RMGNet: The Progressive Relationship-Mining Graph Neural Network for Text-to-image Person Re-identification

Xin Zhang, Kun Liu, Xinwang Wang, Zhong Zhou, Haiyong Chen

***Abstract*—The Text-to-Image Person Re-identification (TI-ReID) task objective is to precisely identify the person's images with the textual description of the person. The mainstream research methods focus on cross-modal aligning local features, and overlook the learning of intra-modal and cross-modal relationships between different features. This renders the person features lacking in high-level semantic information. To resolve such issues, we propose the Progressive Relationship-Mining Graph Network (RMGNet), including the Intra-Modal Relationship-Mining (IMRM) and the Cross-Modal Relationship-Mining (CMRM) module. These modules are employed to model and mine semantic relationship information among different features. Specifically, the IMRM module models and mines the high-level semantic interrelationships inherent in the image and text features. The CMRM module introduces the nearest neighbor method to model cross-modal semantic relationships to enhance the cross-modal semantic correspondence capabilities of person features. On this basis, we design the Adaptive Corner Center (Acc) loss and the Coarse-to-Fine Learning (C2FL) strategy. These ensure the network receives consistent and effective metric learning supervision throughout the entirety of the training process. To validate the efficacy of the proposed method, extensive experiments are conducted on three prevalent datasets: CHUK-PEDES, ICFC-PEDES, and RSTPReid. The achieved mAP of 70.59%, 41.62%, and 49.58% surpassed those current state-of-the-art methods.**

***Index Terms*—Person Re-identification, Multi-Modal, Text-to-Image Retrieval, Relationship-Mining Graph, Graph Neural Network.**

I. Introduction

FOR the research of Intelligent Transportation Systems, Text-to-Image Person Re-identification [1, 2] has broad application prospects. It can achieve the recognition of different person utilizing various devices and algorithms without images of the target person. This facilitates the in-depth integration and application of computer vision technology in object tracking [3–7], action recognition [8, 9] and autonomous driving [10–12]. Due to the significant modal gap, the primary challenge lies in the efficient extraction of cross-modal discriminative features in both person images and text. This necessitates the exploration of their hidden semantic correspondence relationship.

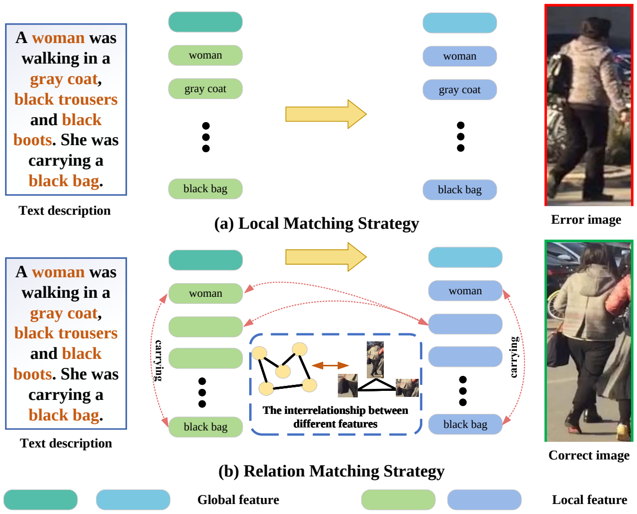

A woman was walking in a gray coat, black trousers and black boots. She was carrying a black bag.
Text description
woman
gray coat
black bag
woman
gray coat
black bag
(a) Local Matching Strategy
Error image
A woman was walking in a gray coat, black trousers and black boots. She was carrying a black bag.
Text description
woman
carrying
black bag
The interrelationship between different features
woman
carrying
black bag
(b) Relation Matching Strategy
Correct image
Global feature
Local feature

Fig. 1. Comparison of different matching strategies. (a) The prevalent local-based matching strategy enhances feature expressiveness by learning and aligning discriminative local features in images and texts. (b) Our relationship-based matching strategy focuses on modeling and mining relationships between different features to further enhance discriminative capabilities and distinguish persons with similar appearances.

In recent years, research on TI-ReID has advanced, with existing methods primarily following two strategies: global matching [13–15] and local matching [16–18]. The global matching methods map person image and text features into a unified space to reduce modal disparities interference, same as the image-text retrieval methods [19, 20]. However, relying solely on global features for matching the person text and image may lead to the omission of crucial discriminative local features. This makes it difficult to accurately identify different person with similar appearances. While local matching methods focus on mining salient regions within images and discriminative words in the text, enabling fine-grained matching between person images and textual descriptions, as illustrated in Figure 1(a). Despite the similarity in appearance between the two people depicted in the image, accurate discrimination is still achievable through distinct local features, such as differences in shoe color and trouser type.

While the aforementioned methods have achieved some progress, the actual recognition results remain suboptimal.

Xin Zhang, Kun Liu, and Haiyong Chen are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, 300130, China.(e-mail: zhangxin; KunLiu;HaiyongChen@hebut.edu.cn); XinWang Wang is with the School of Instrument Science and Engineering, Southeast University, Nanjing 210018, Jiangsu, China and the School of Integrated Circuit, Wuxi Institute of Technology, Wuxi, 214121, China (e-mail: 230189684@seu.edu.cn); Xin Zhang and Zhong Zhou are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: zhangxin8275; zz@buaa.edu.cn)

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3532685

These methods focus on mining highly discriminative local features and achieve cross-modal alignment. The learning of intrinsic semantic connections and interactions between features is overlooked. Specifically, the variations in feature relationships will lead to changes in semantic meaning, consequently altering the correspondence between images and text. These methods ignore fully mining the complex and varied semantic relationships between different features. This defect results in errors in recognition outcomes, as depicted in Figure 1(b). The two pedestrian images and their corresponding description texts are basically the same, with nearly identical local features. However, the relationship between the backpack and the pedestrian is different in the image and the textual description. The textual description of the image on the left reads ”person carrying a bag”, while the right image is ”person wearing a bag”. It can be seen that through the modeling and mining of feature relationships, the discriminative capability of features can be further enhanced.

To effectively model and mine the relationships between different features, we propose the Progressive Relationship Mining Graph Neural Network (RMGNet). It enhances feature expression capabilities by learning the interrelationship between different features within intra-modal and inter-modal. In the RMGNet, the Intra-Modal Relationship-Mining (IMRM) module encodes high-level semantic concepts associated with local features of images and texts. This optimizes features by aggregating semantic contextual information and latent interrelationships between features. The Cross-Modal Relationship-Mining (CMRM) module models the cross-modal feature semantic interrelationship by fusing the GNN and the nearest neighbor strategy. Utilizing the powerful relational reasoning capabilities of GNN to learn the semantic relationships between different modal features and aggregation enhancement. In addition, we convert the image-text cross-modal matching task into the binary classification task. The classification probability output by the network serves as auxiliary discriminant information to improve the accuracy of TI-ReID.

Furthermore, the hetero-center triplet (Hc-Tri) loss [21], tends to easily satisfy the relative distance constraint early in the training process and converges quickly. Consequently, there is a lack of supervision in the later stages of network training. Therefore, we propose the coarse-to-fine learning (C2FL) strategy and the novel adaptive corner center (Acc) loss to train the network.

The proposed method improves the recognition accuracy with the following four contributions:

- We propose the Progressive Relationship Mining Graph Neural Network (RMGNet), which is used to model and mine the hidden inter-relationship between features within the intra-modal and inter-modal. The network is the first to apply the GNN to learn the mutual semantic relationships between features in the TI-ReID task.
- We design the Intra-Modal Relationship-Mining (IMRM) module, which is used to model and mine the hidden fine-grained semantic relationships between different features within the intra-modal.
- We design the Cross-Modal Relationship-Mining (CMRM) module, which is employed to model and learn the semantic correlation and affinity relationship between person features within the inter-modal by introducing the nearest neighbor strategy.
- The new Coarse-to-Fine Learning (C2FL) strategy and Adaptive Corner Center (Acc) loss are proposed to enable the network to receive effective metric learning supervision throughout the training process.

II. Related Works

As computer vision technology has evolved, the person ReID task has achieved great advancements in both academic research and practical applications [22–25]. However, the image-based person ReID necessitates at least one image of the target pedestrian in the application. Therefore, the practical application of image-based person ReID is limited. To address this limitation and enhance the practicality of person ReID, Li et al.[26] proposed the text-to-image person re-identification task. Researchers have proposed a variety of re-identification frameworks, which can be primarily divided into global matching methods [27–29] and local matching methods [30–32].

A. Global matching method

The global matching method was the main research approach in the early TI-ReID task. It focuses on learning the correspondence between person images and text descriptions holistically, calculating similarity based on global features [13, 14, 28, 29, 33]. In [26], Li et al. proposed the recurrent neural network with a gated neural attention mechanism network (GNA-RNN) and used the Visual Geometry Group (VGG) network to extract text and image features respectively for similarity measurement. In addition, they proposed the CUHK-PEDES dataset. In[27], Zhang et al. posited that accurately measuring feature similarity across different modalities is crucial for matching images and texts. To tackle this, they proposed the cross-modal projection matching (CMPM) loss and the cross-modal projection classification (CMPC) loss to effectively enhance the compactness between each person features. Li et al.[34] proposed the visual semantic reasoning network (VSRN), which uses the GCN to capture the semantic relationship of salient regions in the image. It then utilizes the gating and memory mechanism for global reasoning, enhancing the performance of image-text matching tasks. With the development of large pre-trained models, some research attempts to use visual, language, and other pre-trained models to enhance the expression ability of the person features and distinguish discriminate between different persons[13, 35]. Ye et al [35]. leveraged the cross-modal image-text alignment capability of the Contrastive Language-Image Pre-training (CLIP) model solely for enhancing performance using global features. While such methods have achieved certain results, the absence of constraints on local features limits their effectiveness in real-world application scenarios.

