Progressive Attribute Embedding for Accurate Cross-modality Person Re-ID

Aihua Zheng
IMIS Lab of Anhui Province, School of Artificial Intelligence, Anhui University
Hefei, China
ahzheng214@foxmail.com

Peng Pan
Anhui Provincial Key Lab of MCC, School of Computer Science and Technology, Anhui University
Hefei, China
anlepanp@foxmail.com

Hongchao Li
Anhui Provincial Key Lab of MCC, School of Computer Science and Technology, Anhui University
Hefei, China
lhc950304@foxmail.com

Chenglong Li*
IMIS Lab of Anhui Province, School of Artificial Intelligence, Anhui University
Hefei, China
lcl1314@foxmail.com

Bin Luo
Anhui Provincial Key Lab of MCC, School of Computer Science and Technology, Anhui University
Hefei, China
ahu_lb@163.com

Chang Tan
iFLYTEK Co., Ltd.
Hefei, China
changtan2@iflytek.com

Ruoran Jia
iFLYTEK Co., Ltd.
Hefei, China
jiaruoran@foxmail.com

ABSTRACT

Attributes are important information to bridge the appearance gap across modalities, but have not been well explored in cross-modality person ReID. This paper proposes a progressive attribute embedding module (PAE) to effectively fuse the fine-grained semantic attribute information and the global structural visual information. Through a novel cascade way, we use attribute information to learn the relationship between the person images in different modalities, which significantly relieves the modality heterogeneity. Meanwhile, by embedding attribute information to guide more discriminative image feature generation, it simultaneously reduces the inter-class similarity and the intra-class discrepancy. In addition, we propose an attribute-based auxiliary learning strategy (AAL) to supervise the network to learn modality-invariant and identity-specific local features by joint attribute and identity classification losses. The PAE and AAL are jointly optimized in an end-to-end framework, namely, progressive attribute embedding network (PAENet). One can plug PAE and AAL into current mainstream models, as we implement them in five cross-modality person ReID frameworks to further boost the performance. Extensive experiments on public datasets demonstrate the effectiveness of the proposed method against the state-of-the-art cross-modality person ReID methods.

*Corresponding authors: Chenglong Li.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548336

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**.

KEYWORDS

Person Re-identification, Cross-modality, Attribute Embedding

ACM Reference Format:
Aihua Zheng, Peng Pan, Hongchao Li, Chenglong Li, Bin Luo, Chang Tan, Ruoran Jia. 2022. Progressive Attribute Embedding for Accurate Cross-modality Person Re-ID. In *Proceedings of the 30th ACM Inter-national Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161. 3548336

1 INTRODUCTION

Cross-modality visible-infrared person ReID (RGB-IR ReID) [31] aims to match images of people captured by visible and infrared cameras. In addition to the common challenges such as view changes, illumination, and background clutter, it brings extra challenges to match the modality heterogeneous data of the same person. As shown in the oval in Fig. 1 (a), the ubiquitous heterogeneity results in a large distance between the feature distributions of the same person in two modalities. Meanwhile, due to the influence of light or occlusions, RGB-IR ReID still suffers from large inter-class similarities (the blended distributions in green, black and pink), and intra-class discrepancy (the scattered distributions in pink and black), as shown in the black box in Fig. 1 (a),

Existing methods can be divided into two main categories: 1) GAN-based methods [24, 28, 35], which try to bridge the modal differences by generating corresponding modality images. However, the inherent differences between the modalities may destroy the local structure and introduce unavoidable noise during the generation

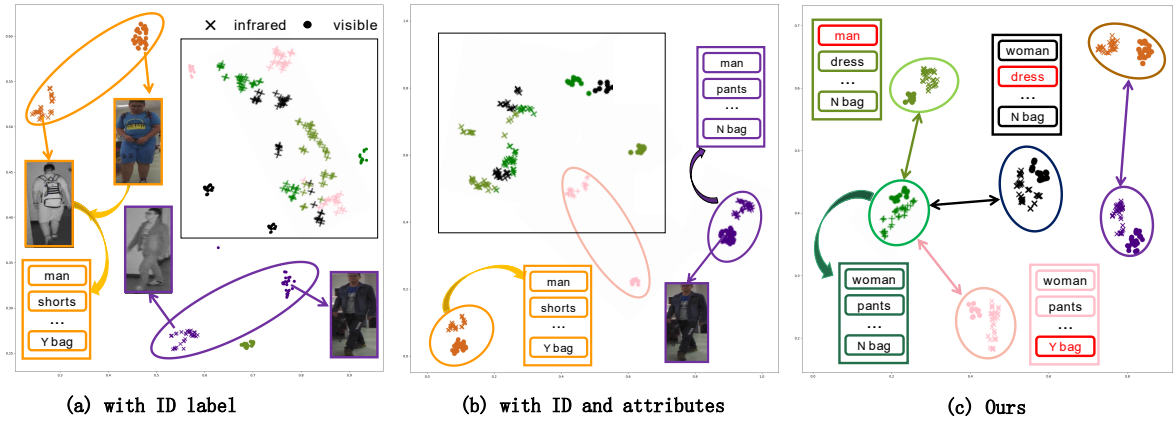


Figure 1: Feature distribution of six IDs selected from SYSU-MM01 [31] dataset randomly through three different methods. The dot and cross denote the features in the visible and infrared modalities respectively and different colors represent different identities. (a) baseline [12]: training using only identity labels. (b) baseline+ATTR [42]: using identity labels and attribute labels, and training with additional attribute loss. (c) baseline+Ours: training with attributes embedded by our proposed module.

