Interact, Embed, and EnlargE (IEEE): Boosting Modality-specific Representations for Multi-Modal Person Re-identification

Zi Wang[3], Chenglong Li[1,2,4], Aihua Zheng[1,2,4] *, Ran He[5], Jin Tang[1,2,3]

[1]Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei, 230601, China
[2]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Hefei, 230601, China
[3]School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[4]School of Artificial Intelligence, Anhui University, Hefei, 230601, China
[5]NLPR, CRIPAC, CASIA, Beijing, China
{ziwang1121, lcl1314, ahzheng214}@foxmail.com, rhe@nlpr.ia.ac.cn, tangjin@ahu.edu.cn

Abstract

Multi-modal person Re-ID introduces more complementary information to assist the traditional Re-ID task. Existing multi-modal methods ignore the importance of modality-specific information in the feature fusion stage. To this end, we propose a novel method to boost modality-specific representations for multi-modal person Re-ID: Interact, Embed, and EnlargE (IEEE). First, we propose a cross-modal interacting module to exchange useful information between different modalities in the feature extraction phase. Second, we propose a relation-based embedding module to enhance the richness of feature descriptors by embedding the global feature into the fine-grained local information. Finally, we propose multi-modal margin loss to force the network to learn modality-specific information for each modality by enlarging the intra-class discrepancy. Superior performance on multi-modal Re-ID dataset RGBNT201 and three constructed Re-ID datasets validate the effectiveness of the proposed method comparing with the state-of-the-art approaches.

Introduction

With the development of infrared cameras in daily surveillance, RGB and near-infrared cross-modal Re-ID evolves to a new branch in Re-ID, which can relieve the limitation of the conventional visible single modality Re-ID in low illuminations. However, it brings additional heterogeneous challenge between modalities to person Re-ID. To overcome the imaging limitations of complex visual situations, there emerges the attempt of RGB-NI-TI multi-modal Re-ID (Zheng et al. 2021) by providing complementary three-modality for each person sample.

The first issue of multi-modal Re-ID is to effectively fuse the complementary information from the multi-modality data. Existing multi-modal fusing schemes mainly fall into three categories: 1) Early fusion, which generally fuses information on image-level before the convolution operation (Wang et al. 2020; Li and Wu 2018). 2) Late fusion, which merges different modality features into one fused feature in the middle or at the end of the network to enhance the feature representation (Li et al. 2018; George and Marcel 2021). 3) Progressive fusion, which incorporates multi-modal information in a progressive way (Zheng et al. 2021) to integrate the information of different scales and modalities. However, only the fixed fusion result participates in the subsequent training, which significantly limits the sensitivity of the network to learn the independent modality-specific information. **To boost the modality-specific information via incorporating the information of other modalities in the fusing phase, we propose a cross-modal interacting module.** We argue to interact/absorb complementary information from other modalities while preserving independent/specific modality information instead of directly fusion, as shown in Fig. 1. Meanwhile, we propose to introduce channel-attention to boost the useful information of the feature from other branches during the interaction. Specifically, we save the modality-specific information by retaining each modal feature extracted by the backbone, and then merge the features from the other modalities followed by channel-attention to strengthen important regions. The fused feature and the original feature are added in the final stage of interaction. It is worth noting that the three modality features are still independent rather than fused into a single one. Each modality branch will no longer affect the others after interaction while continuing independent training.

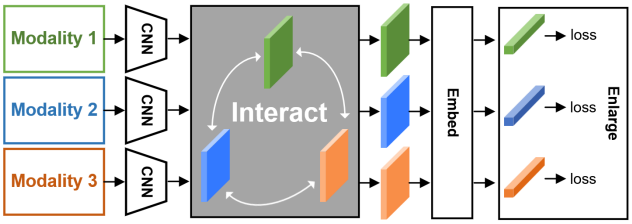

Figure 1: The proposed cross-modal interaction (in the gray box), which exchanges information through the interaction of different modalities.

Second, the deeply-learned local feature has been successfully explored in both Re-ID (Sun et al. 2018; Wei et al. 2017; Yao et al. 2019; Zhao et al. 2017) and other computer vision tasks (Zhang et al. 2014; Cao et al. 2017). Representative local feature based Re-ID methods, including PCB (Sun et al. 2018) and PL-Net (Yao et al. 2019) mainly emphasize the local part feature learning and concat all part features to

*Corresponding author
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

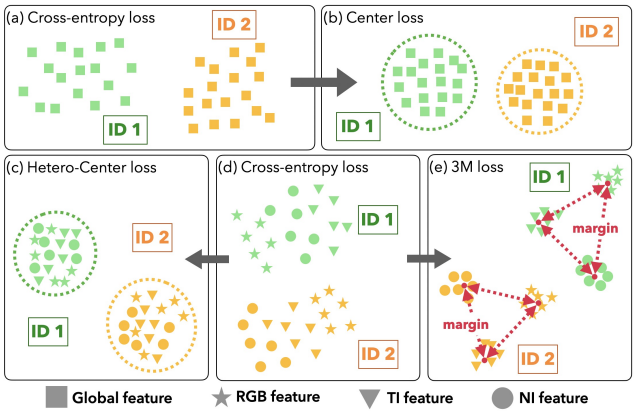

Figure 2: Schematic diagrams of feature distribution under different loss constraints, (a)-(b): Global feature distributions. (c)-(e): Modality feature distributions.