B. Local matching methods

Motivated by traditional person ReID methods based on local matching [36–38], researchers have proposed local

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3532685

matching-based TI-ReID methods. This type of method primarily focuses on learning local discriminative information in person images and texts, which key is to performing cross-modal alignment [16–18, 39]. Jing et al. [40] performed cross-modal recognition by introducing pose estimation information to align local features. In [41], Aggarwal et al. designed the cross-modal attribute-aided matching framework (CMAAM) which approach introduces and preserves high-level semantic information in pedestrian features through an additional attribute prediction model. This helps alleviate the modal gap interference and improves the effect of feature learning. Ding et al.[39] raised the semantically self-aligned network (SSAN) to mine semantically aligned local features from person images and texts. They also establish the correspondence between person parts and word phrases through a multi-view non-local network. This effectively alleviates significant modal gaps and intra-class differences. Furthermore, they proposed the widely used dataset ICFG-PEDES. Yang et al.[42] redesigned the cross-attention module to limit the gap between different modality features and introduced direct constraints in local feature matching progress. In [43], Han et al. adopted the graph convolutional network (GCN) to extract and fuse multi-modal features. And proposed an asymmetric multi-level alignment module to extract ”local” information more accurately from a ”global” perspective. The local matching method further improves the effectiveness of TI-ReID. However, there is an asymmetry in the amount of information contained in images and texts, that is, the semantic information in images is relatively redundant, while the semantic information in texts is relatively lacking. Therefore, forced alignment of local features will disrupt the feature-extracting process. Consequently, it is difficult to further improve the feature expression ability and distinguishability.

Hence, we start by learning the mutual relationships between different features and propose the RMGNet. The network employs the IMRM module and the CMRM module to progressively model and mine potential semantic relationships between different features. Following this, the interrelationship between features is utilized as a guide to aggregate and strengthen contextual information, to augment the expressiveness and discriminative of features.

III. Method

In this section, we provide a detailed introduction to the proposed TI-ReID method. Firstly, we introduce the proposed RMGNet, which includes the IMRM and the CMRM module. Next, we propose the C2FL training strategy and the Acc loss. Finally, we explain how to calculate the similarity between the person images and texts.

A. Overview of Framework

The overall architecture of the proposed RMGNet is illustrated in Figure 2(a). The network mainly consists of three parts: the single-modal feature extraction module, the IMRM module, and the CMRM module. In the single-modal feature extraction module, we utilize the pre-trained Vision Transformer (ViT) [44] and Bidirectional Encoder Representations from Transformers (BERT) network [45] to extract image and text features of the person, respectively. It should be noted that, unlike other TI-ReID methods, when extracting person text features, we extract forward and backward text features through forward-order and reverse-order input, respectively. Subsequently, the extracted local features are fed into the IMRM module to learn the relationships between different local features.

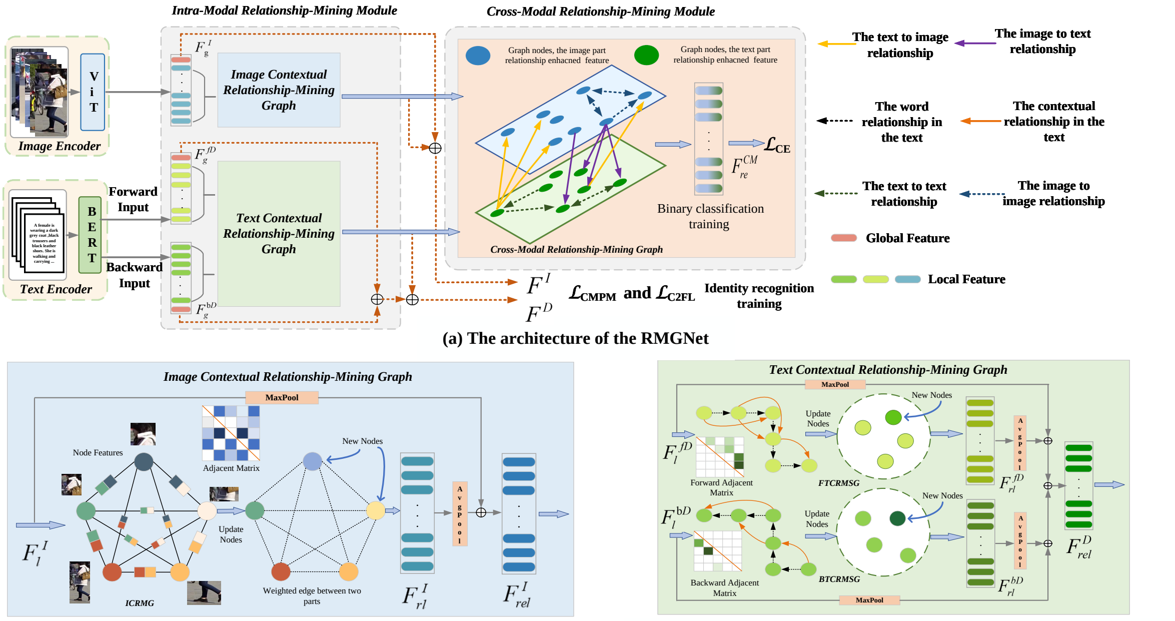
Specifically, the IMRM module comprises the image contextual relationship-mining graph (ICRMG) and text contextual relationship-mining graph (TCRMG), which encode the mutual semantic relationships between image and text local features within the intra-modal. The model optimizes local features by aggregating their semantic contextual information and interrelationships. Then, on the one hand, we fuse the relationship-enhanced local features with the global features as the final image and text feature expression to calculate similarity. On the other hand, the CMRM module adopts the relationship-enhanced local features to mine the interrelationships between different modal local features. The CMRM module employs the nearest neighbor method to model the semantic relationships between local features within different modalities. Consequently, the cross-modal discriminative information has been learned, enhancing the cross-modal semantic correspondence and expression ability of person features. Finally, through binary classification training, it is directly determined whether the image and text are the same person.

B. Intra-Modal Relationship-Mining Module

1) Image contextual relationship-mining graph: In the person ReID task, the distinguishable local features in person images have played an important role. But only relying on these local features is not enough. It is also very important to model and mine the semantic relationships between different features. As mentioned in the introduction, the left description of Figure 1(b) is ’A woman was walking in a gray coat. She was carrying a black bag.’, the right description of Figure 1(b) is ’A middle-aged woman was wearing a gray coat, walking and wearing a black bag.’. It can be observed that key information such as ’woman,’ ’gray coat,’ and ’black bag’ are detected in both person images. If we only use these local features for direct matching, it may lead to recognition errors. Clearly, the interrelationship between the bag and other local features, such as ’carrying the bag’ or ’wearing the bag,’ is crucial for accurately distinguishing different person. To this end, we design the Image Contextual Relationship-Mining Graph (ICRMG), which leverages the GNN to model and mine the potential interrelationship between these person local features, the detailed architecture of the ICRMG as shown in Figure 2(b). We employ the KNN graph as the graph structure that allows for more efficient information aggregation to better capture relationships between different local features and explore their semantic relevance and similarity. Meanwhile, this also allows preventing the introduction of excessive noise.

Specifically, firstly, the person image is divided into n patches, and ViT is employed to extract the local features F_l^I and global features F_g^I of the person, where $F_l^I =$

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3532685


Intra-Modal Relationship-Mining Module
Cross-Modal Relationship-Mining Module
Image Encoder
ViT
F_g^I
Image Contextual Relationship-Mining Graph
Forward Input
Backward Input
BERT
Text Encoder
F_g^{fD}
F_g^{bD}
Text Contextual Relationship-Mining Graph
Graph nodes, the image part relationship enhacned feature
Graph nodes, the text part relationship enhacned feature
$\mathcal{L}_{CE}$
F_{re}^{CM}
Binary classification training
Cross-Modal Relationship-Mining Graph
F^I
F^D
$\mathcal{L}_{CMPM}$ and $\mathcal{L}_{C2FL}$
Identity recognition training
The text to image relationship
The image to text relationship
The word relationship in the text
The contextual relationship in the text
The text to text relationship
The image to image relationship
Global Feature
Local Feature
Image Contextual Relationship-Mining Graph
MaxPool
Node Features
Adjacent Matrix
New Nodes
Update Nodes
ICRMG
Weighted edge between two parts
Avgpool
F_l^I
F_{rl}^I
F_{rel}^I
Text Contextual Relationship-Mining Graph
MaxPool
Update Nodes
New Nodes
F_l^{fD}
Forward Adjacent Matrix
FTCRMSG
F_{rl}^{fD}
F_l^{bD}
Backward Adjacent Matrix
BTCRMSG
F_{rl}^{bD}
F_{rel}^D

(a) The architecture of the RMGNet

(b) The architecture of the Image Contextual Relationship-Mining Graph

(c) The architecture of the Text Contextual Relationship-Mining Graph

Fig. 2. (a) shows the architecture of RMGNet, while (b) and (c) represent the architectures of ICRMG and TCRMG, respectively. Given N image-text pairs as input, we first use the ViT and BETR models to extract global and local features. The local features are then fed into the IMRM module to mine the mutual relationships among different local features by the ICRMG and the TCRMG. This can generate the relationship-enhanced image and text local features, F_{rel}^I and F_{rel}^D, respectively. In the third step, we fuse global features with relationship-enhanced local features to perform identity recognition training with the proposed C2FL strategy and CMPM loss. Concurrently, F_{rel}^I and F_{rel}^D are sent to the CMRM module to learn semantic interrelationships between different features across different modalities. Finally the binary classification training is conducted on the generated relationship-enhanced cross-modal feature F_{re}^{CM}.

