Robust Multi-Modality Person Re-identification

Aihua Zheng, Zi Wang†, Zihan Chen†, Chenglong Li*, Jin Tang
Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology, Anhui University, Hefei, China
{ahzheng214, ziwang1121, zhchen96, lcl1314}@foxmail.com, tangjin@ahu.edu.cn

Abstract

To avoid the illumination limitation in visible person re-identification (Re-ID) and the heterogeneous issue in cross-modality Re-ID, we propose to utilize complementary advantages of multiple modalities including visible (RGB), near infrared (NI) and thermal infrared (TI) ones for robust person Re-ID. A novel progressive fusion network is designed to learn effective multi-modal features from single to multiple modalities and from local to global views. Our method works well in diversely challenging scenarios even in the presence of missing modalities. Moreover, we contribute a comprehensive benchmark dataset, RGBNT201, including 201 identities captured from various challenging conditions, to facilitate the research of RGB-NI-TI multi-modality person Re-ID. Comprehensive experiments on RGBNT201 dataset comparing to the state-of-the-art methods demonstrate the contribution of multi-modality person Re-ID and the effectiveness of the proposed approach, which launch a new benchmark and a new baseline for multi-modality person Re-ID.

Introduction

The last decade has witnessed an exponential surge in person re-identification (Re-ID). However, primary efforts on a single visible modality faces severe challenges in adverse illumination and weather conditions like total darkness and dense fog, which restrict its applications in all-day and all-weather surveillance. For example, as shown in Fig. 1 (a) and (b), RGB images almost become invalid in harsh lighting conditions and the performance of RGB-based person Re-ID would thus be limited.

To overcome imaging limitations of visible sensors, Wu *et al.* (2017) propose an RGB and Near Infrared (RGB-NI) dataset SYSU-MM01 for cross-modal person Re-ID, which has been drawn much more attention in recent years from both academic and industrial communities (Dai et al. 2018; Hao et al. 2019a; Wang et al. 2019; Ye et al. 2018; Li et al. 2020). However, the heterogeneous properties across RGB and NI modalities caused by distinct wavelength ranges bring additional challenges to person Re-ID. In addition, the imaging quality of NI is still limited in some challenging scenarios, such as high illumination. For instance, for the ID1 and ID2 in Fig. 1 (a) and (b), their NI images are significantly affected.

To incorporate the complementary information among the multi-modality information, there emerges the attempt of RGBD dual-modality Re-ID (Barbosa et al. 2012; Munaro et al. 2014b) by introducing the depth information. However, the existing depth data are captured in indoor conditions which significantly limit its applications and attractions in both research and industry communities.

In this paper, we propose a new task named multi-modality person Re-ID by integrating RGB, TI (Thermal) and NI (Near Infrared) source data for robust person Re-ID. These three kinds of data have diverse ranges of spectrum, as shown in Fig. 1 (a), and could provide strong complementary benefits in person Re-ID. For example, NI information can overcome the low illumination therefore provides more visible information especially in low illumination, as the person ID3 in Fig. 1 (b). Comparing to NI images, TI information is insensitive to lighting conditions, and has a strong ability to penetrate haze and smog, even in a long distance surveillance, which provides more discriminative information between human bodies and surrounding environments, or auxiliary costumes, as the person ID1 and ID2 in Fig. 1 (b). Introducing NI and TI cameras/modalities into RGB one has perspective applications including all-day all-weather security monitoring, long distance drone investigation, autonomous vehicle in complex environments, etc.

The new task of multi-modality person Re-ID raises three major problems. 1) **How to design a suitable baseline** algorithm to effectively leverage the complementary benefits of all modalities for robust multi-modality person Re-ID even in the presence of missing modalities. 2) **How to create a reasonable size benchmark dataset** for the comprehensive evaluation of different multi-modality person Re-ID algorithms. 3) **How much do each modality, each or diverse combinations of multiple modalities contribute** in multi-modality person Re-ID.

To address above problems, we first **design a progressive fusion network (PFNet)** to learn robust RGB-NI-TI features for multi-modality person Re-ID. Specifically, we employ three individual branches to extract single-modality features and integrate the spatial attention operations to capture more meaningful regions within each modality. Then we fuse the features of each two modalities in part-level to