process. 2) Modality-shared feature learning methods [5, 18, 34] devote to projecting heterogeneous modal features into the unified space to reduce the cross-modality difference. Nevertheless, both categories tend to learn global image-level information for modality feature representation. The large cross-modality heterogeneity significantly hinders the discriminative feature representation. Furthermore, the large intra-class discrepancy and inter-class similarity in the visual appearance across the non-overlapping cameras also bring huge challenges to cross-modality ReID.

As auxiliary information, the attributes have also been proved as a kind of effective information to boost the vision tasks, including person search [1], vehicle ReID [22, 33] and face recognition [8]. In cross-modality ReID, Zhang *et al.* [42] point out that some color-independent person attributes are unchanged across modalities. They propose an end-to-end network that uses additional attribute labels as auxiliary information to bridge the cross-modality gap. By predicting person attributes through additional attribute classification branches, it can learn modality variables and identity-specific local features under the joint supervision of attribute and identity classification losses. However, there are still two major problems. 1) It only fine-tunes the network jointly with the attribute loss, which fails to take into account the internal connections between attributes and images, as well as the potential interactions among attributes. 2) Unlike global image identity information, attributes are semantically fine-grained information, which is very easy to lose during network training. As a result, the challenges of intra-class discrepancy and inter-class similarity (inside the rectangular box in Fig. 1 (b)) are still not well addressed.

To address the above problems, we propose a novel Progressive Attribute Embedding Network (PAENet) to comprehensively integrate attributes with image information for cross-modality ReID. Specifically, PAE includes three levels of embeddings. The first-level embedding relies on the cross-attention scheme, which can learn the complementary information between different modalities by the interaction among key, value and query. Through this embedding, the gap in semantic space between images and attributes is effectively reduced. The regions associated with the attributes can adaptively provide discriminative details to achieve fine-grained matching. Therefore, we design the second-level embedding by the attribute-guided attention to dynamically selects attribute-related appearance regions within each modality for fine-grained matching. Finally, a certain area may contain multiple attributes, which have a different impact on identification. Therefore, the three-level embedding is used to collaboratively employ the connections between different attributes by using channel attention as an element-wise gating function to select key attributes. By cascading these embeddings, PAENet can achieve the effective integration of attributes and images gradually. Meanwhile, it can learn the relationship between different attributes collaboratively.

Guided by the attributes, the network generates more discriminative modality-aware features and dynamically mines the modality-invariant fine-grained information, which effectively reduces modality differences, shown as the brown and purple IDs in Fig. 1 (c). By using attributes to guide feature generation, the network can further relieve the inter-class similarity (green and black IDs) and intra-class discrepancy (black and pink IDs).

In addition to fusing attributes and images in PAE, we also propose an attribute-based auxiliary learning scheme to further boost the discriminative representation of images guided by attributes. In particular, we design an attribute classification module in the training stage to guide image representation learning by the attributes. It is worth noting that AAL is only used as an auxiliary learning strategy to assist in obtaining modality-invariant feature representation, we remove it in the testing phase.

The main contributions of this paper include:

- We propose a novel progressive attribute embedding method to effectively employ the internal connections between attributes and images, as well as the potential interactions

among attributes for cross-modality ReID performance boosting. In addition, the fine-grained information of attributes are well leveraged in network training.
- We propose an effective attribute-based auxiliary learning scheme to further boost the discriminative representation of images guided by attributes while maintaining the efficiency.
- The proposed progressive attribute embedding scheme is generic and easily integrated with existing ReID frameworks as we implemented in the experiments, and the results validate the superiority of our scheme against the state-of-the-art methods.

2 RELATED WORK

2.1 RGB-IR Cross-modality Person ReID

In RGB-IR ReID, Wu *et al.* [31] first contribute a large benchmark dataset (SYSU-MM01) and propose a one-stream zero-padding network for RGB-IR image matching. Current researches mainly devote the shared feature learning methods to dealing with modality differences. Ye *et al.* [38] design a new baseline for cross-modality ReID, which uses the non-local attention block to achieve competitive performance. Lu *et al.* [18] propose a novel cross-modality shared-specific feature transfer algorithm to explore the potential of both the modality-shared information and the modality-specific characteristics. Meanwhile, other works [12, 16, 29, 36] have investigated effective loss functions to handle the modality gap. However, most of the above methods focus on improving the intra-class cross-modality similarity, while ignoring the enlarging inter-class discrepancy of features.