$\{f_1^I, f_2^I ... f_n^I\}$, $F_l^I \in R^{n*512}$. After that, the local features F_l^I are transformed into the d-dimensional feature space, as:

$$H^I = w_1 F_l^I + b_1 \tag{1}$$

where w_1 and b_1 are learning parameters of network. As a result, the local features of the pedestrian image can be expressed as $H^I = \{h_1^I, h_2^I, h_3^I ... h_n^I\}$. In the second step, we use the local features of the person image to construct an undirected weighted relationship graph, that is, the image contextual relationship-mining graph denoted as $G^I = (H^I, E^I)$, where H^I is nodes within the graph, E^I represents edges i.e. the interrelationship between two connected nodes and regularized by weighted adjacent matrix. In this way, the semantic relationship between two nodes (two local features) can be modeled through the weight of the edges in the graph, and calculated as follows:

$$e_{i,j}^I = ReLu\left(\left(w_1^I \cdot h_i^I + b_1^I\right) \cdot \left(w_2^I \cdot h_j^I + b_2^I\right)\right) \tag{2}$$

where w_1^I and w_2^I are the parameters of the fully connected layer respectively, b_1^I and b_2^I are the parameters of the BN layer respectively, which are used to determine the interrelationship between two nodes h_i^I and h_j^I. Further, the weighted adjacency matrix $A_{(i,j)}^I$ of the ICRMG can be expressed as:

$$A_{(i,j)}^I = \begin{cases} e_{i,j}^I, & if\ i \neq j \\ 0, & else \end{cases} \tag{3}$$

After constructing the graph G^I. The weighted edge E^I guides the aggregation of hidden interrelationship information in other nodes that are semantically relevant to the local feature H^I. The entire process is as follows:

$$f_{rli}^I = ReLu\left(\sum_{j=1}^{n} A_{(i,j)}^I \times \left(w^I \cdot h_j^I + b^I\right) + h_i^I\right) \tag{4}$$

where f_{rli}^I represents the local feature that has fused relationship information between other local features. In this way, the relationship-enhanced person image local features F_{rl}^I can be generated. Finally, in order to prevent oscillation interference in the early stage of model training. We aggregate the initial local features and enhance local features through maximum pooling and average pooling:

$$F_{rel}^I = \alpha \,.\, MaxPool\left(F_l^I\right) + (1-\alpha) \,.\, AvgPool\left(F_{rl}^I\right) \tag{5}$$

where α is the adjustment parameter, F_{rel}^I represent the final relationship-enhanced image loacl features.

2) Text contextual relationship-mining graph: After obtaining the relationship-enhanced person image local features, in order to obtain the relationship-enhanced person text local features. We design the text contextual relationship-mining graph (TCRMG) as shown in Figure 2(c). Since text is a type of data with sequential attributes, its reading order has a significant impact on feature learning and matching. For example, the text description 'A woman was walking in a gray coat. She was carrying a black bag.' and 'A middle-aged woman was wearing a gray coat, walking and wearing a black bag.' corresponds to different person images. When extracting features forward, the first two features described by the two sentences are ”woman” and ”gray coat”, which are similar. This interferes with the matching of image-text, that the person images corresponding to the two sentences are considered to be the same. In contrast, when extracting features backward, the network will first extract two distinct features: ’carrying a black bag’ and ’wearing a black bag’. This can effectively distinguish the description text of different person images. Therefore, we model the interrelationship of person text local features from both forward and backward perspectives, respectively.

Different from the method of directly using the bidirectional long short-term memory (Bi-LSTM) network [46] for text feature learning. We build two sub-graphs in TCRMG, namely the forward text contextual relationship-mining sub-graph (FTCRMSG) and the backward text contextual relationship-mining sub-graph (BTCRMSG). The two sub-graphs model and learn the correlation of the text local features from the forward and reverse perspectives, respectively. Specifically, we first input the description text into the BERT model in forward sequence to learn the forward text local features F_l^{fD} and global features F_g^{fD}, where $F_l^{fD}=\left\{f_1^{fD},f_2^{fD}...f_n^{fD}\right\}$, $F\in R^{n*512}$. After that, the local features F_l^D are transformed to fed into FTCRMSG, as:

$$F_l^{fD}=w_2\cdot BERT\left(\overrightarrow{W^f}\right)+b_2 \tag{6}$$

where $\overrightarrow{W^f}$ represents the forward input word vector. Similarly, by inputting person description text in reverse, we can obtain backward text local features F_l^{bD}, as follows:

$$F_l^{bD}=w_3\cdot BERT\left(\overleftarrow{W^b}\right)+b_3 \tag{7}$$

Second, we employ GNN to model and mine the interrelationships between person text local features. We need to construct two directed weighted graphs, that is FTCRMSG and BTCRMSG, denoted as $G^{fD}=(H^{fD},E^{fD})$ and $G^{bD}=(H^{bD},E^{bD})$ respectively. Taking the forward local feature relationship mining as an example, H^{fD} represents the set of forward person text local features, that is, the nodes in the FTCRMSG. E^{fD} represents the relationship between two nodes in the FTCRMSG, which is determined as follows:

$$e_{i,j}^{fD}=ReLu\left(\left(w_1^{fD}\cdot h_i^{fD}+b_1^{fD}\right)\cdot\left(w_2^{fD}\cdot h_j^{fD}+b_2^{fD}\right)\right) \tag{8}$$

where h_i^{fD} and h_j^{fD} represent two different forward person text local features, $e_{i,j}^{fD}$ represents the mutual semantic relationship between them. On this basis, the weighted adjacency matrix of the FTCRMSG can be obtained as:

$$A_{(i,j)}^{fD}=\begin{cases}e_{i,j}^{fD}, & if\ i\neq j\\ \quad 0, & else\end{cases} \tag{9}$$

where $A_{(i,j)}^{fD}$ is the weight adjacency matrix. The edges and weights in the graph can represent the semantic correlation between two words. Thereby, we can model and extract the interrelationship between different text local features as follows:

$$f_{rli}^{fD}=ReLu\left(\sum_{j=1}^{n}A_{(i,j)}^{fD}\times\left(w^{fD}\cdot h_j^{fD}+b^{fD}\right)\right)+h_i^{fD} \tag{10}$$

where f_{rli}^{fD} represents the updated forward text local feature that has been enhanced with the relationship information. Therefore, the relationship guide updated backward text local features can be calculated as:

$$\begin{gathered}e_{i,j}^{bD}=ReLu\left(\left(w_1^{bD}h_i^{bD}+b_1^{bD}\right)\left(w_2^{bD}h_j^{bD}+b_2^{bD}\right)\right)\\ A_{(i,j)}^{bD}=\begin{cases}e_{i,j}^{bD}, & if\ i\neq j\\ \quad 0, & else\end{cases}\\ f_{rli}^{bD}=ReLu\left(\sum_{j=1}^{n}A_{(i,j)}^{bD}\times\left(w^{bD}h_j^{bD}+b^{bD}\right)\right)+h_i^{bD}\end{gathered} \tag{11}$$

Similar to the ICRMG, the final forward and backward relationship-enhanced person text local features, F_{rel}^{fD} and F_{rel}^{bD}, are obtained by fusing the original and the relationship-enhanced text local features. Finally, we weighted and fuse the forward and backward relationship-enhanced text local features to generate the final text relationship enhancement local features F_{rel}^D as:

$$\begin{gathered}F_{rel}^{fD}=\alpha\,.\,MaxPool\left(F_l^{fD}\right)+(1-\alpha)\,.\,AvgPool\left(F_{rl}^{fD}\right)\\ F_{rel}^{bD}=\alpha\,.\,MaxPool\left(F_l^{bD}\right)+(1-\alpha)\,.\,AvgPool\left(F_{rl}^{bD}\right)\\ F_{rel}^{D}=\left\|\frac{F_{rel}^{fD}+F_{rel}^{bD}}{2}\right\|_2\end{gathered} \tag{12}$$

C. Cross-Modal relationship-mining Module

In order to further model and mine the correspondence between these features within different modals, we have designed the Cross-Modal Relationship-Mining module (CMRM). First, we construct the Cross-Modal Relationship-Mining Graph (CMRMG) based on the obtained image and text relationship-enhanced features, defined as $G=(H,E)$, where $H=\left\{F_{rel}^I,F_{rel}^D\right\}=\left\{f_{rel1}^I,f_{rel2}^I\cdots f_{reln}^I,f_{rel1}^D,f_{rel2}^D\cdots f_{reln}^D\right\}$ represent the nodes on the CMRMG. E represents the edge in the graph, which is used to model and describe the high-level semantic relationships between different person local features within different modals. To accurately identify effective relationships and filter out irrelevant interfering relationships. We determine the weighted adjacency matrix by computing the nearest neighbor space of features. Specifically, the cosine distance between different features is calculated and ranked. Afterward, it is assumed that there is a semantic relationship between the two features only if f_i and f_j are among each other’s top K neighbors. This is because when the two different local features are close to each other, their semantic meanings can be considered to correspond and have interrelationships with each other. It should be noted that in the process of determining semantic associations, the intra-modal and inter-modal neighbor spaces are calculated separately and then fused. The weighted adjacency matrix of the CMRMG can be calculated as:

$$A_{cm}=\begin{cases}e_{i,j}, if\ f_i\in N_{intra}\left(f_j\right)\ and\ f_j\in N_{intra}\left(f_i\right)\ , i\neq j\\ \qquad\qquad 0, \qquad\qquad\qquad else\end{cases} \tag{13}$$

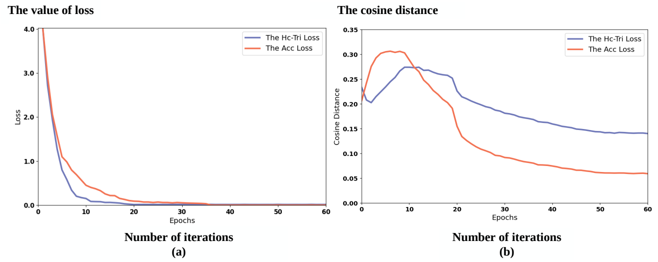


Fig. 3. (a) The variation curve of the Hc-Tri loss and the Acc loss. (b) The variation curve of the cosine distance between positive samples under the Acc loss and the Hc-Tri loss guiding.

where A_{cm} is the weighted adjacency matrix, $N_{intra}(\cdot)$ is the nearest neighbor space. The GNN is used to model the semantic relationships between different local features in the two modalities and performs learning updates as:

$$f_{reli} = ReLu\left(\sum_{j=1}^{N} A_{cm} \times (w \cdot h_j + b) + h_i\right) \tag{14}$$

where f_{reli} is the cross-modal relationship-enhanced person local feature. Further, we fuse the local features after cross-modal relationship enhancement to obtain the final cross-modal relationship-enhanced feature F_{re}^{CM}. Finally, F_{re}^{CM} is used for binary classification training to determine directly whether the person image matches the description text.