form the final representation for Re-ID while overlooking the crucial global information. **To boost the fine-grained information by global feature while capturing both local and global information during feature learning, we propose a novel relation-based embedding module for Re-ID.** Specifically, we embed the global information into the fine-grained part features to enhance the expressive ability of each part feature. After this embedding module, the feature of each branch contains both local details from part features and the absorbed global information.

At last, as a classification task, most of the Re-ID networks are trained by the commonly used cross-entropy loss (CE loss). Furthermore, Center loss (Wen et al. 2016) gather all samples in each category by reducing the intra-class distance. To reduce the heterogeneity between different modalities, HC loss (Zhu et al. 2020) constrains the distance between modality centers sharing the same identity. However, CE loss and center loss only use the identity information thus lack of consideration of the relationship among modalities. HC loss only focuses on enforcing the heterogeneity while ignoring the complementarity among the multi-modality data. **To boost the feature discrimination among modalities, while simultaneously considering the intra-class cross-modal incongruities and inter-class discrepancies in multi-modal Re-ID, we propose a novel multi-modal margin loss (3M loss).** As shown in Fig. 2 (e), except for enforcing the inter-class discrepancies on the identity level via conventional CE loss, we further enlarge the center distance of the same identity in different modalities to enforce the intra-class cross-modal incongruities. 3M loss can further force the network to learn modality-specific features rather than modality-shared information.

In summary, we aim to boost the modality-specific representation of multi-modal data in three steps: Interact, Embed and EnlargE (IEEE), then concat the features of the three branches to form the final representation to complete the multi-modal person Re-ID task. The main contributions of this paper can be summarized as follows:

- Interact. To absorb complementary information from other modalities and simultaneously maintain the modality-specific information, we propose a cross-modal interacting module (CIM) to exchange the information between modalities during the modality-specific feature learning.
- Embed. To jointly utilize both the local and global information, we design a relation-based embedding module (REM) by embedding the global information into fine-grained local features to enhance the modality-specific feature representation.
- Enlarge. To learn the discriminative representation on multiple modality-specific features, we propose a novel multi-modal margin loss (3M loss) to enlarge the center distance of different modal features and reduce the intra-class cross-modal similarity.

Related Work

Multi-modal Fusion Schemes

General methods of multi-modal fusion can be divided into the following categories: early fusion, late fusion and progressive fusion. 1) Early fusion, also called image-level fusion. It fuses multi-modality data with consistent information into a single input, followed by the subsequent network training (Wang et al. 2020; Li and Wu 2018). However, this type of fusion methods emphasize the subjective visual effect and only fuse the shallow texture information. 2) Late fusion, also called feature-level fusion, which is widely used in face anti-spoofing (George and Marcel 2021), salient object detection (Fu et al. 2020), object tracking (Zhu et al. 2019) and other fields (Joze et al. 2020). They generally gather the multi-modality features extracted by the corresponding branch to obtain the final feature representation. However, they fail to focus on specific modality information due to the large heterogeneity between multiple modalities. 3) Progressive fusion method (Zheng et al. 2021) fuses the multi-modality features progressively instead of at once, which can improve the positive effects of complementary information and better aggregate details from each modality. However, it ignores the heterogeneity between modalities, which may contain crucial discriminative information for distinguishing different persons.

Cross-modal and Multi-modal Re-ID

Visible light cameras may fail to capture person images due to environmental factors such as low illumination. Thus, cross-modal person Re-ID dataset SYSU MM01 (Wu et al. 2017) and RegDB (Nguyen et al. 2017) are proposed to solve this problem by introducing near-infrared and thermal modality, respectively. The images in query and gallery have different modalities, thus the significant gap between visible and near-infrared/thermal images is the main challenge of cross-modal person Re-ID. Li et la. (2020a) use an X-modality generated by a self-supervised learning network as the aid to reduce the heterogeneity between RGB and NI. Zhao et la. (2019) modify the RGB-based models to fit the cross-modal Re-ID task by applying pentaplet loss and selecting difficult pentaplet pairs during the training phase.