*Corresponding author, † Equal contribution

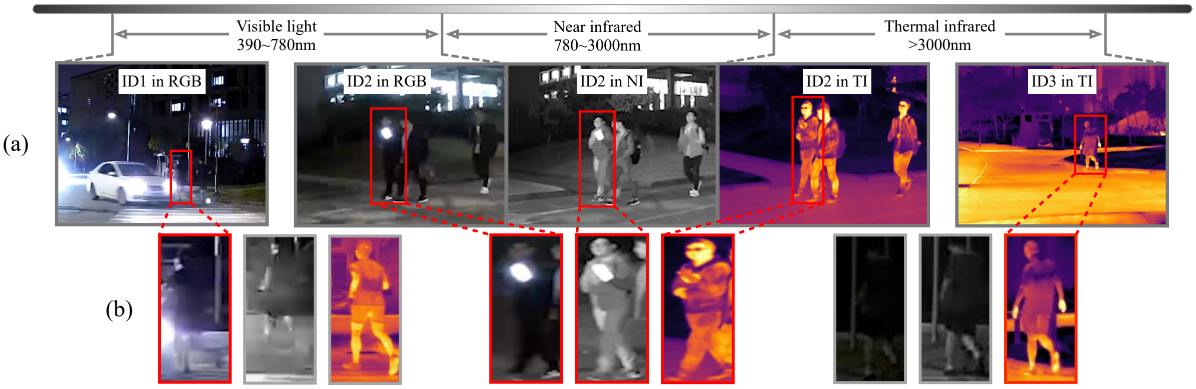

Figure 1: Demonstration of RGB, NI (Near Infrared) and TI (Thermal Infrared) multi-modality person re-identification. (a) Examples of three person identities captured in RGB, NI and TI modalities in severe conditions. (b) Corresponding multi-modality person image triples (ordered by RGB, NI and TI) of the three identities in (a).

capture the complementary local information among modalities such as the body parts and accessories of pedestrian. At last, we progressively fuse the features of three modalities to take both advantages of local-global and multi-modality information. In addition, to handle the missing modality issue when one or two modalities are not available during the test in real-world applications, we propose to transfer features of available modalities to the missing ones by introducing a cross-modality transfer module (Xu et al. 2017) to our progressive fusion network. By this way, we can still leverage the learnt multi-modal representations for robust person Re-ID.

Second, we **contribute a reasonable size dataset RGBNT201** for comprehensive evaluation in multi-modality person Re-ID. RGBNT201 contains 201 identities of person, with 4787 aligned image triplets of three modalities (RGB, NI and TI) captured by four non-overlapping cameras in real-world scenarios. It contains most of challenges in person Re-ID task, including various changes of poses, occlusion, view, illumination, resolution, background and so on. Most importantly, it contains more challenges in adverse environmental conditions, shown in the **supplementary file** due to space limitation, which launches a fair platform for Re-ID and related communities.

Finally, we **perform comprehensive evaluation on the proposed RGBNT201 dataset** with prevalent backbones on various combinations and fusion schemes across RGB, NI and TI modalities, to explore the contribution of each modality. We further evaluate PFNet against the state-of-the-art methods justify the effectiveness of the proposed progressive fusion network and provide a baseline for multi-modal Re-ID. In addition, the compatibility of missing modality scenarios further evidences the multi-modal benefit and simultaneously expands the diverse applications in real-life.

To our best knowledge, this is the first work to launch the task of RGB-NI-TI multi-modality Re-ID and the corresponding benchmark dataset. This paper makes the following contributions to person Re-ID and related applications.

- We create a new task, called multi-modality person Re-ID, by introducing the multi-modality information to handle the problem of imaging limitations of single or dual modalities in person Re-ID.
- We propose an effective progressive fusion network to achieve adequate fusion of different source data in multi-modality person Re-ID.
- We introduce a cross-modality transfer module in our framework to allow the missing modality issue by transforming the existing modality representations to the missing ones.
- We build a new benchmark dataset for multi-modality person Re-ID with 201 different persons in a wide range of viewpoints, occlusions, environmental conditions and background complexity.
- We carry out a comprehensive evaluation of different state-of-the-art approaches and in-depth experimental analysis of our progressive fusion network on the newly created benchmark dataset.

Related Work

Recent efforts on RGB infrared cross-modality and RGB Depth dual modality Re-ID provide a new solution for RGB-based single modality person Re-ID in challenging environment.

RGB-Infrared Person Re-identification

To overcome the illumination limitations in RGB-based single modality Re-ID, Wu *et al.* (2017) first propose the RGB-NI cross-modality Re-ID problem and contribute a cross-modality Re-ID dataset SYSU-MM01. Subsequently, Ye *et al.* (2018) propose a metric learning model for cross-modality Re-ID via triplet loss to supervise the training of network instead of contrastive loss. Dai *et al.* (2018) propose a generative adversarial network to learn common representations of two modalities. Feng *et al.* (2019) employ modality-specific networks to tackle with the heterogeneous matching problem. Wang *et al.* (2019) introduce a network to handle the two discrepancies separately which translates