D. Training

Network training plays a crucial role in ReID research [47, 48]. During the training process of the method, first, we use the CMPM loss [27], the Hc-Tri loss [21], and the proposed Adaptive Corner Center (Acc) loss to guide the network to learn to extract pedestrian features. Specifically, we use the CMPM loss to optimize the learning of person image and text features to alleviate the interference of modal gaps. For a small batch of training data, its features and labels can be expressed as $\left\{(F_i^I, F_j^D), y_{i,j}\right\}_{i,j=1}^{K}$, where F_i^I is the ith person image feature, F_j^D is the jth person text feature, that is generated by fusing the global features and relationship-enhanced local features. When $y_{i,j} = 1$, it means that the two features are correctly matched and belong to the same person. The probability $p_{i,j}$ that F_i^I and F_j^D match can be defined as:

$$p_{i,j} = \frac{exp\left(F_i^{ID}\overline{F_j^D}\right)}{\sum_{t=1}^{K} exp\left(F_i^{ID}\overline{F_t^D}\right)}, \quad \overline{F_j^D} = \frac{F_j^D}{\left\|F_j^D\right\|} \tag{15}$$

where $\overline{F_j^D}$ represents the normalized text features, and $F_i^{ID}\overline{F_j^D}$ represents the projection of the image features in the text feature space. There are multiple correctly matched images and texts in a batch of training samples. Therefore, the normalized matching probability is calculated as follows:

$$q_{i,j} = \frac{exp\left(y_{i,j}\right)}{\sum_{t=1}^{K} exp\left(y_{i,t}\right)} \tag{16}$$

where $q_{i,j}$ is the normalized final matching probability. Moreover, we normalize through the softmax function, which can also play a role in label smoothing. By calculating the KL divergence of the image-to-text projection probability and the true matching probability, the final image-to-text matching loss function can be obtained, specifically:

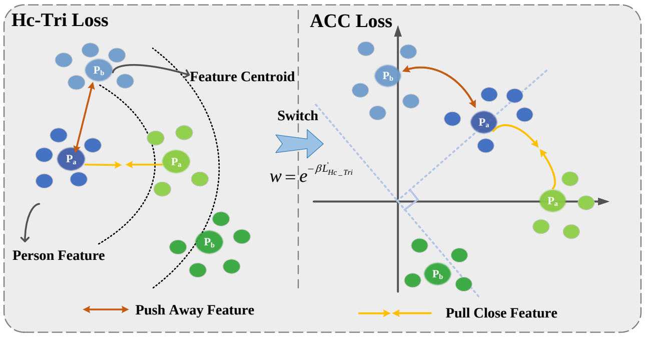


Fig. 4. The illustration of the proposed Coarse-to-Fine Learning strategy.

$$L_{I2T} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j} + \varepsilon}\right) \tag{17}$$

where ε is the parameter to prevent numerical overflow. After that, the image and text features in eq:15 also need to be exchanged to calculate the text-to-image matching loss. Therefore, the final CMPM loss function is expressed as:

$$L_{CMPM} = L_{I2D} + L_{D2I} \tag{18}$$

In addition, to reduce the intra-class distance, increase the inter-class distance. We also introduce the Hc-Tri loss based on the CMPM loss, calculated as follows:

$$L_{Hc\_Tri} = \sum_{i=1}^{K}\left[\rho + D\left(F_i^M, F_i^{\overline{M}}\right) - \min_{j \neq i} D\left(F_i^M, C_j^N\right)\right]_+ \tag{19}$$

where M and $\overline{M}$ respresent the modal and $M \neq \overline{M}$. F_i^M, $F_i^{\overline{M}}$, and C_j^N represent anchor sample features, positive sample features, and negative sample feature centre respectively. Although the loss can effectively increase the distance between different person features. However, during training, we find that the Hc-Tri loss converges quickly in the early stages, as shown in Figure 3 (a). It can be seen that the Hc-Tri loss converges to 0 in the 20th iteration. This shows that most central triples can easily satisfy the boundary constraints. As a result, effective supervision signals cannot be generated in the later stage of training. To overcome the limitations of Hc-Tri loss, we propose the more stringent Acc loss, which is calculated as follows:

$$L_{Acc} = \frac{\mu}{K}\sum_{i=1}^{K}\left(\left(1 - \cos\left(F_i^M, F_i^{\bar{M}}\right)\right) + \frac{1}{K}\sum_{j \neq i}\left[\cos\left(F_i^M, C_j^N\right)\right]_+\right) \tag{20}$$

where $cos(\cdot)$ is the cosine similarity, μ is the adaptive weight. We utilize the cosine space which with the smaller value range to calculate sample distances, which can better utilize the training idea of triplet loss. It ensures that the objective function remains continuous during the optimization process, enhancing the overall training effectiveness of the network, as shown in Figure3 (a). However, as can be seen from Figure3 (b), the cosine distance of Acc loss decreases significantly faster than the Hc-Tri loss after the warm-up stage. This is because the network is guided by strict metric learning loss when the initial learning ability is weak. This will cause the network to be biased towards early learning samples and affect the generalization ability. Therefore, we design the Coarse-to-Fine Learning strategy (C2FL) to achieve better network training effect. The strategy uses the relatively loose Hc-Tri loss for coarse guidance in the early stages of training. When the Hc-Tri loss converges, the strategy automatically switches to the strictly Acc loss to further reduce modal differences, as shown in Figure 4. We use the value of the Hc-Tri loss as an indicator in the C2FL to control the switching, as follows:

$$L_{C2FL} = \lambda_1 \cdot L_{Hc-Tri} + \lambda_2 \cdot wL_{Acc}, w = e^{-\beta L^{'}_{Hc-Tri}} \tag{21}$$

where $L^{'}_{Hc-Tri}$ is the cumulative average of L_{Hc-Tri} in each epoch, β is the adjustment parameter, λ_1 and λ_2 are balance parameters. Finally, because we directly use cross-modal relationship-enhanced features to process binary classification tasks, that is, directly determine whether two samples belong to the same person. Therefore, we use cross-entropy loss for binary classification training, calculated as follows:

$$L_{CE} = -\left(y \centerdot \log \hat{y}\right) + (1-y) \centerdot \log\left(1-\hat{y}\right) \tag{22}$$

where y is the ground truth label, $\hat{y}$ is the prediction results. The total loss function of the method can be expressed as follows:

$$L = \omega \cdot L_{CE} + L_{CMPM} + L_{C2FL} \tag{23}$$

E. Similarity calculation

In order to accurately match the person image and text, we calculate their similarity from two perspectives. First, we fuse the global features of the person image extracted by ViT and the enhanced person image local features to generate the final person image feature F^I. In the same way, the final person text features F^D can be obtained. After that, we calculate the cosine distance D between the two features as part of the similarity score. Second, we also introduce the prediction probability generated in the binary classification task P_{same}, the total similarity score S_{sim} as follows:

$$S_{sim} = \theta_1 \cdot D\left(F^I, F^D\right) + \theta_2 \cdot P_{same} \tag{24}$$

IV. Experiments and Evaluation

In this section, we will conduct a series of experiments to evaluate the performance and effectiveness of the proposed method on three benchmark TI-ReID datasets.

A. Datasets, Metrics, and Implementation Details

We utilize three publicly available TI-ReID datasets: CUHK-PEDES [26], ICFG-PEDES [39], and RSTPReid [31], to validate the effectiveness of our approach through experiments. To comprehensively evaluate the efficacy of different methods, we adopt Cumulative Matching Characteristic (CMC) [49] and mean Average Precision (mAP) [50] as evaluative metrics in our experiment.

CUHK-PEDES: The dataset is collected from the screenshots of video and movie camera street shooting images, comprising 13003 person with distinct identities, along with 40206 images of different person and 80440 description texts. The training set consists of 11003 person, 34054 person images, and 68108 person description texts. The verification set includes 2000 person, 3078 person images, and 6156 person descriptions texts.

ICFG-PEDES: The person images in the dataset are all from the MSMT17 dataset, including 4102 different identities, and 54522 image-text pairs, and the description text contains an average of 37.2 words. The dataset is partitioned into training and test sets. The training set includes 34674 image-text pairs of 3102 person, while the test set consists of 1,000 person,19,848 images, and their corresponding text descriptions.

