.e.*, IMPM loss and CMPM loss) to the overall constraint, we conducted corresponding ablation experiment on some encoder combinations. As shown in the Table 7, not using IMPM loss (0 IMPM epoch) or using IMPM loss throughout training results in a decline in performance, while applying IMPM in the early epochs shows the best balance of intra-modal and cross-modal alignment. For example, using IMPM for 12 epochs brings 10.74% and 23.64% improvements compared to not using IMPM and using IMPM throughout the training process, respectively. This indicates that IMPM loss has the ability to quickly converge intra-class distances, and using it in the early epochs of training makes a great contribution to model effect.

4.4.2. Influence of each hyper-parameter

The hyper-parameter Q in the Relation-based Feature Filter has a significant impact on retrieval performance, which determines the degree of high-relevant information filtering. As reported in Table 8, we compare the performances of different Q settings on the CUHK-PEDES dataset. We can observe that our method achieves the best accuracy when Q is set to 48. That is mainly because the extracted features may fuse some weak-relevant noises when Q is too large, and the discriminative information can be mistakenly discarded when Q is too small.

Besides, we also analyze the influence of hyper-parameter λ on the performance, which determines the training weight of the IMPM loss in the ICMP strategy. From Table 9, we can find that our method reaches the peak performance when $\lambda = 0.1$, which means the training weight of the IMPM loss is much smaller than the CMPM loss. The main reason is that the intra-modal gap is much smaller than the cross-modal gap, consequently, the convergence speed of the IMPM loss is much faster than the CMPM loss.

Table 8
Performance comparison of different hyper-parameter Q on CUHK-PEDES dataset.

Q	Rank-1	Rank-5	Rank-10
16	65.21	84.24	89.46
32	65.77	83.74	89.73
48	**65.84**	**84.24**	**90.22**
64	65.25	83.90	89.65

Table 9
Performance comparison of different hyper-parameter λ on CUHK-PEDES dataset.

λ	Rank-1	Rank-5	Rank-10
0	64.65	84.28	89.78
0.01	65.12	**84.31**	89.93
0.1	**65.84**	84.24	**90.22**
1	65.16	84.03	89.77
10	63.35	82.62	88.42

4.5. Visualization

4.5.1. Feature distribution

As shown in Fig. 7, we adopt t-SNE to visualize the feature distributions with different training strategies, including a single IMPM loss, single CMPM loss and joint use of both losses. Before training, there is a large domain gap between visual and textual modalities, and samples inside each modality are unordered. Through the training with the IMPM loss, we can find that each modality's samples achieve effective clustering, but the cross-modal domain gap still exists. Furthermore, through training with the CMPM loss, the feature distributions of the two modalities are highly overlapped, but a few samples still deviate from clustering. Finally, after the jointly training of these losses, our method achieves a better feature distribution, which simultaneously reduces the intra-modal and cross-modal gaps.

4.5.2. Retrieval results

Fig. 8 lists some retrieval examples using our method and the compared baseline. We can find that both methods perform well on retrieving, but our method shows more capability in capturing details. As shown in the 4th row of Fig. 8, our method successfully finds the correct person with 'panda' pattern and 'glasses', but the baseline only focuses the rough features such as 'shirt' and 'blue bag'.

5. Conclusion

In this work, we presents a novel Local-Enhanced Representation Framework (LERF) for text-based person search task, which takes local denoising and intra-cross modal matching into consideration to enhance the discriminative ability of representations. On the one hand, we design a Relation-based Local-Enhanced Fusion (RLEF) Module to restrain local noises during the feature fusion, so that the high-relevant information can be fully utilized. On the other hand, we introduce an Intra-Cross Modal Projection (ICMP) Strategy to jointly conduct intra-modal and cross-modal matching during the distance measurement. Finally, through extensive experiments on three mainstream datasets, our method demonstrates superior performance, bringing a new solution to text-based person search tasks from local enhancement perspective.