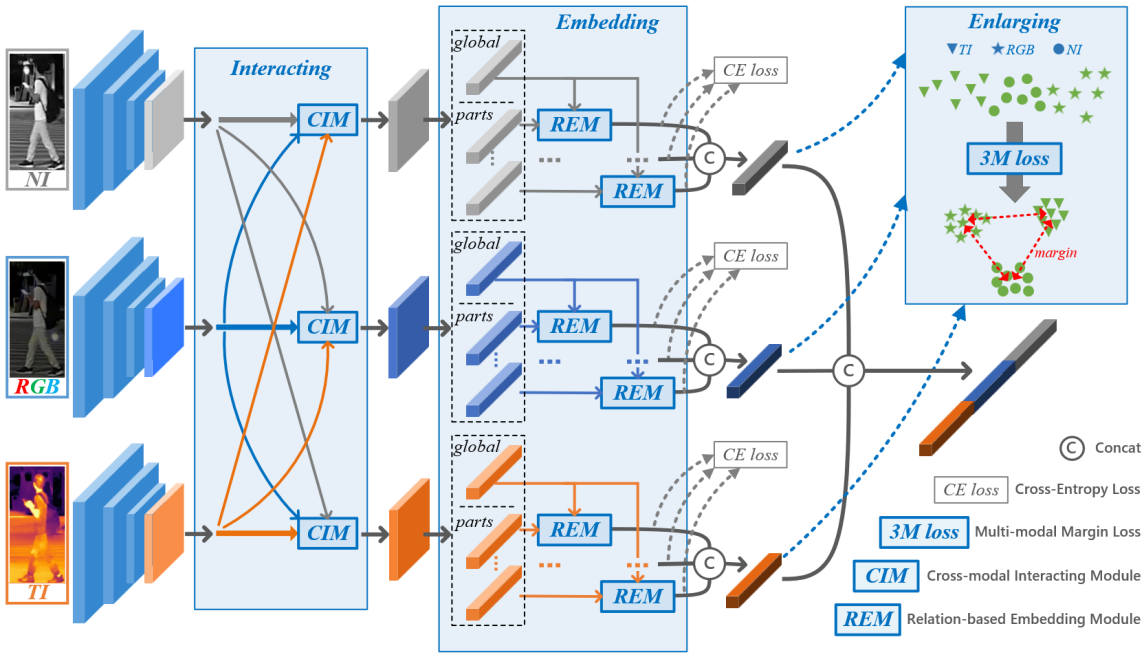

Figure 3: An overview of our proposed Interact, Embed and EnlargE (IEEE). First, we extract features of multi-modal images by using three independent branches. Second, we send the extracted features to the cross-modal interacting module for information exchanging. Then, we propose a relation-based embedding module to enhance the part features by considering the global information of global feature. Finally, we concat all the three modality features as the final person representation. The whole training processing is under the constraints of cross-entropy loss and the proposed multi-modal margin loss.

To incorporate additional modalities to complement the RGB information, some RGB-Depth multi-modal datasets are proposed (Barbosa et al. 2012; Munaro et al. 2014). However, since depth image quality is largely affected by distance, depth data is only applicable to indoor scenes. Zheng et al. (2021) create a multi-modal dataset RGBNT201 containing visible, near-infrared and thermal modalities. To mine the complementary information in multi-modal Re-ID, PFnet (Zheng et al. 2021) launches a baseline by extracting different modal features via multi-branch network and then progressively fuses two modalities (RGB-NI and RGB-TI) in part-level. However, it neglects to exploit the diverse heterogeneity among the modalities.

Proposed Method

The proposed method *IEEE (Interact, Embed and EnlargE)* consists of three main components, cross-modal interacting module (CIM), relation-based embedding module (REM), and multi-modal margin loss (3M loss) with the three-stream backbone architecture, as shown in Fig. 3.

Three-stream Feature Extracting Network

In order to obtain the features of each modality and ensure the maximum retention of modality-specific information, we use a three-stream network structure to process the multi-modality data separately. As shown in Fig. 3, we select ResNet50 (He et al. 2016) as backbone network to extract features from the original input multi-modality data. Due to the different imaging principles, each modality has its own concerned area. The diversity of these modalities is critical for multi-modal person Re-ID. We first use three independent (without parameter sharing) ResNet50 networks to retain the high-quality representations in the feature extraction stage. After ResNet50 extraction stage, the features contain the information of their respective modality since the parameters of the three branches are not shared.

Cross-modal Interacting Module (CIM)

As shown in Fig. 4, after obtaining the three modality features via the three-stream feature extract stage, we propose a cross-modal interacting module (CIM) to help the current modality to absorb the information of other modalities. Specifically, we integrate the information of other modalities through the summation, then enhance the useful information from other modalities by using channel-attention. Finally, in order to maintain the current modal information, we combine the boosted feature with the current modal feature by summation. The interaction will be repeated in K times, where K represents the total number of modalities, and $K = 3$ in our paper.

First, we implement pixel-level summation on RGB feature f_{RGB}^{ori} and NI feature f_{NI}^{ori} extracted by the backbone to obtain the interacted feature f_{RN}^{sum}. Then, we implement 1×1 convolutional layer on f_{RN}^{sum} and TI feature f_{TI}^{ori} to obtain $f_{RN}^{sum'}$ and $f_{TI}^{ori'}$.