RSTPReid: Person images in this dataset are gathered from the MSMT17 dataset too. However, this dataset is more closely aligned with real-world application scenarios. It contains 4101 different person, 20505 person images, and 41010 corresponding text descriptions. The training set includes 3701 people, 18505 images, and 37010 text descriptions corresponding to images. The validation set consists of 200 people, 1000 images, and 2000 text descriptions.

Implementation Details: First, we employ the Pytorch framework for training and finetuning the proposed method. During the processing, we use ImageNet pre-trained ViT-Base and pre-trained BERT-Base-Uncase network to extract image and text features, respectively. We set the input image size of the network to $384 * 128$, and each text length is unified to 64. The dimensions of image and text features are set to 512. In addition, to better train the network, we employ data augmentation operations, including random erasing, flipping, and random cropping. Each training batch consists of 32 image-text pairs and a total of 80 epochs of iterative training are conducted. The Stochastic Gradient Descent (SGD) optimizer is initialized to $1e-4$ at the beginning of the training process and subsequently reduced by 10% after 50 epochs. We set the hyperparameter α of the fusion feature to 0.4. Additionally, the adjustment parameters θ_1 and θ_2 for similarity calculation are configured as 0.6 and 0.4 respectively. For the L_{C2FL} loss, the adjustment parameters λ_1 and λ_2 are set to 0.35 and 0.65 respectively. The adjustment parameter ω in the total loss function is set to 0.7. All the training and experiments are performed with GeForce RTX 3090Ti GPU.

B. Ablation study

1) Analysis of Intra-Modal Relationship-Mining Module: The experiments are conducted under the CUHK-PEDES dataset. The comparison methods mainly select the representative excellent TI-ReID methods. We adopt ViT and BERT networks as Baseline methods. For comparison, we select the global matching-based TBPSLD [13] method, the local matching-based TIPCB [18], and the LGUR [51] method. To mitigate interference from external variables, all methods utilized ViT and BERT as feature extractors. During experiments, we match person images and texts by computing the cosine distance between features as the measure of similarity. We presented the results in Table I. It can be seen from the experiment

TABLE I
The ablation experiment result (%) of the IMRM module. The best result for each indicator will be bolded.

Method	Rank-1	Rank-5	mAP
Baseline	64.02	82.15	58.27
TBPSLD[13]	64.40	81.27	61.19
TIPCB[18]	64.26	83.16	-
LGUR[51]	65.25	83.12	-
+IPRMG	67.91	85.87	64.64
+TPRMG	68.07	85.53	64.31
+IMRM(IPRMG&TPRMG)	**69.29**	**86.05**	**65.02**

results that the proposed method attained the best performance, achieving 69.29% in Rank-1, 86.05% in Rank-5, and 65.02% in mAP. The following conclusion can be got: 1) Compared with the global matching-based and the local matching-based methods, our TI-ReID method exhibits superior performance, significantly outperforming them. 2) The introduction of the interrelationship information between different local features has effectively enhanced the performance of TI-ReID and the best performance achieved following the introduction of the IMRM module. Consequently, it's from the results that the IMRM module can enhance feature expressiveness and boost the accuracy of re-identification by mining the mutual semantic relationships between different features.

Moreover, to ascertain the influence of the modeling method of text local feature relationships on the TI-ReID performance. An analysis of the text contextual relationship-mining graph is conducted, (a) the relation learning graph that is modeled using only forward text local features (F-TCRMG), (b) the relation learning graph that is modeled using only reverse text local features (B-TCRMG), and (c) the relational learning graph (F&B-TCRMG) that jointly utilizes both forward and backward text local features, i.e. the method used in this paper. We chose BERT as the baseline and Bi-LSTM[46] as the comparison method. Table II shows the experiment result. The F&B-TCRMG achieves the best performance with 68.05% in Rank-1, 84.51% in Rank-5, and 64.32% in mAP. This shows that the best effect is achieved when jointly using the text local features in both directions for modeling and learning semantic relationships. It can be seen from the comparison that although the Bi-LSTM has learned some reverse text features. However, it still cannot achieve the learning effect obtained by inputting text in reverse. Furthermore, when using backward text local features exclusively, the absence of supplementation and guidance from forward text local features would introduce noise, leading to a decrease in recognition effectiveness. In summary, the TCRMG proposed in this article adequately learns the relationships between different features and enhances the distinguishability of person text features by jointly utilizing forward and backward text local features.

2) Analysis of Cross-Modal Relationship-Mining Module: In order to verify the effectiveness and contribution of the design CMRM module, we conduct comparative experiments on the CUHK-PEDES dataset. In the experiment, we selected TBPSLD [13], IVT [52], TIPCB [18], and LGUR [51] for comparison. The ViT and BERT are also selected as baselines, and other experimental settings are the same as before. The experiment results are shown in Table III.

TABLE II
The ablation experiment result (%) of the TPRMG.

Method	Rank-1	Rank-5	mAP
Baseline	64.09	82.16	58.26
Bi-LSTM [46]	64.26	82.58	58.87
+F-TPRMG	66.73	83.72	62.03
+B-TPRMG	64.21	82.05	58.15
+TPRMG(F&B-TPRMG)	**68.05**	**84.51**	**64.32**

TABLE III
The ablation experiment result (%) of the CMRM module.

Method	Rank-1	Rank-5	mAP
Baseline	64.02	82.15	58.27
TBPSLD[13]	64.40	81.27	61.19
IVT[52]	65.23	83.01	-
TIPCB[18]	64.26	83.19	-
LGUR[51]	65.25	83.12	-
+CMRM	69.36	85.29	63.93
+CMRM&IMRM	**73.18**	**88.95**	**68.23**

It can be seen from Table III that the effectiveness of TI-ReID has improved after the introduction of the CMRM module. Compared to the best-performing local matching-based method LGUR, our method outperforms 7.93% in Rank-1 and 5.83% in Rank-5. After the introduction of the IMRM module, the TI-ReID performance has been further improved with 3.82% in Rank-1, 3.66% in Rank-5, and 4.3% in mAP. The best results have been achieved when the IMRM and CMRM are used together. Consequently, it's from the results that the designed CMRM module can effectively model and mine the mutual semantic relationships between different features within different modals. This relationship information can be leveraged to enhance the expressive ability of features, which effectively improves the performance of TI-ReID.

3) Analysis of Training Strategy: First, we perform ablation experiments on the loss function. We choose ViT and BERT as baseline methods that use CMPM Loss for training, to extract person image and text features respectively. In the experiment, the extracted features are directly used for similarity calculation to match person images and texts. Table IV shows the TI-ReID performance of the RMGNet on different training loss. It can be seen that the Acc loss proposed in this paper achieved the second-best re-identification result. When the Acc loss and Hc-Tri loss are used together, the best recognition effect is achieved. Specifically, compared with other loss and baseline, it achieves 68.13% on Rank-1, 84.87% on Rank-5, and 63.03% on mAP. Table V shows the ablation experiment result. We can see that under the guidance of Hc-Tri loss, the TI-ReID model's Rank-1 increased to 65.73%, Rank-5 increased to 83.17% and mAP increased to 60.58%. It is evident from the results that the network can bring different modal features of the same person closer with the aid of the Hc-Tri loss, consequently improving the feature learning ability of the network. When Acc loss is introduced, the search accuracy is further improved. The best results are achieved after introducing two loss functions simultaneously. Compared with the Baseline method, there is an improvement of 3.99% on Rank-1, 2.04% on Rank-5, and 4.02% on mAP. This shows

TABLE IV
THE COMPARISON RESULT (%) OF THE DIFFERENT LOSS FUNCTION.

Loss Function	Rank-1	Rank-5	mAP
Baseline	64.12	82.63	58.83
Circle Loss	65.72	82.96	60.51
Pair-wise Contrastive Loss	65.03	82.83	60.12
Contrastive Loss	64.36	82.14	59.31
Acc Loss	67.52	84.72	61.45
Hc-Tri Loss	65.69	83.04	60.55
Hc-Tri Loss + Acc Loss	**68.13**	**84.87**	**63.03**

TABLE V
THE ABLATION EXPERIMENT RESULT (%) OF THE LOSS FUNCTION.

Strategy	Loss Function		Rank-1	Rank-5	mAP
	Hc-Tri	Acc			
1	×	×	64.15	82.87	59.02
2	✓	×	65.73	83.17	60.58
3	×	✓	67.55	84.76	61.47
4	✓	✓	**68.14**	**84.91**	**63.04**

that the designed Acc loss can effectively alleviate the defects in Hc-Tri loss and provide effective supervision information for the network during the entire training process.

In addition, in order to verify the effectiveness of the proposed Coarse-to-Fine Learning strategy. We perform ablation experiments on the training process, and select TIPCB [18] for comparison. Table VI shows the experimental results.

It can be observed from the table that the least favorable effect is achieved when using Hc-Tri loss alone. The primary reason is that the constraints of Hc-Tri loss are relatively loose. When the majority of triples meet the constraints in the early stages of training, they can no longer offer supervision information for network training. After replacing the Hc-Tri loss with the Acc loss, the Rank-1, Rank-5, and mAP of our proposed method increased by 1.87%, 1.67%, and 0.82% respectively. However, due to its relatively strict constraints, it will cause the method to overfit some early training samples, thereby hindering the achievement of optimal results. When the two losses are comprehensively utilized through the proposed C2FL strategy, the best TI-ReID performance is achieved, with 69.16% in Rank-1, 86.23% in Rank-5, and 64.57% in mAP. Moreover, the TIPCB method, being a TI-ReID approach based on the PCB network, exhibits high sensitivity to constraint strength. Based on the observed changes in results, directly applying Acc loss leads to a decrease in performance. In contrast, after introducing the designed C2FL strategy, such issue can be effectively resolved and further improve accuracy. In summary, the proposed C2FL strategy not only adaptively adjusts the method training process that can provide continuous and effective supervision information, but also demonstrates generalization capabilities.