However, our proposed method relies heavily on the accuracy and quality of text descriptions. Errors or ambiguities in text input may lead to incorrect retrieval results. In addition, the method tends to utilize the first Q word-level features, and ignores the utilization of filtered parts, which may cause the loss of discriminative information that is not closely related to other words. In the future, we aim to

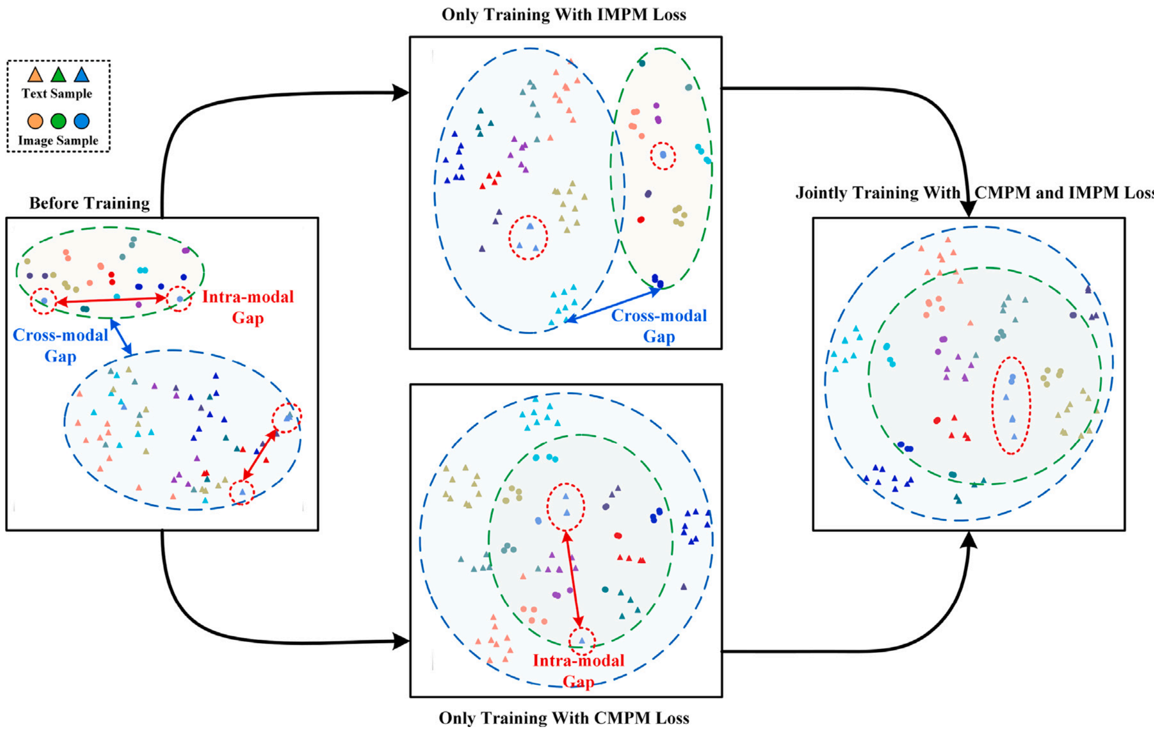

Fig. 7. Visualization of feature distributions with different training strategies by t-SNE. The features of images and texts are marked into the circles and rectangles, respectively. Different colors represent the identities of these samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