In order to emphasize the positive effects of the fea-

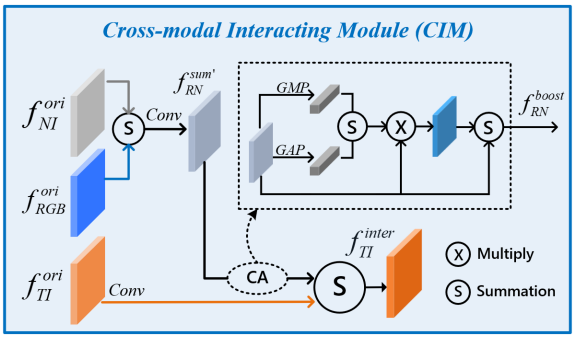

Figure 4: The details of cross-modal interacting module (CIM), taking TI modality interacts with RGB and NI as an example.

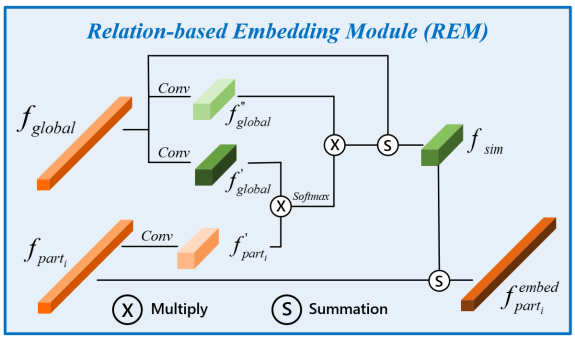

Figure 5: The details of relation-based embedding module (REM), taking the global feature to enhance the i-th part feature as an example.

tures from other modalities, we further propose to introduce channel-attention (Woo et al. 2018) on $f_{RN}^{sum'}$ to obtain the boosted feature f_{RN}^{boost}:

$$f_{RN}^{boost} = CA(f_{RN}^{sum'}) * f_{RN}^{sum'} + f_{RN}^{sum'}, \tag{1}$$

where $CA(\cdot)$ denotes the channel-attention. Sequentially, we add the TI feature $f_{TI}^{ori'}$ to the boosted feature to obtain the interacted feature of TI modality f_{TI}^{inter}:

$$f_{TI}^{inter} = f_{TI}^{ori'} + f_{RN}^{boost}. \tag{2}$$

In the same manner, we can obtain the interacted features of RGB and NI modalites f_{RGB}^{inter} and f_{NI}^{inter}. After CIM, the interacted feature of each modality not only integrates useful information from others, but also retains the modality-specific information.

Relation-based Embedding Module (REM)

Relation-based embedding module (REM) devotes to embed context information from global feature into each local feature by considering their relation. We first implement adaptive average pooling on the feature obtained from CIM to get f_{global} and P part features f_{part_i}, $i \in [1, ..., P]$. To absorb global information for each local feature, we embed the global information into each fine-grained part feature. For the sake of convenience, we use the i-th part feature f_{part_i} and the global feature f_{global} in TI branch to introduce REM module. First, we implement three 1×1 convolution layers on part and global features to obtain $f_{global}^{'}$, $f_{global}^{''}$ and $f_{part_i}^{'}$ as shown in Fig. 5. Then we perform dot product operation on $f_{part_i}^{'}$ and $f_{global}^{'}$ followed by softmax to obtain their similarity as a weight to measure the relationship between global and part. This process can formulated as:

$$V_{sim} = f_{part_i}^{'} \odot f_{global}^{'}. \tag{3}$$

Finally, to integrate the global and fine-grained local information, we implement summation operation on f_{sim} and the original part feature to obtain the feature $f_{part_i}^{embed}$:

$$\begin{aligned} f_{sim} &= f_{global} + V_{sim} * f_{global}^{''}, \\ f_{part_i}^{embed} &= f_{part_i} + f_{sim}. \end{aligned} \tag{4}$$

Note that each part feature has different relationship with the global feature, we enhance each part by global feature individually. Then we concat all part features to obtain the embedded feature of TI modality f_{TI}^{embed}, which is ensured to contain the fine-grained information of body parts.

In the same manner, we can obtain the embedded features of RGB modality f_{RGB}^{embed} and NI modality f_{NI}^{embed}. We integrate diverse information from different modalities by concatting all the features to form the final representation of the individual person:

$$f_{final} = C(f_{TI}^{embed}, f_{RGB}^{embed}, f_{NI}^{embed}), \tag{5}$$

where $C(\cdot)$ denotes the concatenation operation.

The final descriptor f_{final} contains the diverse modality-specific information from different modalities, and the modality-specific features extracted from corresponding modal branches have interacted with other modalities and boosted by the global information through embedding.