Meanwhile, some methods explore cross-modality representation learning from the perspective of generative adversarial training by GAN technology. The cmGAN[5] is the first effort in GAN-based cross-modality person ReID. Dai *et al.* [5] propose an end-to-end generated network, which consists of a generator to extract features from two different modalities and a discriminator to distinguish the modality features. Wang *et al.* [25] propose to generate cross-modality paired images and perform both global set-level and fine-grained instance-level alignments, which can reduce the modality variation well. Although these methods generate corresponding cross-modal images or features to reduce the modal heterogeneity, the generated are unreliable with inevitable noise. At the same time, the infrared images lack the rich color texture information in the visible image [29], therefore it is not reasonable to directly convert the cross-modal images/features.

2.2 Attributes for ReID

With the gradual progress of research, attributes (such as gender, age and clothing) have been noticed as a kind of effective auxiliary information for ReID. Attributes can provide additional annotations and have been introduced into person ReID. Compared with individual personal identification, attributes can provide a higher level of semantic identification information. Liu *et al.* [17] have labelled the two largest datasets, i.e., Market-1501 and DukeMTMC-reID, with attribute labels, and then simultaneously learned a ReID model to predict the semantic attributes of the pedestrian. Recently, deep learning methods [14, 21] use the attributes to help the supervision of joint training, so as to increase the distinction of identity

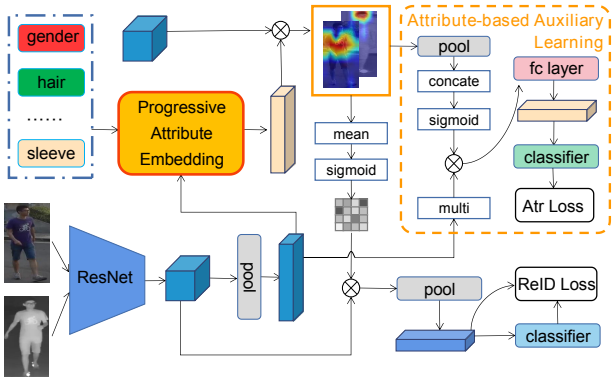


Figure 2: Framework of PAENet– It involves novel key components: the progressive attribute embedding (PAE) and attribute-based auxiliary learning (AAL).

features and strengthen the relevance of image pairs. Zhang *et al.* [40] use the feature aggregation strategy to make full use of attribute information. To reduce the reliance on attribute annotation, unsupervised methods [20, 26, 28] have been proposed.

Although both attribute recognition and ReID are classification tasks, the former favors fine-grained recognition while the latter belongs to global visual information recognition. However, most of the methods mentioned above ignore this discrepancy between these two tasks as well as the internal relationship among attributes.

3 METHOD

Preliminary. The overview of Progressive Attribute Embedding Net (PAENet) is illustrated in Fig. 2. The input images, including the visible and infrared images, are first fed into the two-stream network to extra the image features. Then we propose the progressive attribute embedding (PAE) to fuse attributes and image features, promoting the learning of discriminative modal-irrelevant features and assigning more accurate local features. Meanwhile, to avoid the misidentification of identity caused by excessive interference of attribute information, the attribute-based auxiliary learning (AAL) is proposed to assist in generating better attribute feature representation. The two components are integrated into a unified framework and can facilitate each other.

3.1 Baseline

The conventional two-stream network is used as the backbone to extra features. Specifically, we denote the modality-specific features network as $\boldsymbol{conv}^m, m \in [v, t]$, which separately extract visible and infrared features. The feature embedding network $\boldsymbol{conv}^s$ projects modality-specific person features into the shared common feature space. Given a visible image $\boldsymbol{I}^v \in R^{3\times H\times W}$ and an infrared image $\boldsymbol{I}^t \in R^{3\times H\times W}$, the learned 3D person features $\boldsymbol{F}^v$ and $\boldsymbol{F}^t$ in the common space can be represented as,

$$\boldsymbol{F}^m = \begin{cases} \boldsymbol{conv}^s\left(\boldsymbol{conv}^v\left(\boldsymbol{I}^v\right)\right), \\ \boldsymbol{conv}^s\left(\boldsymbol{conv}^t\left(\boldsymbol{I}^t\right)\right), \end{cases} \tag{1}$$

where $\boldsymbol{F^m} \in R^{C\times H\times W}$, C, H and W are the channel number, image height and width, respectively. We adopt ResNet-50 as the backbone, in which each branch contains a pre-trained model, which inherits the architecture of ResNet-50 before the global average pooling layer. At the same time, the last down-sampling operation is removed to enrich the granularity of feature. Then, we use Gem-Pooling [38] to obtain fine-grained features ($\boldsymbol{f}^m$).