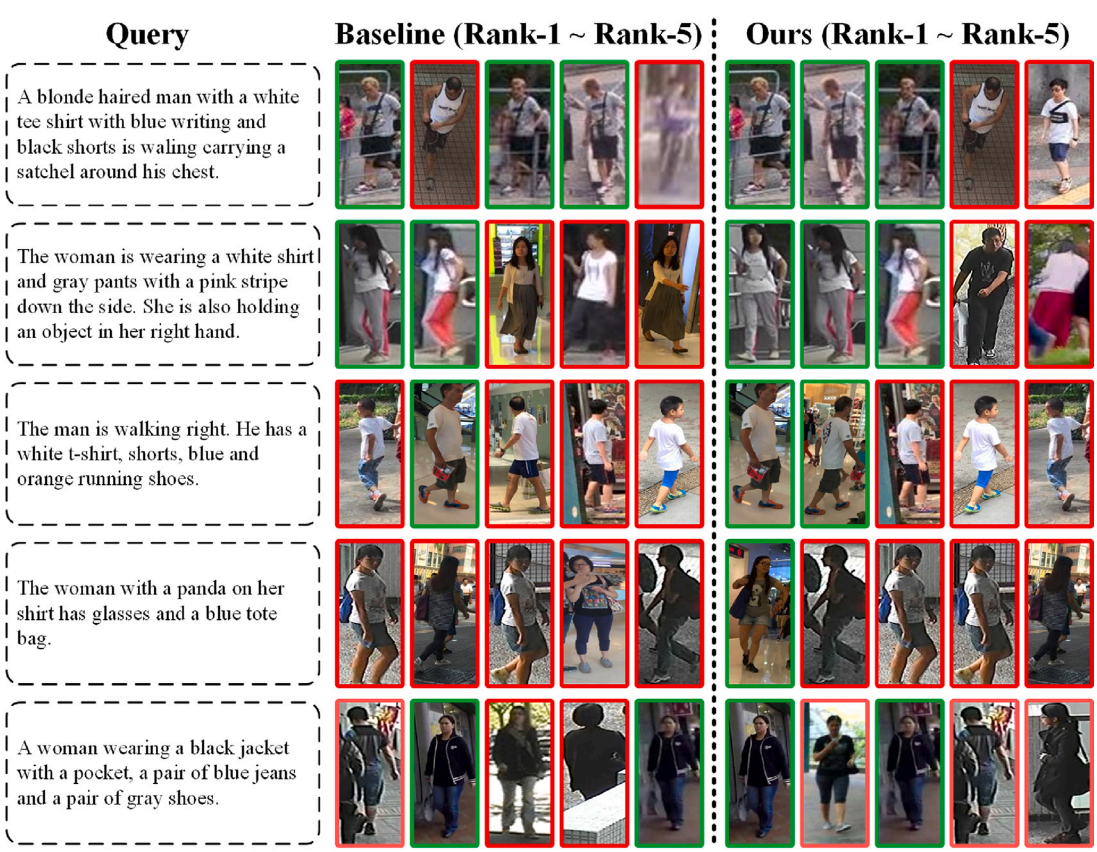

Fig. 8. Visualization of retrieval results on CUHK-PEDES dataset. The green and red bounding boxes indicate correct and incorrect matches, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

handle potential erroneous image-text pairs in the dataset and adjust the proportion of filtered and unfiltered information in text features for joint optimization.

CRediT authorship contribution statement

Guoqing Zhang: Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Conceptualization. **Yuhao Chen:** Software, Investigation. **Yuhui Zheng:** Writing – review & editing, Supervision, Funding acquisition. **Gaven Martin:** Writing – review & editing, Supervision. **Ruili Wang:** Writing – review & editing, Supervision.

Declaration of competing interest

We authors confirm that

The work described is not under consideration for publication elsewhere;

All the necessary files have been uploaded by online;

Each author has participated sufficiently;

All the authors listed have approved the manuscript that is enclosed.

Acknowledgment

This research is supported by National Natural Science Foundation of China under Grant 62172231, 92470202 and U20B2065; and by Natural Science Foundation of Jiangsu Province under Grant BK20220107; and by Preliminary Research Project on Leading Technologies by Wuxi Industrial Innovation Research Institute-Visual Intelligent Analysis of Worker Behavior and Anomaly Warning; and by 2020 Catalyst: Strategic New Zealand - Singapore Data Science Research Programme Fund by MBIE, New Zealand.

Data availability

Data will be made available on request.