TABLE VI
THE COMPARISON RESULT (%) OF THE DIFFERENT TRAINING STRATEGY.

Strategy	Rank-1	Rank-5	mAP
Ours(w/o C2FL) + Hc-Tri Loss	65.65	83.15	60.61
Ours(w/o C2FL) + Acc Loss	67.52	84.82	61.43
Ours(w/ C2FL)	**69.16**	**86.23**	**64.57**
TIPCB(w/o C2FL) + Hc-Tri Loss	64.78	83.65	59.83
TIPCB(w/o C2FL) + Acc Loss	64.04	83.94	59.45
TIPCB(w/C2FL)	**66.43**	**84.47**	**61.21**

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

To evaluate the superiority of our proposed TI-ReID method, we compare our method with several state-of-the-art TI-ReID techniques. In these experiments, we conducted the comparisons on the three datasets, CUHK-PEDES, ICFG-PEDES, and RSTPReid. Table VII, Table VIII, and Table IX report the comparisons, and G, L and R represents the global feature-based, the local feature-based and the relationship-based matching methods, respectively. To mitigate interference from other factors, post-processing methods such as re-ranking are not introduced in the experiment. All approaches adhere to the same protocol to ensure equitable comparison.

From the experiment result in Table VII, it can be seen that on the CUHK-PEDES dataset, the proposed method achieves 77.19% Rank-1, 92.18% Rank-5, and 70.59% mAP. Secondly, compared to the global matching-based methods, those local matching-based methods demonstrate superior performance in TI-ReID. This can be attributed to the fact that local matching-based methods are able to extract more discriminative information from images and texts of person. In the experiment, the proposed method achieves the optimum. Especially compared to the best local matching-based method RaSa[53], our method has achieved 0.68% in Rank-1, 1.89% in Rank-5, and 1.21% in mAP improvement. Compared with the best performing global matching method TBPS[54], the RMGNet is 3.65% higher in Rank-1, 3.99% higher in Rank-5, and 5.21% higher in mAP. Compared with the UMUMSA[55] which also employ the relationship between features to perform person matching, since it pays more attention to the semantic alignment of local and global features and ignores the interrelationship between different features, its performance is reduced by 2.94% in Rank-1, 2.35% in Rank-5, and 4.44% on mAP. Meanwhile, compared with IRRA[29], which also pays attention to the interrelationships between features, the proposed method has also significantly improvement, improving 3.81% on Rank-1, improving 2.25% on Rank-5, and improving 4.46% on mAP. This can be attributed to the fact that the graph network, in contrast to the Transformer network, is more suitable for learning the mutual semantic relationships between different features and can also prevent the introduction of other noise information. Therefore, the IMRM and CMRM modules can better model and mine the mutual semantic relationships between different features. Additionally, under the guidance of the proposed C2FL strategy, the designed method enhances expressive ability and discriminative features. Furthermore, it is noteworthy that the proposed method solely utilizes pre-trained ViT and BERT networks to extract image and text features, yet it still achieves the best recognition results. This further proves the effectiveness of our method.

Then, the experiment results in the ICFG-PEDES dataset, which has the more complex TI-ReID scene, are shown in Table VIII. We can learn that the proposed method achieves the most satisfactory performance among all approaches,

TABLE VII
COMPARISON RESULT(%) WITH OTHER METHODS ON CUHK-PEDES.

Type	Method	Reference	Rank-1	Rank-5	mAP
G	GNA-RNN[26]	CVPR 17	19.05	-	-
	Dual Path[28]	TOMM 20	44.40	66.26	-
	CMKA[56]	TIP 20	54.69	73.65	-
	TBPSLD[13]	BMCV 21	64.40	81.27	61.19
	TCTPR [35]	JIG 23	71.70	87.95	-
	BDNet [57]	PR 23	66.27	85.07	57.04
	FLAN [58]	ESWA 24	71.89	88.54	-
	TBPS [54]	AAAI 24	73.54	88.19	65.38
L	MIA[59]	TIP 20	53.10	75.00	-
	MGEL [60]	IJCAI 21	60.27	80.01	-
	ACSA [61]	TMM 22	63.56	81.40	-
	CAIBC [32]	MM 22	64.43	82.87	-
	AXM-Net [62]	AAAI 22	64.44	80.52	58.73
	TIPCB [18]	Neuro 22	64.26	83.19	-
	MMGCN [43]	TMM 23	69.41	87.07	62.49
	APTM [63]	MM 23	76.17	89.47	65.52
	RaSa [53]	IJCAI 23	76.51	90.29	69.38
	CTL [64]	TCSVT 23	69.47	87.13	60.56
	BCALF [65]	EAAI 24	66.39	83.48	-
	SCVD [66]	TCSVT 24	76.72	90.38	-
R	IRRA [29]	CVPR 23	73.38	89.93	66.13
	UMUMSA [55]	AAAI 24	74.25	89.83	66.15
	Ours	This Paper	**77.19**	**92.18**	**70.59**

TABLE VIII
COMPARISON RESULT(%) WITH OTHER METHODS ON ICFG-PEDES.

Type	Method	Reference	Rank-1	Rank-5	mAP
G	TCTPR [13]	JIG 23	60.23	78.09	-
	BDNet [57]	PR 23	57.31	76.15	-
	FLAN [58]	ESWA 24	61.39	78.65	-
	TBPS [54]	AAAI 24	65.05	80.34	39.83
L	MIA [59]	TIP 20	46.49	67.14	-
	ViTTA [67]	ECCV 20	50.98	68.79	-
	TIPCB [18]	Neuro 22	54.96	74.72	-
	SRCF [68]	ECCV 22	57.18	75.01	-
	MMGCN [43]	TMM 23	60.20	76.75	37.56
	APTM [63]	MM 23	68.22	82.87	39.58
	RaSa [53]	IJCAI 23	65.28	80.40	41.29
	CTL [64]	TCSVT 23	57.69	75.79	36.07
	BCALF [65]	EAAI 24	59.31	75.94	-
R	IRRA [29]	CVPR 23	63.46	80.25	38.06
	UMUMSA [55]	AAAI 24	65.62	80.54	38.78
	Ours	This Paper	**68.35**	**83.06**	**41.62**

TABLE IX
COMPARISON RESULT(%) WITH OTHER METHODS ON RSTPREID.

Type	Method	Reference	Rank-1	Rank-5	mAP
G	Dual Path [28]	TOMM 20	43.51	65.44	-
	FLAN [58]	ESWA 24	57.79	80.82	-
	TBPS [54]	AAAI 24	61.95	83.55	48.26
L	IMG-Net [69]	JEI 20	37.60	61.15	-
	AMEN [70]	PRCV 21	38.45	62.40	-
	LBUL [30]	MM 22	45.55	68.20	-
	IVT [52]	ECCVW 22	46.70	70.00	-
	ACSA [61]	TMM 22	48.40	71.85	-
	MMGCN [43]	TMM 23	52.95	75.30	41.74
	BCALF [65]	EAAI 24	47.83	71.12	-
	SCVD [64]	TCSVT 24	62.18	84.26	-
R	IRRA [29]	CVPR 23	60.20	81.30	47.17
	UMUMSA [55]	AAAI 24	63.40	83.30	49.28
	Ours	This Paper	**63.67**	**84.59**	**49.58**

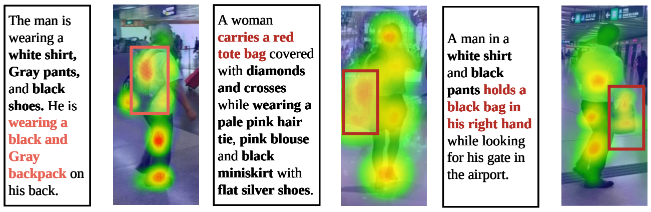

The man is wearing a white shirt, Gray pants, and black shoes. He is wearing a black and Gray backpack on his back.
A woman carries a red tote bag covered with diamonds and crosses while wearing a pale pink hair tie, pink blouse and black miniskirt with flat silver shoes.
A man in a white shirt and black pants holds a black bag in his right hand while looking for his gate in the airport.

Fig. 5. The visualization of cross-modal heatmap of TI-ReID result. The red bold font indicates the information describing of the interrelationship between local features. The red boundingbox area is the key area that contains local feature interrelationship information.

with Rank-1 = 68.35%, Rank-5 = 83.06%, and mAP = 41.62%. Similarly, compared with the second-ranked method APTM[63], our method has 0.13% gains in Rank-1, 0.19% improvement in Rank-5, and 2.04% in gains in mAP. Compared with the TBPS[54], our method has achieved 3.3% in Rank-1, 2.72% in Rank-5, and 1.79% in mAP improvement. Compared with the relationship-based method UMUMSA[55], our method improves 2.73% in Rank-1, 2.52% in Rank-5, and 2.84% in mAP. It shows that the proposed method can perform better TI-ReID in more complex scenarios.

Finally, we also compared our method with state-of-the-art methods on the RSTPReid dataset. The experiment results are shown in Table IX. The experimental results reveal that our method achieves best performance, yielding 63.67% in Rank-1, 84.59% in Rank-5, and 49.58% in mAP. In comparison with the UMUMSA [55], our approach yields improvements of 0.27%, 1.29%, and 0.3% in Rank-1, Rank-5, and mAP, respectively. The comparative experiments conducted across three datasets fully prove the effectiveness and generalization of the proposed method in the TI-ReID task.