Multi-modal Margin Loss

In multi-modal person Re-ID, each person consists of a triplet of three modality images. To learn the modality-specific information, we propose a novel multi-modal margin loss (3M loss) to enlarge the distance between modality centers in each triplet. The proposed 3M loss aims to increase the diversity of intra-class modal information while ensuring the discrimination between IDs. Our multi-modal margin loss can be formulated as:

$$\begin{aligned} L_{3M} = &Max\left(|margin - d(m_j, m_k)|\right), \\ &m_j, m_k \in \{RGB, NI, TI\}, \end{aligned} \tag{6}$$

where the $|\cdot|$ denotes the absolute value, $margin$ denotes the expected center distance between each two modalities. $d(m_j, m_k)$ denotes the $L2$ distance between the centers of m_j and m_k modalities, and we choose the largest distance among the distances between each two modalities as the

value of 3M loss. We also implement cross-entropy loss (CE loss) to extract identity information for classification. CE loss of each person is calculated as follow:

$$L_{CE} = -\sum_{m=1}^{M}\sum_{p=1}^{P} y^{GT}\log\left(y_p^m\right), \tag{7}$$

where P and M denote the number of part (strip), and the total number of modalities respectively, while y^{GT} and y_p^m are the ground truth label and the predicted identity label of the p_{th} part (local) feature in m_{th} modality respectively.

The final loss in the training phase can be formulated as:

$$L_{final} = L_{CE} + \delta * L_{3M}, \tag{8}$$

where δ is the balance hyperparameter between CE loss and 3M loss set as δ = 1 in our training phase. 3M loss forces the network to focus on the modal-specific information of each modality to improve the diversity of feature.

Difference from Cross-modal Re-ID

Compared to the widely explored cross-modal person Re-ID task, which aims to query one single modality image from gallery in another modality, multi-modal Re-ID devotes to query the multiple multi-modal images (in triplet for three-modality case) from the gallery with the same multi-modal fashion for each sample. Therefore, cross-modal Re-ID focuses on the heterogeneity across two modalities while multi-modal Re-ID dedicates to the complementarity among the multi-modality data. One can construct the cross-modal scenario by deleting some modality data from the multi-modal Re-ID scenario. However, the performance of the state-of-the-art cross-modal Re-ID methods significantly decline due to the lack of the multi-modal complementary and the presence of the cross-modal heterogeneity, as we evaluated in Table 2.

Implementation Details

The implementation platform is Pytorch with a GeForce RTX 3090 GPU. We use three ResNet50 networks pre-trained on ImageNet as the backbone to extract features. The original learning rate is set as 0.001, and we reduce the learning rate by 10 times in epoch 20 and epoch 40. The number of mini-batches is 8. The feature maps after CIM module are equally split into 6 stripes. The dimension of each part feature is reduced to 128 by the FC layer. Thus, the feature dimension of each modality (f^{embed}) is 6 $\times$ 128 = 768, and the final feature (f_{final}) of the individual person is 768 $\times$ 3 = 2304-dim. Both cross-entropy loss and multi-modal margin loss are used in training phase, we set the $margin$ in multi-modal margin loss to 1, and the δ in final loss to 1. We use Stochastic Gradient Descent (SGD) with the momentum of 0.9 and weight decay of 0.0005 to fine-tune the network.

Experiments

We evaluate the proposed method IEEE on the benchmark multi-modal person Re-ID datasets RGBNT201 (Zheng et al. 2021) and constructed multi-modal dataset based on Market1501 (Zheng et al. 2015), comparing to state-of-the-art methods.

Dataset and Evaluation Protocols

RGBNT201 (Zheng et al. 2021) is the first multi-modal person Re-ID dataset. It contains 4787 image triplets of 201 persons, while 141 identities for training, 30 identities for validation, and 30 identities for testing. RGBNT201 dataset is collected on campus in four non-overlapping views, each of which consists of a triplicated cameras to simultaneously record RGB, NI and TI data. It offers diverse information and challenges for multi-modal person Re-ID task.

Market1501 (Zheng et al. 2015) is a scalable RGB single-modal person Re-ID dataset. To construct the multi-modal Market 1501, We first generate the TI modality images from the RGB images via cycleGAN (Zhu et al. 2017). Then we transfer the RGB images into gray ones as the supplementary NI modality. Last, to simulate the night scene, we reduce 60% of the brightness of all the images in RGB modality. The training and testing splitting is consistent with the original Martket1501, with 750 identities for training and 751 ones for testing.

Evaluation Protocols. In our experiments, we measure the similarity between two features by euclidean distance. We employ the mean Average Precision (mAP) and Cumulative Matching Characteristic curve (CMC) to compare the performance of our proposed method with other methods, where rank-n indicates the first n closest samples to the $query$ with the same ID from different cameras according to the distance measurement (Euclidean Distance).