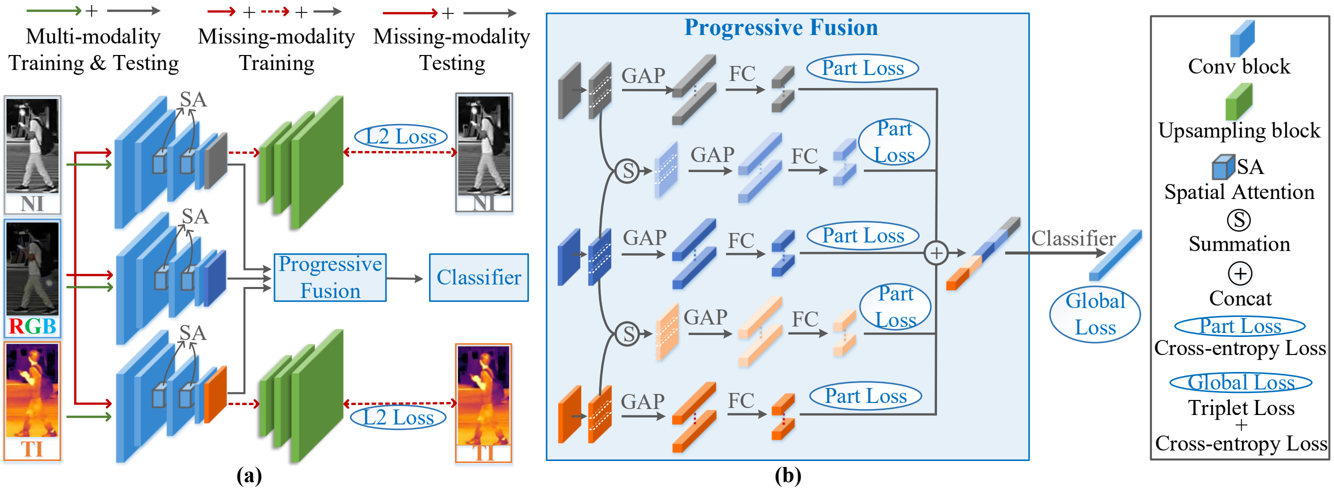

Figure 2: Framework of the proposed PFNet. (a) Multi-modality feature learning. For multi-modality situation, we develop a three-stream network to extract the features of RGB, NI and TI modality respectively as shown in the green lines. When one or two modalities are missing, taking NI and TI for instance, we use three-modal data to train the cross-modal representation transforming from the existing (RGB) modality to the missing (NI and TI) modalities by two convolutional networks and upsampling modules as shown in the red solid and dotted lines. Then we use the existing (RGB) modality data and the transformed missing modality representation learnt by the trained convolutional network during testing, as shown in the red solid lines. (b) In progressive fusion phase, we fuse the three branches into a summation branch. In particular, in the summation branch and each single branch, we divide the fusion tensor into part-level. Then we concatenate the features of all branches as the testing feature for Re-ID.

different modalities to unify the representations for images. Hao *et al.* (2019b) use sphere softmax to learn a hypersphere manifold embedding and constrain the intra-modality and cross-modality variations on this hypersphere. Li *et al.* (2020) introduce an auxiliary intermediate modality and reformulate infrared-visible dual-mode cross-modal learning as an Infrared-Intermediate-Visible three-mode learning problem to reduce the gap between RGB-NI modalities. Wang *et al.* (2020) propose to generate cross-modality paired-images and perform both global set-level and fine-grained instance-level alignments. Note that, Nguyen *et al.* (2017) propose a dual modality person dataset with paired RGB and Thermal data of each person. Since the dataset is captured by only one single camera, it is commonly used for cross-modality person Re-ID evaluation. Although the infrared data (both near infrared and the thermal infrared) can provide better visible information in adverse lighting conditions, cross-modality Re-ID confronts additional challenges due to the heterogeneous appearance in different modalities as shown in Fig. 1, which limits the performance of person Re-ID.

RGB-Depth Person Re-identification

Integrating multiple sources has been widely explored in various computer vision and multimedia tasks, including RGBT tracking (Li et al. 2017b, 2019; Zhu et al. 2019), RGBT saliency detection (Li et al. 2017a; Tu et al. 2019), etc. To make full use of the complementary among different modality resources, the depth data has been introduced to offset the RGB information. Representative RGBD Re-ID datasets include PAVIS (Barbosa et al. 2012), BIWI (Munaro et al. 2014b), and so on. Based on above datasets, Pala *et al.* (2015) combine clothing appearance with depth data for Re-ID. Xu *et al.* (2015) propose a distance metric using RGB and depth data to improve RGB-based Re-ID. Wu *et al.* (2017) propose a kernelized implicit feature transfer scheme to estimate the Eigen-depth feature from RGB images implicitly when depth device was not available. Paolanti *et al.* (2018) combine depth and RGB data with multiple k-nearest neighbor classifiers based on different distance functions. Ren *et al.* (2019) exploit a uniform and variational deep learning method for RGBD object recognition and person Re-ID. Mogelmose *et al.* (2013) propose a tri-modal (RGB, depth, thermal) person Re-ID dataset and extract color, soft body biometrics to construct features for multi-modality Re-ID. However, existing depth data is captured in the indoor condition which limits its application in more common outdoor real-world environments.

PFNet: Progressive Fusion Network

To fully leverage the complementary information in multi-modality resources, we propose a progressive fusion network (PFNet) for multi-modality person Re-ID, as shown in Fig. 2. PFNet aims to fuse from local view to global view in terms of both multi-modal cues and spatial contexts. On one hand, we employ the illumination-insensitive textures in NI and thermal-aware contexts in TI to supplement RGB modality and thus first fuse NI and TI to RGB respectively then all of them. On the other hand, we fuse the appearance feature from both local body parts and accessories level and global body shape level.

Single-Modality Feature Extraction

To obtain the high-quality representation of single modality, we first design three branches to capture the representation of person image in each modality based on ResNet50 (He et al. 2016). Along each branch, we further propose to introduce a spatial attention (SA) layer (Woo et al. 2018), to enhance the meaningful information in the feature maps.