Following the state-of-the-art methods [3, 12, 16], we use the pooled features for subsequent recognition tasks. We use the prevalent MMD [12] as our baseline, which utilizes identity loss $\mathcal{L}_{id}$, the proposed the Maximum Mean Discrepancy loss $\mathcal{L}_{MMD}$ and hetero-center triplet loss $\mathcal{L}_{Hc-Tri}$ to constrain the network, the baseline learning loss is denoted as $\mathcal{L}_b$,

$$\mathcal{L}_b = \mathcal{L}_{id} + \mathcal{L}_{\text{MMD}} + \mathcal{L}_{Hc-Tri}. \tag{2}$$

.

3.2 Progressive Attribute Embedding

To address the differences between images and attributes, we propose the PAE module that blends these two types of information in a progressive embedding way to bridge the cross-modality gap. To be brief, the first embedding aims to reduce the difference in semantic space between images and attributes; the second embedding dynamically selects attribute-related appearance regions through attribute-guided attention; the third is used to collaboratively explore the connections between different attributes and the rich contextual information.

We use one-hot vector $\boldsymbol{A}^m$ to represent the attributes, $\boldsymbol{A}^m = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}, \boldsymbol{a}_i \in [0, 1]$, where n represents the number of attributes, and $m \in [v, t]$ representing the RGB/IR modality. According to the given attributes ($\boldsymbol{A}^m$), we first project them into a 2048-dimensional vector ($\boldsymbol{f}_a^m$). Given an image feature ($\boldsymbol{f}^m$) and attribute embedding feature ($\boldsymbol{f}_a^m$), we aim to learn an attribute enhanced feature to learn both global visual information and local detail information. The overview of the progressive attribute embedding is illustrated in Fig. 3, which mainly consists of three progressive embeddings.

Embedding-I. Attributes are fine-grained semantic information, while images belong to global structural visual information. Therefore, there is a big gap between attributes and images. We believe that the cross-attention mechanism [6] can discover the hidden relationships between different information by using a simple but powerful reasoning mechanism. Herein, the first embedding aims to extract useful information from images and attributes through the powerful and robust cross-attention mechanism to mine the critical information.

Specifically, we first use linear mapping to align the dimensions of attribute and image features, which are sequentially sent to the cross attention module. In order to integrate attributes and images more effectively, the image feature ($\boldsymbol{f}^m$) serves as query ($\boldsymbol{Q}$). Meanwhile, the image feature ($\boldsymbol{f}^m$) and the attribute feature ($\boldsymbol{f}_a^m$) perform the concatenating operation, and subsequently as key ($\boldsymbol{K}$) and value ($\boldsymbol{V}$), then use the following expressions to realize the fusion operation,

$$\begin{cases} \boldsymbol{f}_{ca}^m = \text{Norm}(Att(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + \boldsymbol{Q}), \\ \text{Att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{V}, \end{cases} \tag{3}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ are query, key and value, respectively, d is the embedding dimension. The cross-attention is based on the trainable associative relation between query and key. Followed by two residual connections, a normalization layer and a simple feed-forward network, finally, the network can learn clear structural information and subtle pixel-level features ($\boldsymbol{f}_{e1}^m$).

$$\begin{cases} z_1 = \text{Norm}(\boldsymbol{f}_{ca}^m + \boldsymbol{f}^m), \\ \boldsymbol{f}_{e1}^m = \text{Norm}\left(z_1 + FFN\left(z_1\right)\right). \end{cases} \tag{4}$$

Embedding-II. The Embedding-I uses long-term dependence instead of the local spatial method to fuse attribute and image features. However, it cannot use attribute features to guide the transfer of image features, and attributes are detailed information that is very easy to lose as the network is trained. To this end, we cascade the attribute-guided attention mechanism to help deal with the lack of attribute information. We argue that different attributes correspond to different positions on the image. For attribute features, we only need to focus on specific relevant areas. For this reason, to perceive attribute-related regions and deliver the most discriminative detail information adaptively, we subsequently introduce a second embedding by using a spatial attention mechanism with the guide of specific attributes. Specifically, we first process the attributes through the linear layer and spatial duplication (1*1 convolution layer and reshape operation). And then, we use a 1*1 convolution layer on the fused features ($\boldsymbol{f}_{e1}^m$) to unify its dimension and size with the attribute features after spatial duplication operation. For the convenience of representation, the processing of attributes and embedding features are denoted as p_1 and p_2, respectively. After feature mapping, the attention weight is obtained,

$$\begin{cases} \boldsymbol{f}_{ST}^m = \left(\text{conv}\left(p_1\left(\boldsymbol{f}_{e1}^m\right) \odot p_2(\boldsymbol{f}_a^m)\right)\right), \\ \boldsymbol{f}_{e2}^m = \boldsymbol{f}_{e1}^m \odot softmax(\boldsymbol{f}_{ST}^m), \end{cases} \tag{5}$$

where $\odot$ represents element multiplication, conv is 1*1 convolution layer, and $softmax$ is used to obtain adaptive attention weights, which are multiplied with image feature to obtain spatial attention-guided feature ($\boldsymbol{f}_{e2}^m$). After this embedding, the model adaptively focuses on specific areas of the image.