Comparison with State-of-the-Art Methods

First of all, we compare our method (IEEE) with the existing multi-modal person and vehicle Re-ID methods PFNet (Zheng et al. 2021) and HAMNet (Li et al. 2020b) respectively. To verify the robustness of our method on multi-modal Re-ID task, we further extend three state-of-the-art single-modal methods, MLFN (Chang, Hospedales, and Xiang 2018), HACNN (Li, Zhu, and Gong 2018), OSNet (Zhou et al. 2019) for comparison. Specifically, we implement the single-modal method on each one of three branches to extract the feature of each modality, then concat all features as the representation for Re-ID.

As shown in Table 1, our method achieves superior performance than all the compared methods. The single-modal Re-ID methods generally work overshadowed by multi-modal methods. The main reason is the single-model methods lack of ability to deal with the heterogeneous multi-modal information. Moreover, these methods can not effectively mine the complementary information among diverse modalities, which is crucial to multi-modal Re-ID.

As the representative multi-modal Re-ID method, HAMNet designs a heterogeneity-collaboration loss to enforce the prediction results of each branch. However, it uses the addition operation on the outputs of the multi-stream backbone, which ignores the importance of feature interaction between different modalities. Although PFNet aims to discover the complementary information among multi-modality data and proposes a progressive fusing scheme, it pays too much attention to the fused feature while losing the sensitivity to the modality-specific information. Therefore, both HAMNet

Methods		RGBNT201				Market1501 (multi-modal version)			
		mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Single-modal	MLFN	24.66	23.68	38.52	49.52	42.69	68.11	87.11	91.95
	HACNN	19.34	14.71	25.48	32.78	42.90	69.15	86.64	92.22
	OSNet	22.12	22.85	37.20	45.93	39.71	69.33	86.67	91.30
Multi-modal	HAMNet	27.68	26.32	41.51	51.67	59.96	82.84	92.50	94.96
	PFNet	38.46	38.88	52.03	58.37	60.93	83.61	92.84	95.49
Ours	**IEEE**	**46.42**	**47.13**	**58.49**	**64.23**	**64.32**	**83.93**	**92.96**	**95.69**

Table 1: Experimental results of our method on RGBNT201 and Market1501 (multi-modal version) comparing with state-of-the-art single-modal and multi-modal methods (in %).

Methods		RGB to TI		TI to RGB		RGB to NI		NI to RGB	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Cross-modal	HC loss	16.74	14.25	16.53	19.32	20.54	22.19	21.80	22.93
	DDAG	18.09	14.79	20.01	18.05	21.39	20.37	24.76	21.66
	MPANet	19.00	20.82	20.28	23.14	25.04	30.96	26.03	26.54
Ours (Multi-modal)	**IEEE**	mAP: **40.41**		Rank-1: **40.79**		mAP: **40.72**		Rank-1: **41.51**	

Table 2: Experimental results of state-of-the-art cross-modal person Re-ID methods on two cross-modal datasets reconstructed by RGBNT201. *RGB*, *TI* and *NI* indicate *visible*, *thermal* and *near infrared* respectively in cross-modal Re-ID.

and PFNet present limited improvement on multi-modal person Re-ID.

Our method (IEEE) significantly boosts the performance on both mAP and ranking scores, which verifies the effectiveness of our method of fusing complementary information while suppressing the heterogeneous issue among the multi-modality data.

Comparison with Cross-modal Re-ID Methods

To verify the necessity of multi-modal Re-ID comparing with the cross-modal scenario, we construct the multi-modal dataset RGBNT201 into cross-modal setting. Specifically, we retain RGB and TI images to achieve the same modality setting as RegDB (Nguyen et al. 2017), and keep RGB and NI photos to imitate SYSU dataset (Wu et al. 2017). For fair comparison, we evaluate our method on corresponding two-modality scenarios. We compare the our model with three state-of-the-art cross-modal person Re-ID methods including HC loss (Zhu et al. 2020), DDAG (Ye et al. 2020) and MPANet (Wu et al. 2021) as shown in Table 2. Due to the huge heterogeneity across modalities, all the three cross-modal Re-ID methods present modest performance. On the contrary, our method achieves promising performance via utilizing the complementary multi-modal information, which verifies the effectiveness of multi-modal Re-ID.

Ablation Study

Our method consists of three key components, cross-modal interacting module (CIM), relation-based embedding module (REM) and multi-modal margin loss (3M loss). To evaluate the contribution of each component in our model, we conduct an ablation experiment on RGBNT201 by progressively introducing each component as shown in Table 3.

	Modules			RGBNT201			
	CIM	3M	REM	mAP	R-1	R-5	R-10
a	×	×	×	21.55	21.89	33.97	41.27
b	✓	×	×	31.94	30.26	46.05	53.47
c	✓	✓	×	39.24	39.59	54.67	64.83
d	✓	✓	✓	**46.42**	**47.13**	**58.49**	**64.23**

Table 3: Ablation study on Cross-modal Interacting Module (CIM), Multi-modal Margin Loss (3M) and Relation-based Enhance Module (REM) on RGBNT201 (in %). Results in line *a* refer to the three-stream extracting network supervised by cross-entropy loss.