References

[1] H. Miao, J. Lin, J. Cao, X. He, Z. Su, R. Liu, SMPR: Single-stage multi-person pose regression, Pattern Recognit. 143 (2023) 109743.
[2] M. Ling, T. Pan, Y. Ren, K. Wang, X. Geng, Motional foreground attention-based video crowd counting, Pattern Recognit. 144 (2023) 109891.
[3] G. Zhang, J. Liu, Y. Chen, Y. Zheng, H. Zhang, Multi-biometric unified network for cloth-changing person re-identification, IEEE Trans. Image Process. 32 (2023) 4555–4566.
[4] G. Zhang, H. Zhang, W. Lin, A.K. Chandran, X. Jing, Camera contrast learning for unsupervised person re-identification, IEEE Trans. Circuits Syst. Video Technol. 33 (8) (2023) 4096–4107.
[5] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, F. Wu, Diverse part discovery: Occluded person re-identification with part-aware transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2898–2907.
[6] G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, S. Chen, Deep high-resolution representation learning for cross-resolution person re-identification, IEEE Trans. Image Process. 30 (2021) 8913–8925.
[7] G. Zhang, Z. Luo, Y. Chen, Y. Zheng, W. Lin, Illumination unification for person re-identification, IEEE Trans. Circuits Syst. Video Technol. 32 (10) (2022) 6766–6777.
[8] V.D. Nguyen, P. Mantini, S.K. Shah, Contrastive clothing and pose generation for cloth-changing person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7541–7549.
[9] Y. Chen, G. Zhang, Y. Lu, Z. Wang, Y. Zheng, Tipcb: A simple but effective part-based convolutional baseline for text-based person search, Neurocomputing 494 (2022) 171–181.
[10] S. Aggarwal, V.B. Radhakrishnan, A. Chakraborty, Text-based person search via attribute-aided matching, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2617–2625.
[11] Z. Wang, Z. Fang, J. Wang, Y. Yang, Vitaa: Visual-textual attributes alignment in person search by natural language, in: Proceedings of the European Conference on Computer Vision, ECCV, 2020, pp. 402–420.
[12] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Pose-guided multi-granularity attention network for text-based person search, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 11189–11196.
[13] J. Zhou, B. Huang, W. Fan, Z. Cheng, Z. Zhao, W. Zhang, Text-based person search via local-relational-global fine grained alignment, Knowl.-Based Syst. 262 (2023) 110253.
[14] A. Farooq, M. Awais, J. Kittler, S.S. Khalid, Axm-net: Implicit cross-modal feature alignment for person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 4477–4485.
[15] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 686–701.
[16] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1970–1979.
[17] Z. Ding, C. Ding, Z. Shao, D. Tao, Semantically self-aligned network for text-to-image part-aware person re-identification, 2021, arXiv preprint arXiv:2107.12666.
[18] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, G. Hua, Dssl: Deep surroundings-person separation learning for text-based person retrieval, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 209–217.
[19] J. Gu, J. Cai, S.R. Joty, L. Niu, G. Wang, Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7181–7189.
[20] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Pecognition, 2020, pp. 10921–10930.
[21] S. Yang, Q. Li, W. Li, S. Li, A. Liu, Dual-level representation enhancement on characteristic and context for image-text retrieval, IEEE Trans. Circuits Syst. Video Technol. 32 (2023) 8037–8050.
[22] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, Y. Li, Caibc: Capturing all-round information beyond color for text-based person retrieval, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5314–5322.
[23] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, C. Wang, See finer, see more: Implicit modality alignment for text-based person retrieval, in: European Conference on Computer Vision, 2022, pp. 624–641.
[24] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving description-based person re-identification by multi-granularity image-text alignments, IEEE Trans. Image Process. (2020) 5542–5556.
[25] N. Sarafianos, X. Xu, I.A. Kakadiaris, Adversarial representation learning for text-to-image matching, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5814–5824.
[26] F. Shen, X. Shu, X. Du, J. Tang, Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval, IEEE Trans. Multimed. (2023) 8922–8931.
[27] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, C. Ding, Learning granularity-unified representations for text-to-image person re-identification, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5566–5574.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR, 2015.
[30] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, in: NIPS 2014 Workshop on Deep Learning, 2014.
[31] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[33] C. Wang, Z. Luo, Y. Lin, S. Li, Text-based person search via multi-granularity embedding learning, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 1068–1074.
[34] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, Y. Li, Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1984–1992.
[35] L. Gao, K. Niu, B. Jiao, P. Wang, Y. Zhang, Addressing information inequality for text-based person search via pedestrian-centric visual denoising and bias-aware alignments, IEEE Trans. Circuits Syst. Video Technol. (2023).
[36] K. Niu, L. Huang, Y. Long, Y. Huang, L. Wang, Y. Zhang, Comprehensive attribute prediction learning for person search by language, IEEE Trans. Image Process. 33 (2024) 1990–2003.
[37] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Trans. Multimed. Comput. Commun Appl. (TOMM) 16 (2) (2020) 1–23.