Embedding-III. Although the second embedding can adaptively focus on specific image regions, a particular area may be associated with multiple attributes. Moreover, some attributes may have a negative effect on the recognition performance, while others are positive. To distinguish the importance of different attributes, we further propose the third embedding by using channel attention as an element-wise gating function, which can choose the positive effective one on the network performance among the different attributes.

Concretely, we first employ a linear layer to embed attributes ($\boldsymbol{A}^m$) into an embedding vector. We concatenate the attribute embedding vector ($\boldsymbol{f}_a^m$) and the image features ($\boldsymbol{f}_{e2}^m$) after the previous two embeddings, followed by n fully connected layers and the $sigmoid$ function to obtain the channel attention weights,

$$\boldsymbol{w}^m = sigmoid\left(fc_i^m\left[\boldsymbol{f}_a^m, \boldsymbol{f}_{e2}^m\right]\right), \tag{6}$$

where $i \in [1, 2, \cdots, n]$, and n is the number of attributions. Then we multiply the weights $\boldsymbol{w}^m$ and feature map $\boldsymbol{F}^m$, and finally obtain n different feature,

$$\boldsymbol{F}_e^m = \left[\boldsymbol{F}_1^m, \boldsymbol{F}_2^m, \cdots, \boldsymbol{F}_n^m\right]. \tag{7}$$

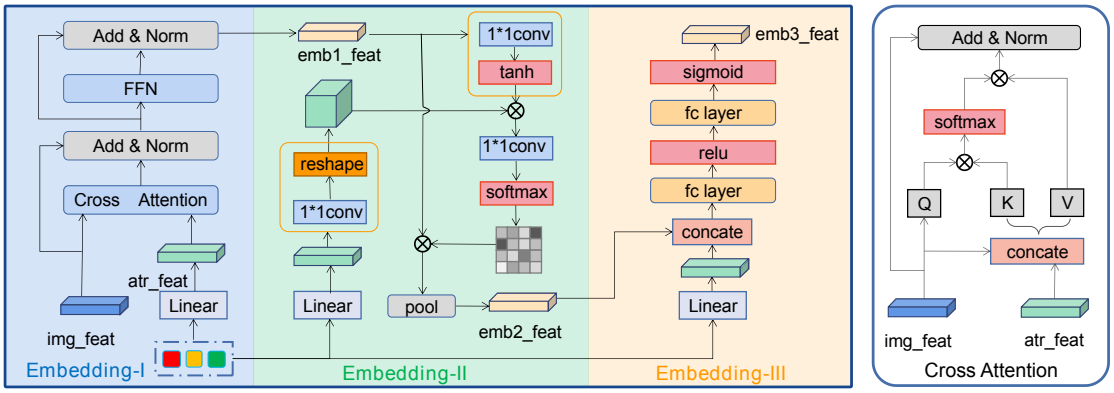


Figure 3: Progressive Attribute Embedding– We take attributes as input and achieve progressive fusion with image features through three different embeddings.

3.3 Attribute-based Auxiliary Learning

After three layers of progressive attribute embedding, the fused features F_e^m can well integrate attribute information into the image features. However, it may be exceedingly biased to the attribute information while weakening the identity information. Therefore, we propose an attribute-based auxiliary learning (AAL) module only in the training phase.

The purpose of this module is to use an auxiliary attribute classification task to help learn more detailed identity information learning. In this way, attributes and images are able to utilize their respective useful information to complement each other and enhance the feature representation. In addition, this module introduces original image features, which are used to help generate better attribute features for attribute classification tasks. We first concatenate all the fused feature (f_i^m) after pooling, and then calculate the attention weight,

$$W_1 = sigmoid\left[f_1^m, f_2^m, \cdots, f_n^m\right]. \tag{8}$$

Then we copy n copies of the pooled features f^m obtained by the feature extractor, and then multiply them with W_1 to avoid losing global information. Next, we use the fully connected (FC) layer to obtain the attribute feature representation (f_{atr}^m), which can well reflect the information of the attribute-related area.

Attribute Loss. We add attribute classification branches for $f_{atr_i}^m$, and set an attribution classifier to obtain the attribution prediction ($\bar{p}_i$) through the constraints of additional attribute labels ($\bar{q}_i$). In our model, the binary cross-entropy loss is used for optimization, and the loss calculation formula is as follows,

$$\mathcal{L}_{atr} = \sum_{i=1}^{M} -\bar{q}_i \log\left(\bar{p}_i\right) \tag{9}$$

where M represents the number of person in a mini-batch.