To capture the spatial relationship of features, we employ the commonly used average-pooling to learn the content of input person image for feature aggregation. To better preserve the textures and select the discriminative feature information, we further introduce a max-pooling operation. We fuse the outputs of average-pooling and max-pooling to generate the descriptor and then forward to a convolution layer to compute the spatial attention map. The attention map A_s of spatial attention module can be formulated as:

$$A_s(F) = \sigma(C^{7\times 7}(Avg(F) + Max(F))), \tag{1}$$

where σ denotes the sigmoid function, and F indicates the features outputted from previous layers. $C^{7\times 7}$ indicates the convolution operation with the kernel size of 7×7.

Part-Level Cross-Modality Fusion

To leverage the complementary information among modalities, we design a part-level cross-modality fusion module to fuse the features from NI and TI modalities into RGB modality. Specifically, we sum the NI and TI features into RGB features respectively to boost the robustness in severe illumination or background conditions. To capture the local information of the person images, we further employ the part scheme dividing each tensor into several parts, and then employ the global average pooling (GAP) on each part in all branches. The fully connected (FC) layer is used to classify the features of each branch.

Particularly, we train each branch with b part loss functions independently. We calculate the difference between ID prediction p and the real labels, and utilize the cross entropy loss as the t-th part loss to optimize the network:

$$\begin{aligned}\mathcal{L}_{part}\left(y, N\right) &= \mathcal{L}_{cross-entropy} \\ &= -\sum_{n=1}^{N}\sum_{i=1}^{K} y^n \log(\hat{y}_i^n),\end{aligned} \tag{2}$$

where N is the number of images in a training batch, y^n is a ground-truth identity label. $\hat{y}_i^n$ is a predicted identity label for each feature in the person representation, defined as:

$$\hat{y}_i^n = \arg\max_{c\in K} \frac{exp((w_i^c)^T x_i)}{\sum_{k=1}^{K} exp((w_i^k)^T) x_i}, \tag{3}$$

where K is the number of identification labels, and w_i^k is classifier for the feature x_i and the label k.

Global Multi-Modality Fusion

To progressively learns the global representations of three modalities, we implement the multi-modality fusion module by combining all above local features in a global way. In particular, we concatenate all part-based feature representations in five streams. The globally concatenated features are fine-tuned by the global loss function, which consists of a triplet loss with hard sample mining (Hermans, Beyer, and Leibe 2017) and a cross entropy loss. For the triplet loss, we randomly select P identities with K images in each batch. The triplet loss function can be denoted as:

$$\begin{aligned}\mathcal{L}_{triplet}\left(X, Y\right) = &\sum_{i=1}^{P}\sum_{a=1}^{K}[m + \overbrace{\max_{p=1,\ldots,K} D(g_a^i, g_p^i)}^{hardest\,positive} \\ &- \overbrace{\min_{n=1,\ldots,K} D(g_a^i, g_n^i)}^{hardest\,negative}],\end{aligned} \tag{4}$$

where m denotes the margin. D indicates Euclidean distance, and g is the output feature of each samples. X denotes the concatenated features consisting of b part features from all the branches.

We concatenate the features from all five branches to optimize the global loss, which can be described as:

$$\begin{aligned}\mathcal{L}_{global}\left(X, Y\right) = &\mathcal{L}_{cross-entropy}\left(y, N\right) \\ &+\mathcal{L}_{triplet}\left(X, Y\right).\end{aligned} \tag{5}$$

At last, the loss of multi-modality training can be formulated as:

$$\begin{aligned}\mathcal{L}_{multi-modality}\left(X, Y\right) = &\mathcal{L}_{global}\left(X, Y\right) \\ &+\sum_{t=1}^{3b}\mathcal{L}_{part}^{t}\left(y_t, N\right).\end{aligned} \tag{6}$$

Cross-Modality Transfer

To handle the issue when one or two modalities are missing in practice, we further propose to learn the feature transforming from existing modality to the missing modalities to keep leveraging the multi-modality information.

As illustrated in the red solid and dotted lines in Fig. 2, taking NI and TI as missing modalities for example, we send the existing RGB data into three individual convolutional branches, one to extract RGB modality feature, and the other two to learn RGB-NI and RGB-TI cross-modal representations respectively. Then we perform the cross-modal transfer on the outputs of each missing-modality branch, alternatively between upsampling operation to enlarge the size of feature map and convolutional block to reduce the number of channels. The upsampling operation is completed via the nearest neighbor interpolation, while each convolutional block contains 1×1, 3×3 and 1×1 convolutional layers. The transferred feature map of each missing modality has 3 channels with size of 128×256, which is the same as the original missing modality image during the training. We use Mean Square Error (MSE) loss to measure the difference between the upsampling results and original missing modality images, which can be described as:

$$\mathcal{L}_{trans}^{sm} = \frac{1}{N}\sum_{i=1}^{N}(I_i^{sm} - \hat{x_i}^{sm})^2, \tag{7}$$