To provide a more intuitive illustration of the recognition effect of our method in the TI-ReID task. We randomly select some person text descriptions from the CUHK-PEDES dataset to retrieve their corresponding pedestrian images. Figure 5 shows the cross-modal heat map of the proposed method. It can be seen that with the assistance of the proposed IMRM and CMRM modules, the network can focus on and learn the interrelationships between the local features of person in the text and the image, enabling accurate cross-modal person re-identification. The recognition results are shown in Figure 6, where the first and second rows are the recognition results of the Baseline and our method, respectively. It can be seen from the retrieval results that the proposed method has obtained correct retrieval results in Rank-1. While only one of the Rank-1 search results of the Baseline method is correct. In addition, the correct and incorrect person image results exhibit notable similarities in both overall appearance and local details. For example, in the 2th search result, all person images are wearing white shorts, black pants, black shoes, and carrying black bags. The difference lies in the mutual semantic relationship between the bag and other features. In this situation, our method can effectively model and extract the mutual semantic relationships between features and enable accurate differentiation of persons

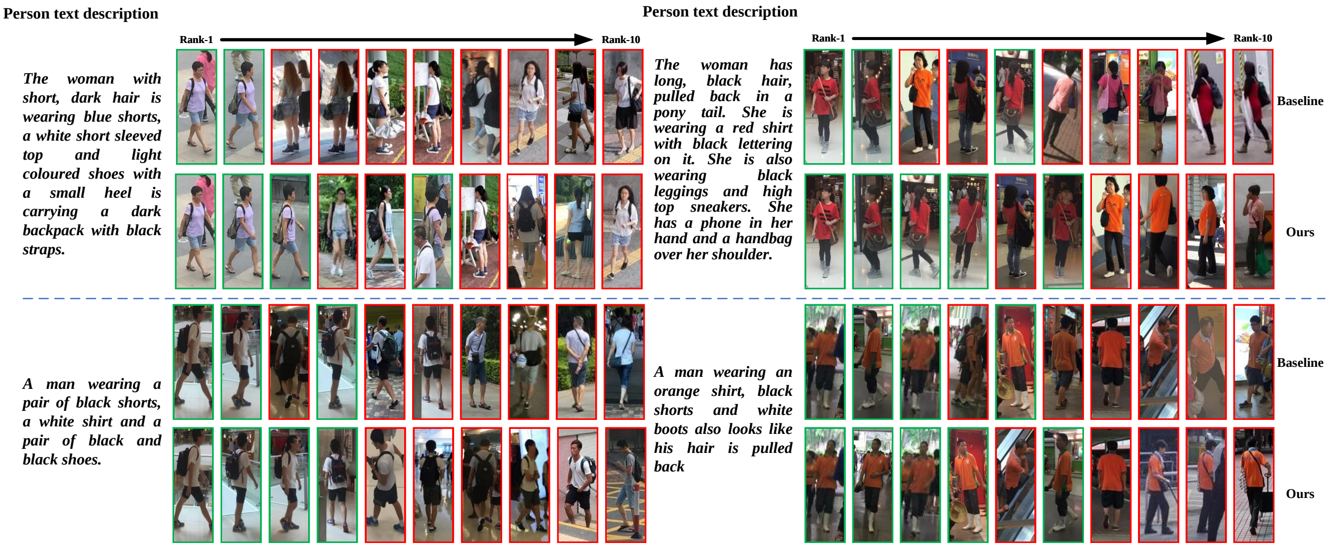

Person text description
Rank-1
Rank-10
The woman with short, dark hair is wearing blue shorts, a white short sleeved top and light coloured shoes with a small heel is carrying a dark backpack with black straps.
Person text description
Rank-1
Rank-10
The woman has long, black hair, pulled back in a pony tail. She is wearing a red shirt with black lettering on it. She is also wearing black leggings and high top sneakers. She has a phone in her hand and a handbag over her shoulder.
Baseline
Ours
A man wearing a pair of black shorts, a white shirt and a pair of black and black shoes.
A man wearing an orange shirt, black shorts and white boots also looks like his hair is pulled back
Baseline
Ours

Fig. 6. Visualization of the text-to-image person re-identification comparison results on the CUHK-PEDES dataset. The first row is the retrieval results of the Baseline method, and the second row is the RMGNet. The green bounding boxes indicate correct results, and the red ones indicate incorrect results.

with similar appearances[71].

V. Conclusion

In this paper, we propose the Progressive Relationship-Mining Graph Network (RMGNet) for the TI-ReID task, which includes the IMRM module, the CMRM module, and the C2FL strategy. Specifically, the IMRM module models and mines hidden relationships between different local features within image and text modals through graph networks. This enhancement the expressiveness and discriminative capabilities of both image and text features. The CMRM module is employed to model the semantic correlation and intrinsic relationship between person features within inter-modal by fusing the nearest neighbor method and GNN. This extracts the affinity relationships of person features in different modalities to reduce the interference caused by the modal gap. The C2FL learning strategy effectively addresses the drawbacks of Hc-Tri loss by employing the Acc loss which offers stricter and more effective supervision information throughout the training process. Furthermore, the strategy enables adaptive adjustments for network optimization. We extensively evaluate our method on three challenging TI-ReID datasets to demonstrate the effectiveness and generalization of the proposed method. In future work, we will try to apply the proposed method to other text-to-image cross-modal tasks, such as text-to-image vehicle re-identification tasks, text-based object tracking tasks, etc. Simultaneously, we aim to optimize the model structure to enhance its adaptability to real-world application scenarios.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3303800, the National Natural Science Foundation of China under Grant U21A20482, 62073117.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021. doi: 10.1109/TPAMI.2021.3054775.

[2] G. Du, T. Gong, and L. Zhang, "Contrastive completing learning for practical text-image person reid: Robuster and cheaper," *Expert Systems with Applications*, p. 123399, 2024. doi: 10.1016/j.eswa.2024.123399.

[3] H. Wang, J. Liu, Y. Su, and X. Yang, "Trajectory guided robust visual object tracking with selective remedy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3425–3440, 2023. doi: 10.1109/TCSVT.2022.3233636.

[4] K. Deng, C. Zhang, Z. Chen, W. Hu, B. Li, and F. Lu, "Jointing recurrent across-channel and spatial attention for multi-object tracking with block-erasing data augmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4054–4069, 2023. doi: 10.1109/TCSVT.2023.3238716.

[5] Y. Zheng, B. Zhong, Q. Liang, G. Li, R. Ji, and X. Li, "Towards unified token learning for vision-language tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3301933.

[6] P. Chen, X. Yu, X. Han, N. Hassan, K. Wang, J. Li, J. Zhao, H. Shi, Z. Han, and Q. Ye, "Point-to-box network for accurate object detection via single point supervision," in *European Conference on Computer Vision*, pp. 51–67, Springer, 2022.

[7] J. Zhao, J. Li, L. Jin, J. Chu, Z. Zhang, J. Wang, J. Xia, K. Wang, Y. Liu, S. Gulshad, *et al.*, "The 3rd anti-uav workshop & challenge: Methods and results," *arXiv preprint arXiv:2305.07290*, 2023.

[8] X. Xiong, W. Min, Q. Wang, and C. Zha, "Human skeleton feature optimizer and adaptive structure enhancement graph convolution network for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 1, pp. 342–353, 2022. doi: 10.1109/TCSVT.2022.3201186.

[9] Y. Mou, X. Jiang, K. Xu, T. Sun, and Z. Wang, "Compressed video action recognition with dual-stream and dual-modal transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3319140.

[10] Z. Ma, Z. Zheng, J. Wei, Y. Yang, and H. T. Shen, "Instance-dictionary learning for open-world object detection in autonomous driving scenarios," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3322465.

[11] C. Tao, J. Cao, C. Wang, Z. Zhang, and Z. Gao, "Pseudo-mono for monocular 3d object detection in autonomous driving," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3962–3975, 2023. doi: 10.1109/TCSVT.2023.3237579.

[12] Z. Wang, J. Zhao, C. Lu, F. Yang, H. Huang, Y. Guo, *et al.*, "Learning to detect head movement in unconstrained remote gaze estimation in the wild," in *Proceedings of the IEEE/CVF Winter Conference on*

Applications of Computer Vision, pp. 3443–3452, 2020.

[13] X. Han, S. He, L. Zhang, and T. Xiang, "Text-based person search with limited data," *arXiv preprint arXiv:2110.10807*, 2021.

[14] S. Yan, N. Dong, L. Zhang, and J. Tang, "Clip-driven fine-grained text-image person re-identification," *IEEE Transactions on Image Processing*, 2023. doi: 10.1109/TIP.2023.3327924.

[15] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1890–1899, 2017.

[16] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, P. Peng, X. Guo, and X. Sun, "Contextual non-local alignment over full-scale representation for text-based person search," *arXiv preprint arXiv:2101.03036*, 2021.

[17] J. Zhou, B. Huang, W. Fan, Z. Cheng, Z. Zhao, and W. Zhang, "Text-based person search via local-relational-global fine grained alignment," *Knowledge-Based Systems*, vol. 262, p. 110253, 2023. doi: 10.1016/j.knosys.2023.110253.

[18] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, "Tipcb: A simple but effective part-based convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171–181, 2022. doi: 10.1016/j.neucom.2022.04.081.

[19] K. Zhang, B. Hu, H. Zhang, Z. Li, and Z. Mao, "Enhanced semantic similarity learning framework for image-text matching," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3307554.

[20] Z. Liu, F. Chen, J. Xu, W. Pei, and G. Lu, "Image-text retrieval with cross-modal semantic importance consistency," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 5, pp. 2465–2476, 2022. doi: 10.1109/TCSVT.2022.3220297.

[21] H. Liu, X. Tan, and X. Zhou, "Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification," *IEEE Transactions on Multimedia*, vol. 23, pp. 4414–4425, 2020. doi: 10.1109/TMM.2020.3042080.