3.4 Optimization

For identity classification task, we first calculate the mean value of the n feature maps F_e^m obtained in the previous module, and then calculate the attention weight through the *sigmoid* function,

$$W_2 = sigmoid\left[\frac{1}{n}\sum_{i=1}^{n} F_i^m\right], \tag{10}$$

where the generated attention weight reflects the correlation between the local area and the corresponding attributes. Finally, we multiply it with the feature map (F^m) extracted by the feature extractor, and the final person feature (f_{id}^m) representation is obtained after pooling.

In this way, the final image features can inherit information from different patterns and capture explicit structural information and subtle pixel-level features. Therefore, we can obtain the attribute features (f_{atr}^m) and image features (f_{id}^m) of a pedestrian, where $m \in [v, t]$, representing the RGB/IR modality.

The image features are used for the subsequent person ReID task, and the whole network is trained by jointly using baseline loss $\mathcal{L}_b$ and the attribute loss $\mathcal{L}_{atr}$ in the AAL module. The overall objective function is,

$$\mathcal{L}_{total} = \mathcal{L}_b + \lambda \mathcal{L}_{atr}, \tag{11}$$

where the λ is a super-parameter. The constraint of different tasks enforce the network to learn both modality-independent and identity-consistent features, which are more robust and discriminative for cross-modality ReID.

4 EXPERIMENTAL RESULTS

4.1 Experimental Setting

We evaluate our method on two benchmark cross-modality ReID datasets SYSU-MM01 [31] and RegDB [19] with widely used metrics: the Cumulative Matching Characteristics (CMC) curve [27], the mean Average Precision (mAP) [44] and the mean Inverse Negative Penalty (mINP) [38]. In SYSU-MM01 dataset, we use the eight attributes annotated by Zhang *et al.* [42], including: *gender* (male, female), *hair length* (long, short), *wearing glasses* (yes, no), *sleeve length* (long, short), *type of lower-body clothing* (dress, pants), *length of lower-body clothing* (long, short), *carrying backpack*

(yes, no), and *carrying satchel*(yes, no). For RegDB dataset, we annotate the same eight attributes for each person. For one certain attribute, the value of positive example is 1 while 0 for negative example.

4.2 Implementation details

Our proposed method is implemented in PyTorch. Following existing cross-modality ReID works, ResNet50 [10] is adopted as our backbone network for fair comparison.The first residual block is specific for each modality, while the other four blocks are shared. The stride of the last convolutional block is set to 1 to obtain a fine-grained feature map. We initialize the convolutional blocks with the pre-trained ImageNet parameters. All the input images are firstly resized to 288×144. We adopt random cropping with zero-padding and horizontal flipping for data augmentation. SGD optimizer is adopted for optimization, and the momentum parameter is set to 0.9. We set the initial learning rate to 0.1 with a warm-up strategy. The learning rate decays by 0.1 at the 30th epoch and 0.01 at the 50th epoch, with a total of 80 training epochs. By default, we randomly select 8 identities, and then randomly select 4 visible and 4 infrared images to formulate a training batch.

4.3 Comparison with State-of-the-Art Methods

Table 1 first presents the quantitative comparison on SYSU-MM01 dataset. Our method significantly outperforms the state-of-the-art methods in all the evaluation metrics in both *all-search* and *indoor-search* scenarios. Note that ATTR [42] first uses attributes however works modestly in cross-modality ReID. The main reason is that it simply constrains the network by attribute labels, thus lacking exploration of the relationship between attributes and images, making it difficult to distinguish some pedestrians with similar attributes and appearances. By flexibly embedding attributes into the network, together with the mutual guidance of both attribute and identity information, our method can better use attributes and further constrain the network, thus achieving a new state-of-the-art performance.

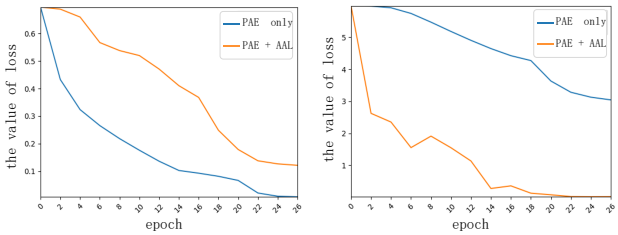
Table 2 reports the comparison results on the RegDB dataset. Since the intra-class differences in RegDB dataset are comparably small, it presents much smaller challenges than the SYSU-MM01 dataset. Therefore the role of attributes embedding is comparably smaller in RegDB. However, our method still effectively improves the performance, compared with our baseline [12].