Dataset		Challenges									ID	Modality			
		PO	HI	BC	MB	LI	LPC	LR	VC	TC		RGB	NI	TI	Depth
Single modality	CUHK03	√	×	√	×	×	√	×	√	×	1467	13164	-	-	-
	iLIDS	√	×	√	√	×	√	√	√	×	300	42495	-	-	-
	Market1501	√	×	√	√	×	√	√	√	×	1501	32668	-	-	-
	MSMT17	√	×	√	√	×	√	√	√	×	4101	126441	-	-	-
	MARS	√	×	√	√	×	√	×	√	×	1261	1191003	-	-	-
Cross modality	SYSU-MM01	√	×	×	×	×	√	√	√	×	491	287628	15792	-	-
	RegDB	√	×	×	×	×	√	√	√	√	412	4120	-	4120	-
Multi modality	BIWI	×	×	×	×	×	√	×	√	×	78	-	-	-	-
	PAVIS	×	×	×	×	×	√	×	√	×	79	395	-	-	395
	IAS_Lab	×	×	×	×	×	√	×	√	×	11	-	-	-	-
	3DPes	×	×	×	×	×	√	×	√	×	200	1012	-	-	1012
	CAVIAR4REID	×	×	×	×	×	√	×	√	×	72	1220	-	-	1220
	RGBNT201	√	√	√	√	√	√	√	√	√	201	4787	4787	4787	-

Table 1: Comparison of RGBNT201 against prevalent Re-ID datasets.The nine columns in the middle represent nine challenges, including part occlusion (PO), high illumination (HI), background clutter (BC), motion blur (MB), low illumination(LI), large pose changing (LPC), low resolution (LR), viewpoint changing (VC) and thermal crossover (TC).

where $sm = \{NI, TI\}$ indicating the missing modality NI or TI. $\hat{x_i}^{sm}$ denotes the transferred feature of the input missing modality (NI or TI) image I_i^{sm}.

The missing-modality training loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{mis-modality} = \mathcal{L}_{multi-modality}(X, Y) \\ + \mathcal{L}_{trans}^{NI} + \mathcal{L}_{trans}^{TI}.\end{aligned} \tag{8}$$

In testing phase with only RGB data, as shown in the red solid line in Fig. 2, we can easily obtain the NI and TI modality information via the two trained cross-modal transfer branches respectively, followed by the proposed progressive fusion scheme. Other missing modality scenarios can be achieved in the same manner.

RGBNT201: Multi-modality Person Re-ID Dataset

To evaluate the proposed PFNet on multi-modality person Re-ID, we propose a multi-modality dataset, RGBNT201, to integrate the complementary information among different modality resources.

Data Acquisition

RGBNT201 dataset is collected on campus in four non-overlapping views, each of which is captured by a triplicated cameras to simultaneously record RGB, NI and TI data. Unlike the most RGB person Re-ID datasets, which are captured only in daytime with favourable lighting, we further capture large number of challenging images in harsh lighting conditions such as darkness in the night, or low visibility weather such as smog and fog. Specifically, we utilize the paired RGB-NI cameras HIKVISION with resolution of 700 × 580 in 15 fps frame rate to capture the RGB and NI modality images. The TI images are simultaneously captured by the FLIR T610 with resolution of 640 × 480 in frame rate of 20 fps. We first implement the frame alignment and then the pixel alignment to produce the multi-modality records.

We record all the three modality data in videos, which span about four months covering early spring to summer providing diverse clothing fashions. The original data contributes more than 9000 seconds, and then we select about 40000 image triplet records from the videos. The bounding boxes are manually annotated with resolution of 256 × 128 for each modality.

Dataset Description

RGBNT201 dataset contains 201 identities in four different viewpoints with diverse illumination conditions and background complexity. For efficient evaluation, we automatically select the person images in every 5-10 adjacent images, followed by manual checking to avoid data redundancy. Each record consists of at least 20 nonadjacent images triplets in the fashion of three-modality, captured with different poses, which forms 4787 images in each modality for the experimental evaluation. We select 141 identities for training, 30 identities for validation, while the remaining 30 identities for testing. In the testing stage, we use the entire testing set as gallery set, while randomly selecting 10 records of each identity as probe.

Comparing with existing prevalent Re-ID datasets as shown in Table 1, RGBNT201 has the following major advantages.

- It contains a large number of person images aligned in three modalities captured by four non-overlapping camera views. To our best of knowledge, RGBNT201 is the largest multi-modality person Re-ID dataset with the most challenge scenarios and modalities comparing with dual-modality person Re-ID datasets including BIWI (Munaro et al. 2014b), PAVIS (Barbosa et al. 2012), IAS_Lab (Munaro et al. 2014a) and CAVIAR4REID (Cheng et al. 2011).
- It includes person images captured in different weathers and illuminations, and conforms to the reality of surveillance systems. Especially the ubiquitous low illumination (LI) and the high illumination (HI) challenges in real-life scenarios, have been significantly ignored in existing single-, cross- and dual-modality datasets.