[38] S. Yan, H. Tang, L. Zhang, J. Tang, Image-specific information suppression and implicit local alignment for text-based person search, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–14.

[39] K. Niu, T. Huang, L. HUang, L. Wang, Z. Yanning, Improving inconspicuous attributes modeling for person search by language, IEEE Trans. Image Process. 32 (2023) 3429–3441.

[40] Z. Wang, J. Xue, A. Zhu, Y. Li, M. Zhang, C. Zhong, Amen: Adversarial multi-space embedding network for text-based person re-identification, in: Pattern Recognition and Computer Vision: 4th Chinese Conference, 2021, pp. 462–473.

Guoqing Zhang received the B.S. and Master degrees in Information Engineering from the Yangzhou University, Yangzhou, China, in 2009 and 2012, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology, Nanjing, China, in 2017. He is currently a Professor with the School of Computer science, Nanjing University of Information Science and Technology (NUIST), Nanjing, China. He is also currently pursuing the Ph.D. degree with the Massey University, Auckland, New Zealand. His current research interests include computer vision, pattern recognition and machine learning.

Yuhao Chen received his B.S. degree from the School of Computer and Science, Nanjing University of Information Science and Technology, Nanjing, China in 2020. He is currently a master degree candidate in the School of Computer and Science, Nanjing University of Information Science and Technology. His research interests include computer vision and pattern recognition.

Yuhui Zheng was born in Shanxi, China, in 1982. He received the B.Sc. degree in pharmacy engineering and the Ph.D. degree in pattern recognition and intelligent system from Nanjing University of Science and Technology, Nanjing, China, in 2004 and 2009, respectively. From 2014 to 2015, he was a visiting professor in the digital media laboratory of the school of Electronic and Electrical Engineering, Sungkyunkwan University, Korea. He is currently a Full Professor at the School of Computer and Software in Nanjing University of Information Science and Technology (NUIST). His main research areas include image and video analysis, scene understanding, visual tracking, and pattern recognition.

Gaven Martin received the Ph.D. degree from Michigan University, Ann Arbor, Michigan, USA, in 1985. He is a New Zealand mathematician. He is Distinguished Professor of Mathematics at Massey University, Auckland, New Zealand, Director of the New Zealand Institute for Advanced Study, former President of the New Zealand Mathematical Society, and former Editor-in-Chief of the New Zealand Journal of Mathematics. He is a former Vice-President of the Royal Society of New Zealand (Mathematics, Physical Sciences, Engineering and Technology). His research involves quasi-conformal mappings, regularity theory for partial differential equations, and the connection between discrete group theory and low-dimensional topology.

Ruili Wang, Fellow of Engineering of New Zealand, received the Ph.D. degree in computer science from Dublin City University, Dublin, Ireland. He is currently a Professor of Artificial Intelligence and Chair of Research in the School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand, where he is the Director of the Centre of Language and Speech Processing. His current research interests include speech processing, language processing, video processing, data mining, and intelligent systems. Dr. Wang serves as a member and an Associate Editor of the editorial boards for international journals, including the journals of IEEE Transactions on Emerging Topics in Computational Intelligence, Knowledge and Information Systems, Neurocomputing, and Applied Soft Computing.