4.4 Ablation Study

Table 3 reports the ablation study on the SYSU-MM01 dataset in *all-search* setting. As shown in Table 3 (b), by only introducing progressive attribute embedding (PAE), it brings -6.08% and -5.28% decline in Rank-1 and mAP, respectively. The reason is, the network pays more attention to attribute information after embedding, resulting in the loss of image identity information. Hence, we propose attribute-based auxiliary learning (AAL) to explore the relationship between attributes and images. Table 3 (d) achieves significant improvement compared with the "baseline", which evidences the contribution of the AAL. By jointly integrating PAE and AAL, it can facilitate the network to generate more informative image feature representations while using attribute information, thus achieving the best performance.

Table 3: Ablation study of PAE and AAL.

	PAE	AAL	R1	R10	R20	mAP	mINP
(a)	✗	✗	66.75	94.16	97.38	62.25	46.08
(b)	✓	✗	60.67	92.97	97.25	56.97	42.00
(c)	✗	✓	69.32	96.83	99.11	66.55	53.43
(d)	✓	✓	**74.22**	**99.03**	**99.97**	**73.90**	**64.29**



(a) changes of $\mathcal{L}_{atr}$ (b) changes of $\mathcal{L}_{id}$

Figure 5: The changes of attribute loss $\mathcal{L}_{atr}$ and ID loss $\mathcal{L}_{id}$.

Visualization of feature distribution. We further utilize t-SNE [23] to visualize the features in 2D plane on SYSU-MM01 dataset, Fig. 4 demonstrates the feature distribution. Different colors represent different categories. Circles and rectangles boxes are used to highlight the significant changes. It can be seen that the inter-class distributions in Fig. 4 (a) are not well discriminated, and the intra-class distributions are also very scattered. Compared with it, our proposed modules can better cluster the features of the same identity together, while also better separating the features of different identities. With respect to Table 3, our method can better aggregate the intra-class features and simultaneously distinguish the inter-class discrimination.

Evaluation of loss changes. To further demonstrate the impact of progressive attribute embedding and the rationality of attribute-based auxiliary learning, the Fig. 5 shows the changes in attribute loss and ID loss during the first 26 iterations. PAE reduces attribute loss to a certain extent. The network with the PAE module only effectively learns the attribute information. However, the decline in identity loss is not so significant. After adding the AAL module, our network avoids this problem. More importantly, through the joint optimization of embedding and interaction, the performance of our method is further improved, reflecting the complementarity of the two modules.

We further evaluate the results by training these two modules independently in All-search mode on SYSU-MM01 dataset. We fix the parameters of one of the two modules (PAE and AAL) and continue optimizing the other one at the 10th epoch, from when the network achieves comparatively stable convergence. Table 4 evidences the effectiveness of the joint training on PAE and AAL in our method.

Table 1: Comparison with the state-of-the-arts on SYSU-MM01 dataset on two different settings. Rank at r accuracy (%) and mAP (%) are reported. Herein, the best, second and third best results are indicated by red, green and blue fonts.

Settings		*All Search*				*Indoor Search*			
Method	Venue	$r = 1$	$r = 10$	$r = 20$	mAP	$r = 1$	$r = 10$	$r = 20$	mAP
cmGAN [5]	IJCAI 18	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
BDTR [36]	IJCAI 18	27.32	66.96	81.07	27.32	31.92	77.18	89.28	41.86
eBDTR [36]	TIFS 19	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
HSME [9]	AAAI 19	20.68	32.74	77.95	23.12	-	-	-	-
D^2RL [28]	CVPR 19	28.9	70.6	82.4	29.2	-	-	-	-
MAC [34]	TIP 20	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
MSR [7]	TIP 19	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
Align[24]	ICCV 19	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
AGW [38]	TPAMI 21	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
ATTR [42]	JEI 20	47.14	87.93	94.45	47.08	48.03	88.13	95.14	56.84
CMSP [30]	IJCV 20	43.56	86.25	-	44.98	48.62	89.50	-	57.50
Xmodal [13]	AAAI 20	49.92	89.79	95.96	50.73	-	-	-	-
DDAG [37]	ECCV 20	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
ssMF [18]	CVPR 20	61.06	89.02	93.9	63.2	70.5	94.9	97.7	72.6
NFS [3]	CVPR 21	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
CICL [43]	AAAI 21	57.20	94.30	98.40	59.30	66.60	98.80	99.70	74.70
HCT [16]	TMM 20	61.68	93.10	97.17	57.51	63.41	91.69	95.28	68.17
MID [11]	AAAI 22	60.27	92.90	-	59.40	64.86	96.12	-	70.12
GLMC [41]	TNNLS 21	64.37	93.90	97.53	63.43	67.35	98.10	99.77	74.02
SPOT [2]	TIP 22	65.34	92.73	97.04	62.25	69.42	96.22	99.12	74.63
MMD [12]	BMVC 21	66.75	94.16	97.38	62.25	71.64	97.75	99.52	75.95
MPANet [32]	CVPR 21	70.58	96.21	98.80	68.24	76.64	98.21	99.57	80.95
PAENet	-	74.22	99.03	99.97	73.90	78.04	99.58	100	83.54