Modalities	OSNet				ResNet50				MobilenetV2			
	mAP	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10
RGB	16.96	34.38	70.68	86.44	13.10	25.16	59.65	78.06	12.41	26.72	64.63	82.28
NI	13.90	27.91	62.86	80.43	12.10	21.74	57.56	75.84	15.29	25.40	60.14	78.74
TI	15.38	26.30	60.24	77.31	15.19	20.97	49.17	68.13	13.09	20.26	54.12	72.34
(RGB-NI)_cat	19.47	29.96	67.06	83.95	14.69	32.89	74.97	89.29	17.42	31.62	66.24	83.10
(RGB-TI)_cat	21.90	32.91	63.88	82.09	16.53	35.15	70.20	86.51	17.67	31.41	70.37	85.11
(RGB-NI-TI)_cat	22.13	38.40	72.57	86.47	24.77	45.85	80.06	91.14	18.13	32.84	71.08	87.12
(RGB-NI)_**PFNet**	22.18	33.29	69.51	88.35	17.95	37.02	80.19	91.76	25.24	35.22	69.59	85.63
(RGB-TI)_**PFNet**	23.50	34.29	69.16	84.47	30.23	51.85	83.48	92.56	20.37	34.42	72.44	89.35
(RGB-NI-TI)_PFNet	**33.65**	**48.94**	**81.27**	**93.36**	**31.76**	**54.59**	**87.06**	**95.49**	**29.47**	**44.08**	**79.48**	**92.58**

Table 2: Experimental comparison of the different modalities and different backbones on RGBNT201 (in %).

Methods	OSNet				ResNet50				MobilenetV2			
	mAP	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10
Baseline	22.13	38.40	72.57	86.47	24.77	45.85	80.06	91.14	18.13	32.84	71.08	87.12
+SA	26.22	47.63	80.12	91.25	26.13	47.32	82.40	92.85	23.74	42.66	77.02	90.10
+CMF	28.12	48.25	81.04	92.22	25.78	49.09	84.55	93.07	26.54	43.45	78.19	91.36
+CMF+SA (PFNet)	**33.65**	**48.94**	**81.27**	**93.36**	**31.76**	**54.59**	**87.06**	**95.49**	**29.47**	**44.08**	**79.48**	**92.58**

Table 3: Ablation study on spatial attention (SA) and cross-modality fusion (CMF) on RGBNT201 with different backbones. Baseline denotes directly concating the three modality features without SA or CMF (in %).

- It involves not only the challenges of RGB-based and RGB-NI cross-modality person Re-ID problems, but also the additional challenges introduced by multiple modalities. Therefore it offers the complementary information to boost the conventional RGB-based and RGB-NI cross-modality Re-ID tasks, and meanwhile launches additional challenges to multi-modality Re-ID research.

Experiments

Implementation Details

The implementation platform is Pytorch with a NVIDIA GTX 1080Ti GPU. We use a ResNet50 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009) as our CNN backbone. The initial learning rate is set as 1×10^{-3}. Consequently, we increase the number of train iterations due to the small learning rate. The number of mini-batches is 8. In the training phase, we use Stochastic Gradient Descent (SGD) with the momentum of 0.9 and weight decay of 0.0005 to fine-tune the network.

Impact of Different Backbones

To verify the contribution of multi-modality information and evaluate the effectiveness of the proposed PFNet for multi-modality Re-ID, we evaluate our method with various combination of modalities on three different backbones, including OSNet (Zhou et al. 2019), ResNet50 (He et al. 2016), and MobilenetV2 (Sandler et al. 2018). As reported in Table 2, i) due to the complex challenges in RGBNT201, none of the three backbones works satisfactory on any modality, which launches a challenging scenario for person Re-ID. ii) More modality scenarios improve the performance of the less modality ones either by simply concatenating ($\_cat$) or proposed progressive fusing (_PFNet), which verifies the contribution of the complementary information among the three modalities. iii) PFNet outperforms concatenating in all the scenarios on all the metrics, which validates the effectiveness of the proposed PFNet while fusing multi-modality information. For the sake of balance between mAP and rank-1 scores, we use ResNet50 as the default backbone for PFNet in the following experiments.

Ablation Studies

To verify the effective contribution of the components in our model, we implement the ablation study on the spatial attention (SA) module and the cross-modality fusing (CMF) scheme in PFNet on RGBNT201, as reported in Table 3. Notice that, both cross-modality fusing scheme and spatial attention module enhance the results of baseline, which demonstrates the contribution of each module. By simultaneously enforcing both modules, our method achieves the best performance.

Comparison with State-of-the-art Methods

Since it is the first work of multi-modality Re-ID, we extend five state-of-the-art single modality Re-ID methods, ABD-Net (Chen et al. 2019), OIM Loss (Xiao et al. 2017), MLFN (Chang, Hospedales, and Xiang 2018), PCB (Sun et al. 2018), and ABD-Net (Chen et al. 2019) by concating the deep features from each of the three modalities for comparison.

As reported in Table 4, ABD-Net integrates the channel aggregation and position awareness attention mechanisms, while MuDeep and PCB consider the multi-scale or part-level respectively for person Re-ID. They achieve remarkable performance while handling multi-modality scenario. However, they are still significantly inferior to the proposed PFNet. Note that both mAP and ranking scores dramatically decline in OIM Loss and MLFN. The main reason is that, MLFN emphasizes on the semantic supervision on

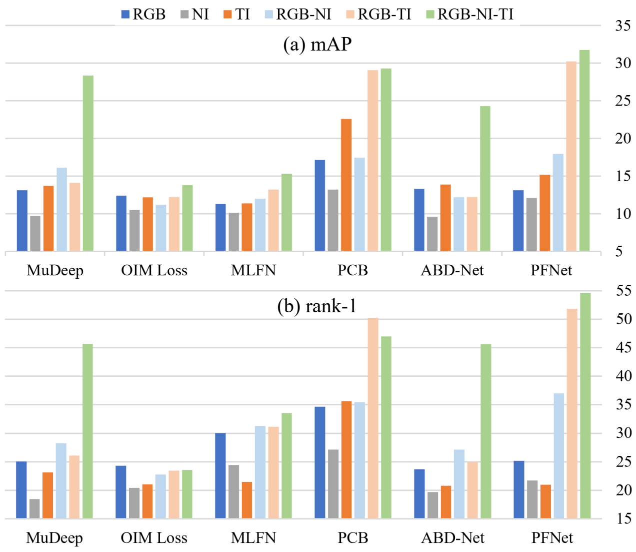

Figure 3: Evaluation results on diverse modality combinations of the proposed PFNet comparing to the state-of-the-art methods (in %).