First of all, the substantial improvement of the results in Table 3 (a) - (b) verifies the effectiveness of CIM. It demonstrates that CIM can exchange meaningful complementary. Second, the significant improvement in Table 3 (c) comparing to Table 3 (b) demonstrates the promising role of our multi-modal margin loss (3M loss) in multi-modal person Re-ID task compared with only cross-entropy loss. Finally, the introducing of REM further enhances the results, as verified in Table 3 (d), which verifies the benefit of the fine-grained information in part features and embedded global information.

Evaluation on Cross-modal Interacting Module

Cross-modal Interacting Module (CIM) aims to exchange modality-specific information and enhance the the advantages of feature from other branches by interaction and channel-attention. To evaluate the effectiveness of the proposed CIM, we compare our CIM with one early fusion scheme (image summation), two late fusion schemes (feature summation and feature aggregation) as JL-DCF (Fu

et al. 2020) and DAPNet (Zhu et al. 2019) respectively, and the progressive fusion schemes as PFNet (Zheng et al. 2021).

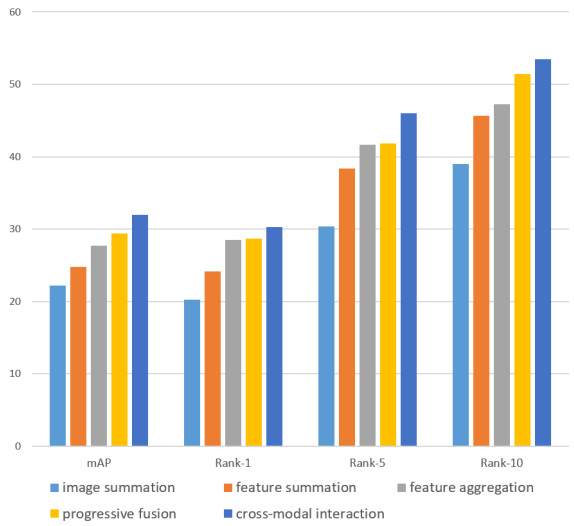

Figure 6: Comparison of our cross-modal interaction module (CIM) against different fusion schemes.

First of all, we can observe from Fig. 6 that both late fusion and progressive fusion are better than early fusion. It is mainly because the early fusion method fuses shallow information on the channel-level of image, and may simultaneously aggregate the noise among different modalities. In contrast, the features processed by convolutional network contain more modal characterized information in late and progressive fusing schemes, which achieve better accuracy than fusing on feature-level. Considering the importance of different regions in the fusion phase, our CIM exchanges the information between modalities simultaneously weight different areas of the feature, thus leads to superior performance.

Evaluation on Multi-modal Margin Loss

The propsoed Multi-modal Margin Loss (3M Loss) aims to improve diversity of modal information by enlarging the distance of multi-modal feature centers. To further evaluate the contribution of the 3M loss, we conduct the comparison experiments by training our model with three state-of-the-art losses, including cross-entropy (CE) loss, hetero center (HC) loss, center loss. Although center loss further suppresses the inter-class differences comparing to the conventional CE loss (Fig 7 (a)), it cannot distinguish the modal level feature distribution due to the lackness of the relationship of multi-modality data, as shown in Fig 7 (b). By considering the heterogeneity between modalities, HC loss aggregates different modality features of the same person, as shown in Fig 7 (c). However, it still ignores the modality-specific information learning in the complementary multi-modality data. As shown in Fig 7 (d), 3M loss simultaneously enlarges the intra-class cross-modal distance improves the inter-class

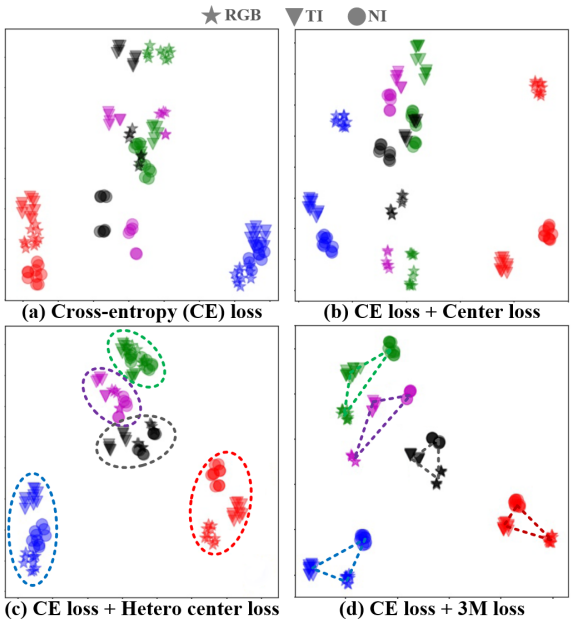

Figure 7: T-SNE (Maaten and Hinton 2008) visualization of the modal level feature distributions under the supervision of different loss combinations. Different colors and shapes represent different IDs and modalities respectively.

difference by learning complementary modality-specific information.