[22] X. Zhang, Y. Ling, Y. Yang, C. Chu, and Z. Zhou, "Center-point-pair detection and context-aware re-identification for end-to-end multi-object tracking," *Neurocomputing*, vol. 524, pp. 17–30, 2023. doi: 10.1016/j.neucom.2022.11.094.

[23] X. Zhang, S. Gao, Y. Yang, C. Chu, and Z. Zhou, "Head point positioning and spatial-channel self-attention network for multi-object tracking," in *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 3500–3506, IEEE, 2022.

[24] M. Xu, H. Guo, Y. Jia, Z. Dai, and J. Wang, "Pseudo label rectification with joint camera shift adaptation and outlier progressive recycling for unsupervised person re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3395–3406, 2022. doi: 10.1109/TITS.2022.3224233.

[25] T. Liang, Y. Jin, W. Liu, T. Wang, S. Feng, and Y. Li, "Bridging the gap: Multi-level cross-modality joint alignment for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. doi: 10.1109/TCSVT.2024.3377252.

[26] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1970–1979, 2017.

[27] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 686–701, 2018.

[28] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020. doi: 10.1145/3383184.

[29] 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.

[30] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and 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*, pp. 1984–1992, 2022.

[31] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, and G. Hua, "Dssl: Deep surroundings-person separation learning for text-based person retrieval," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 209–217, 2021.

[32] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and Y. Li, "Caibc: Capturing all-round information beyond color for text-based person retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5314–5322, 2022.

[33] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5814–5824, 2019.

[34] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4654–4662, 2019.

[35] J. Ding and Y. Mang, "Transformer network for cross-modal text-to-image person re-identification," *Journal of Image and Graphics*, 2023.

[36] W. Zhong, L. Jiang, T. Zhang, J. Ji, and H. Xiong, "A part-based attention network for person re-identification," *Multimedia Tools and Applications*, vol. 79, pp. 22525–22549, 2020. doi: 10.1007/s11042-019-08395-2.

[37] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 902–917, 2019. doi: 10.1109/TPAMI.2019.2938523.

[38] X. Yang, X. Wang, N. Wang, and X. Gao, "Address the unseen relationships: Attribute correlations in text attribute person search," *IEEE transactions on neural networks and learning systems*, 2023. doi: 10.1109/TNNLS.2023.3300582.

[39] Z. Ding, C. Ding, Z. Shao, and D. Tao, "Semantically self-aligned network for text-to-image part-aware person re-identification," *arXiv preprint arXiv:2107.12666*, 2021.

[40] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11189–11196, 2020.

[41] S. Aggarwal, V. B. Radhakrishnan, and A. Chakraborty, "Text-based person search via attribute-aided matching," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2617–2625, 2020.

[42] X. Yang, X. Wang, and D. Yang, "Improving cross-modal constraints: Text attribute person search with graph attention networks," *IEEE Transactions on Multimedia*, 2023. doi: 10.1109/TMM.2023.3297391.

[43] G. Han, M. Lin, Z. Li, H. Zhao, and S. Kwong, "Text-to-image person re-identification based on multimodal graph convolutional network," *IEEE Transactions on Multimedia*, 2023. doi: 10.1109/TMM.2023.3344354.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[46] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja, and C. Du, "A convolution bidirectional long short-term memory neural network for driver emotion recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4570–4578, 2020. doi: 10.1109/TITS.2020.3007357.

[47] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[48] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[49] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1846–1855, 2015.

[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.

[51] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, and C. Ding, "Learning granularity-unified representations for text-to-image person re-identification," in *Proceedings of the 30th acm international conference on multimedia*, pp. 5566–5574, 2022.

[52] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, and X. Wang, "See finer, see more: Implicit modality alignment for text-based person retrieval," in *European Conference on Computer Vision*, pp. 624–641, Springer, 2022.

[53] Y. Bai, M. Cao, D. Gao, Z. Cao, C. Chen, Z. Fan, L. Nie, and M. Zhang, "Rasa: Relation and sensitivity aware representation learning for text-based person search," *arXiv preprint arXiv:2305.13653*, 2023.

[54] M. Cao, Y. Bai, Z. Zeng, M. Ye, and M. Zhang, "An empirical study of clip for text-based person search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 465–473, 2024.

[55] Z. Zhao, B. Liu, Y. Lu, Q. Chu, and N. Yu, "Unifying multi-modal

uncertainty modeling and semantic alignment for text-to-image person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 7534–7542, 2024.

[56] Y. Chen, R. Huang, H. Chang, C. Tan, T. Xue, and B. Ma, "Cross-modal knowledge adaptation for language-based person search," *IEEE Transactions on Image Processing*, vol. 30, pp. 4057–4069, 2021. doi: 10.1109/TIP.2021.3068825.

[57] Q. Liu, X. He, Q. Teng, L. Qing, and H. Chen, "Bdnet: A bert-based dual-path network for text-to-image cross-modal person re-identification," *Pattern Recognition*, vol. 141, p. 109636, 2023. doi: 10.1016/j.patcog.2023.109636.

[58] S. Xie, C. Zhang, E. Ning, Z. Li, Z. Wang, and C. Wei, "Full-view salient feature mining and alignment for text-based person search," *Expert Systems with Applications*, vol. 251, p. 124071, 2024. doi: 10.1016/j.eswa.2024.124071.

[59] K. Niu, Y. Huang, W. Ouyang, and L. Wang, "Improving description-based person re-identification by multi-granularity image-text alignments," *IEEE Transactions on Image Processing*, vol. 29, pp. 5542–5556, 2020. doi: 10.1109/TIP.2020.2984883.

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

[61] Z. Ji, J. Hu, D. Liu, L. Y. Wu, and Y. Zhao, "Asymmetric cross-scale alignment for text-based person search," *IEEE Transactions on Multimedia*, 2022. doi: 10.1109/TMM.2022.3225754.

[62] A. Farooq, M. Awais, J. Kittler, and S. S. Khalid, "Axm-net: Implicit cross-modal feature alignment for person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 4477–4485, 2022.

[63] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," in *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4492–4501, 2023.

[64] H. Wu, W. Chen, Z. Liu, T. Chen, Z. Chen, and L. Lin, "Contrastive transformer learning with proximity data generation for text-based person search," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3329220.

[65] G. Du, H. Zhu, and L. Zhang, "Bottom-up color-independent alignment learning for text–image person re-identification," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109421, 2024. doi: 10.1016/j.engappai.2024.109421.

[66] Z. Wei, Z. Zhang, P. Wu, J. Wang, P. Wang, and Y. Zhang, "Fine-granularity alignment for text-based person retrieval via semantics-centric visual division," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. doi: 10.1109/TCSVT.2024.3392831.

[67] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "Vitaa: Visual-textual attributes alignment in person search by natural language," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 402–420, Springer, 2020.

[68] W. Suo, M. Sun, K. Niu, Y. Gao, P. Wang, Y. Zhang, and Q. Wu, "A simple and robust correlation filtering method for text-based person search," in *European conference on computer vision*, pp. 726–742, Springer, 2022.

[69] Z. Wang, A. Zhu, Z. Zheng, J. Jin, Z. Xue, and G. Hua, "Img-net: inner-cross-modal attentional multigranular network for description-based person re-identification," *Journal of Electronic Imaging*, vol. 29, no. 4, pp. 043028–043028, 2020. doi: 10.1117/1.JEI.29.4.043028.

[70] Z. Wang, J. Xue, A. Zhu, Y. Li, M. Zhang, and C. Zhong, "Amen: Adversarial multi-space embedding network for text-based person re-identification," in *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part II 4*, pp. 462–473, Springer, 2021.

[71] M. Golchoubian, M. Ghafurian, K. Dautenhahn, and N. L. Azad, "Pedestrian trajectory prediction in pedestrian-vehicle mixed environments: A systematic review," *IEEE Transactions on Intelligent Transportation Systems*, 2023. doi: 10.1109/TITS.2023.3291196.

Xin Zhang is a postdoctoral researcher at the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China. He received his B.S. and M.S. degrees from the North University of China, Taiyuan, China, in 2015 and 2018, respectively. He got his Ph.D. degree from Beihang University in 2023. His research interests include Vehicle Re-Identification, Person Re-Identification, Multi-Object Tracking, and Computer Vision.

Kun Liu received the M.S. degree in mechatronic engineering from the Harbin Institute of Technology, Harbin, China, in 2003, and the Ph.D. degree in automation from Tsinghua University, Beijing, China, in 2009. She is currently an Associate Professor with the School of Artificial Intelligence, Hebei University of Technology, Tianjin. Her current research interests include image processing, computer vision, and pattern recognition.

Xinwang Wang received the PhD degree in instrument science and technology from Southeast University, Nanjing, China, in 2024. He is currently an associate researcher of the School of Integrated Circuits in Wuxi Institute of Technology, Wuxi, Jiangsu, China. His research interests include the areas of MEMS inertial devices de-noising, inertial guidance system design and MEMS inertial devi ces..

Zhong Zhou Professor, Ph.D. adviser, State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China. He got his B.S. degree from Nanjing University and Ph.D. degree from Beihang University in 1999 and 2005 respectively. His main research interests include Virtual Reality/Augmented Reality/Mixed Reality, Computer Vision and Artificial Intelligence. He is the member of IEEE, ACM, and CCF.

Haiyong Chen received the M.S. degree in detection technology and automation from the Harbin University of Science and Technology, Harbin, China, in 2005, and the Ph.D. degree in control science and engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently a Professor with the School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin. He is also an expert in the field of photovoltaic cell image processing and automated production equipment. His current research interests include image processing, robot vision, and pattern recognition.