	RGBNT201			
	mAP	rank-1	rank-5	rank-10
MuDeep	28.34	45.65	78.11	90.50
OIM Loss	13.80	23.58	57.14	74.85
MLFN	15.32	33.53	67.34	83.77
PCB	29.30	46.96	83.12	93.66
ABD-Net	24.30	45.61	80.12	91.29
Baseline	24.77	45.85	80.06	91.14
PFNet	**31.76**	**54.59**	**87.06**	**95.49**

Table 4: Experimental results of PFNet on RGBNT201 comparing with state-of-art methods while handling multi-modality Re-ID (in %).

visual factors, which cannot be well deployed without additional semantic annotations especially on infrared data. Our PFNet significantly beats the prevalent methods, which promises the effectiveness of the proposed PFNet while handling multi-modality Re-ID task.

Evaluation on Cross-Modality Scenario

Followed by the data splitting protocol in cross-modality dataset RegDB (Nguyen et al. 2017), we reconstruct our RGBNT201 dataset for six cross-modal Re-ID scenarios between each two modalities, and evaluate two competitive state-of-the-art methods TSLFN+HC (Zhu et al. 2020) and DDAG (Ye et al. 2020).

As reported in Table 5, due to the heterogeneous issue and huge challenges in RGBNT201, the two competitive cross-modality methods result in stumbling performance, comparing with the corresponding multi-modality results as shown in Table 2. This verifies the significance of the new multi-modality Re-ID problem and the effectiveness of the proposed PFNet.

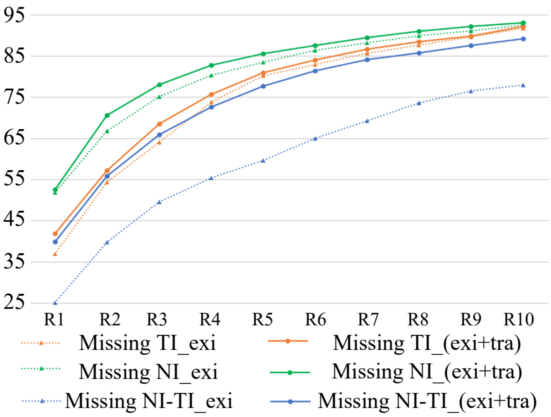

Figure 4: Experimental results of PFNet with diverse missing modality issues during testing (in %). When one or two modalities are absent during the test, “$\_exi$” directly trains and tests on the existing/available modalities. While “$\_(exi + tra)$” introduces the cross-modality transfer in the training thus can use both the existing and the transferred missing modality features for the testing.

Evaluation on Modality Missing Issue

To capture the multi-modal complementary information when some modalities are missing during the test, we evaluate our method with diverse modality absences as explained in Fig. 2. As reported in Fig. 4, comparing to directly training on existing modality data, our method can better capture the multi-modality complementary and thus boost the performance in diverse missing modality scenarios, especially in while both NI and TI modalities are missing, which verifies the effectiveness of the proposed PFNet while handling the missing modality issue.

Evaluation on Diverse Modality Scenarios

To evaluate the effectiveness of the proposed PFNet while handling multi-modality Re-ID task, we further compare

Method		RGBNT201			
		mAP	Rank-1	Rank-5	Rank-10
TSLFN+HC	RGB to TI	15.69	11.28	25.82	37.09
	TI to RGB	16.58	13.04	33.82	45.65
	RGB to NI	22.57	26.41	43.92	51.63
	NI to RGB	22.00	18.36	41.06	57.97
	TI to NI	16.29	14.01	29.95	43.48
	NI to TI	16.98	11.87	29.97	39.17
DDAG	RGB to TI	18.07	18.40	29.38	36.20
	TI to RGB	17.03	15.46	28.99	40.34
	RGB to NI	29.47	35.01	51.63	65.88
	NI to RGB	30.64	34.54	56.76	68.36
	TI to NI	12.81	11.59	25.12	34.78
	NI to TI	12.27	9.79	22.55	32.64

Table 5: Results of state-of-art cross-modality methods on RGBNT201 (in %).

PFNet with the state-of-the-art methods on diverse modality combinations, including single modality, two-modality and three modality scenarios, as shown in Fig. 3. i) Generally, speaking, integrating NI (RGB-NI) or TI (RGB-TI) or both (RGB-NI-TI) improves the untenable performance in RGB modality, which verifies the contribution of our multi-modality sources. ii) Some methods declines after introducing the infrared modalities, such as OIM Loss in both mAP and rank-1 scores and ABD-Net in mAP in RGB-NI and RGB-TI comparing to th single modality scenarios. And the rank-1 of PCB in RGB-NI-TI scenario comparing to RGB-TI scenario. By contrast, our PFNet consistently improves both mAP and rank-1 scores by a large margin in both two and three modality scenarios. This reveals the effectiveness of the proposed fusing scheme for multi-modality Re-ID.