Conclusion

In this paper, we have proposed a novel multi-modal person Re-ID method by boosting the modality-specific representations via inter-modal interacting, part-global embedding and intra-class cross-modal distance enlarging. First, it extracts the multi-modality features by the three-stream network, and exchanges the useful information by cross-modal interaction. In addition, it embeds the the global information to enhance local fine-grained features of individual modality. Furthermore, to explore the cross-modal complementary information for each modality, it learns modality-specific features rather than modality-similar information by enlarging the center distance among different modalities by the proposed multi-modal margin loss. Extensive experiments on challenging multi-modal person Re-ID datasets demonstrate the performance of our proposed method. In the future, we will focus on more effective models to resist drastic changes in the environment for more diverse challenging multi-modal person Re-ID scenarios.

Acknowledgments

This research is supported in part by the Major Project for New Generation of AI under Grant (2020AAA0140002) and the National Natural Science Foundation of China (61976002, 61976003, 62076003 and 61860206004).

References

Barbosa, I. B.; Cristani, M.; Del Bue, A.; Bazzani, L.; and Murino, V. 2012. Re-identification with rgb-d sensors. In *Proceedings of European Conference on Computer Vision*, 433–442.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291–7299.

Chang, X.; Hospedales, T. M.; and Xiang, T. 2018. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2109–2118.

Fu, K.; Fan, D.-P.; Ji, G.-P.; Zhao, Q.; Shen, J.; and Zhu, C. 2020. Siamese network for RGB-D salient object detection and beyond. *arXiv preprint arXiv:2008.12134*.

George, A.; and Marcel, S. 2021. Cross Modal Focal Loss for RGBD Face Anti-Spoofing. *arXiv preprint arXiv:2103.00948*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Joze, H. R. V.; Shaban, A.; Iuzzolino, M. L.; and Koishida, K. 2020. MMTM: multimodal transfer module for CNN fusion. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 13289–13299.

Li, C.; Wu, X.; Zhao, N.; Cao, X.; and Tang, J. 2018. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing*, 281: 78–85.

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020a. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4610–4617.

Li, H.; Li, C.; Zhu, X.; Zheng, A.; and Luo, B. 2020b. Multi-Spectral Vehicle Re-Identification: A Challenge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11345–11353.

Li, H.; and Wu, X.-J. 2018. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, W.; Zhu, X.; and Gong, S. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2285–2294.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Munaro, M.; Basso, A.; Fossati, A.; Van Gool, L.; and Menegatti, E. 2014. 3D reconstruction of freely moving persons for re-identification with a depth sensor. In *Proceedings of IEEE International Conference on Robotics and Automation*, 4512–4519. IEEE.

Nguyen, D. T.; Hong, H. G.; Kim, K. W.; and Park, K. R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3): 605.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of European Conference on Computer Vision*, 480–496.

Wang, X.; Li, S.; Chen, C.; Fang, Y.; Hao, A.; and Qin, H. 2020. Data-level recombination and lightweight fusion scheme for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 30: 458–471.

Wei, L.; Zhang, S.; Yao, H.; Gao, W.; and Tian, Q. 2017. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, 420–428.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *Proceedings of European Conference on Computer Vision*, 499–515.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 5380–5389.

Wu, Q.; Dai, P.; Chen, J.; Lin, C.-W.; Wu, Y.; Huang, F.; Zhong, B.; and Ji, R. 2021. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4330–4339.

Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; and Tian, Q. 2019. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6): 2860–2871.

Ye, M.; Shen, J.; J. Crandall, D.; Shao, L.; and Luo, J. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Proceedings of European Conference on Computer Vision*, 229–247.

Zhang, N.; Donahue, J.; Girshick, R.; and Darrell, T. 2014. Part-based R-CNNs for fine-grained category detection. In *Proceedings of European Conference on Computer Vision*, 834–849.

Zhao, L.; Li, X.; Zhuang, Y.; and Wang, J. 2017. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 3219–3228.

Zhao, Y.-B.; Lin, J.-W.; Xuan, Q.; and Xi, X. 2019. HPILN: a feature learning framework for cross-modality person re-identification. *IET Image Processing*, 13(14): 2897–2904.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust Multi-Modality Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3529–3537.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 1116–1124.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3702–3712.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, Y.; Li, C.; Luo, B.; Tang, J.; and Wang, X. 2019. Dense Feature Aggregation and Pruning for RGBT Tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*, 465–472.

Zhu, Y.; Yang, Z.; Wang, L.; Zhao, S.; Hu, X.; and Tao, D. 2020. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386: 97–109.