Conclusion

To our best knowledge, this is the first work to launch RGB-NI-TI multi-modality person Re-ID problem. We have proposed a novel feature aggregation method, PFNet, to progressively fuse the multi-modality information for person Re-ID, which can better leverage the complementary information in multi-spectral resources for real-life applications. Meanwhile, we have contributed a new benchmark RGB-NI-TI dataset named RGBNT201 for multi-modality person Re-ID at the first time. We further explore the multi-modal complementarity by simply adjusting the training and testing schemes for the missing modality issue. Comprehensive experimental evaluation on proposed RGBNT201 demonstrate the promising performance of the proposed PFNet while handling multi-modality Re-ID task. At last, extensive results evidence that the fusing scheme significantly affect the performance of multi-modality Re-ID task, which turns to be the key research emphasis for our future plan.

Acknowledgments

This research is supported in part by the Major Project for New Generation of AI under Grant (2020AAA0140002), the National Natural Science Foundation of China (61976002, 61976003 and 61860206004), the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2019A0033), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (201900046).

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.

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.

Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; and Wang, Z. 2019. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 8351–8361.

Cheng, D. S.; Cristani, M.; Stoppa, M.; Bazzani, L.; and Murino, V. 2011. Custom pictorial structures for re-identification. *Proceedings of British Machine Vision Conference* 1(2): 6.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-Modality Person Re-Identification with Generative Adversarial Training. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 677–683.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Feng, Z.; Lai, J.; and Xie, X. 2019. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *IEEE Transactions on Image Processing* 29: 579–590.

Hao, Y.; Wang, N.; Gao, X.; Li, J.; and Wang, X. 2019a. Dual-alignment Feature Embedding for Cross-modality Person Re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia*, 57–65.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019b. HSME: Hypersphere Manifold Embedding for Visible Thermal Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8385–8392.

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.

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

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition* 96: 106977.

Li, C.; Wang, G.; Ma, Y.; Zheng, A.; Luo, B.; and Tang, J. 2017a. A unified RGB-T saliency detection benchmark: dataset, baselines, analysis and a novel approach. *arXiv preprint arXiv:1701.02829* .

Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; and Tang, J. 2017b. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of the 25th ACM International Conference on Multimedia*, 1856–1864.

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

Mogelmose, A.; Bahnsen, C.; Moeslund, T.; Clapes, A.; and Escalera, S. 2013. Tri-modal person re-identification with rgb, depth and thermal features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Workshop)*, 301–307.

Munaro, M.; Basso, A.; Fossati, A.; Van Gool, L.; and Menegatti, E. 2014a. 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.

Munaro, M.; Fossati, A.; Basso, A.; Menegatti, E.; and Van Gool, L. 2014b. One-shot person re-identification with a consumer depth camera. *Person Re-Identification* 161–181.

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

Pala, F.; Satta, R.; Fumera, G.; and Roli, F. 2015. Multimodal person reidentification using RGB-D cameras. *IEEE Transactions on Circuits and Systems for Video Technology* 26(4): 788–799.

Paolanti, M.; Romeo, L.; Liciotti, D.; Pietrini, R.; Cenci, A.; Frontoni, E.; and Zingaretti, P. 2018. Person Re-Identification with RGB-D Camera in Top-View Configuration through Multiple Nearest Neighbor Classifiers and Neighborhood Component Features Selection. *Sensors* 18(10): 3471.

Ren, L.; Lu, J.; Feng, J.; and Zhou, J. 2019. Uniform and Variational Deep Learning for RGB-D Object Recognition and Person Re-Identification. *IEEE Transactions on Image Processing* 28(10): 4970–4983.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

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 the European Conference on Computer Vision*, 480–496.

Tu, Z.; Xia, T.; Li, C.; Lu, Y.; and Tang, J. 2019. M3S-NIR: Multi-modal Multi-scale Noise-Insensitive Ranking for RGB-T Saliency Detection. In *Proceedings of IEEE Conference on Multimedia Information Processing and Retrieval*, 141–146.

Wang, G.-A.; Yang, T. Z.; Cheng, J.; Chang, J.; Liang, X.; Hou, Z.; et al. 2020. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12144–12151.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 618–626.

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

Wu, A.; Zheng, W.-S.; and Lai, J.-H. 2017. Robust depth-based person re-identification. *IEEE Transactions on Image Processing* 26(6): 2588–2603.

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.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3415–3424.

Xu, D.; Ouyang, W.; Ricci, E.; Wang, X.; and Sebe, N. 2017. Learning Cross-Modal Deep Representations for Robust Pedestrian Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4236–4244.

Xu, X.; Li, W.; and Xu, D. 2015. Distance metric learning using privileged information for face verification and person re-identification. *IEEE Transactions on Neural Networks and learning systems* 26(12): 3150–3162.

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.

Ye, M.; Wang, Z.; Lan, X.; and Yuen, P. C. 2018. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1092–1099.

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

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.