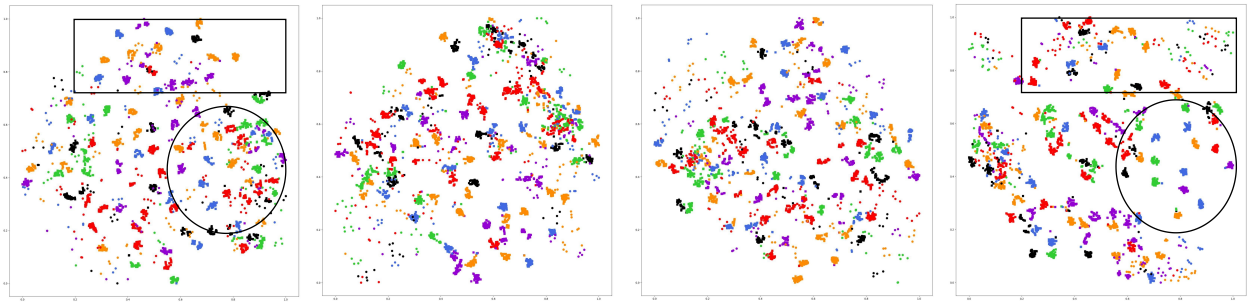


Figure 4: Feature distributions visualized with t-SNE method.

4.5 Evaluation on PAE

In order to better understand the effectiveness of the attribute embedding, we progressively introduce the attribute embedding into the baseline with attribute-based auxiliary learning. Through layer-by-layer embedding, the network adaptively learns the relationship between attributes and the semantic information of different attributes. As shown in Table 5, by progressively imposing the attribute embedding into the baseline with AAL, the performance has been improved gradually, which highlights the effectiveness of the proposed progressive embedding.

4.6 Other Analysis

Visualization of learned features. Fig. 6 visualizes the feature attention maps of the baseline and our model. As shown in Fig. 6 (a), "she is a long-haired woman wearing a long dress and short skirt.", the baseline pays more attention to the texture area on the edge of the top and short skirt. By contrast, our method learns richer and more comprehensive features, avoiding excessive attention to the salient areas. Consistently in Fig. 6 (b), "he is a man with short hair and glasses, wearing short sleeves and shorts," the baseline mainly focuses on the cloth pattern, while our method also focuses on the discriminative face and the under area.

Analysis on balance parameters λ. The auxiliary attribute classification branch in the AAL module helps to learn a more robust discriminative representation. We fine-tune the super-parameter λ in the attribute loss to evaluate the performance in *All-search* mode on SYSU-MM01 dataset. As shown in Fig. 7, we can conclude that the best results are obtained when taking the value of 1.

Different baselines plugin. To evaluate the the generality of our method, which is plug-and-play, we further plug the proposed PAE and AAL modules into five popular open-sourced baselines on SYSU-MM01 dataset. As shown in Table 6, after integrating our modules into the five baselines, it significantly improves all the baselines in all the metrics. This verifies the generality of our method. Note that the final results based on MPANet are comparable to our results on SYSU-MM01 dataset while overshadowed on RegDB dataset, therefore we choose the simple and effective MMD as our baseline.

Table 2: Comparison with the state-of-the-arts on the RegDB dataset. Herein, the best, second and third best results are indicated by red, green and blue fonts.

Method	Venue	*V to T*		*T to V*	
		R1	mAP	R1	mAP
HSME [9]	AAAI 19	41.34	38.82	40.67	37.50
D^{2}RL [28]	CVPR 19	43.40	44.10	-	-
MSR [7]	TIP 19	48.43	48.67	-	-
JSIA [25]	AAAI 20	48.50	49.30	48.10	48.90
AlignGAN [24]	ICCV 19	57.90	53.60	56.30	53.40
CMSP [30]	IJCV 20	65.07	64.50	-	-
CMM+CML [15]	MM 20	-	-	59.81	60.86
Xmodal [13]	AAAI 20	-	-	62.21	60.18
ssMF [18]	CVPR 20	65.40	65.60	63.80	64.20
DDAG [37]	ECCV 20	69.34	63.46	68.06	61.80
Hi-CMD [4]	CVPR 20	70.93	66.04	-	-
HAT [39]	TIFS 21	71.83	67.56	70.02	66.30
CICL [43]	AAAI 21	78.80	69.40	77.90	69.40
NFS [3]	CVPR 21	80.54	72.10	77.95	69.79
SPOT [2]	TIP 22	80.35	72.46	79.37	72.26
MPANet [32]	CVPR 21	82.80	87.70	83.70	80.90
HCT [16]	TMM 20	91.05	83.28	89.30	81.46
GLMC [41]	TNNLS 21	91.84	81.42	91.12	81.06
MID [11]	AAAI 22	87.45	84.85	84.29	81.41
MMD[12]	BMVC 21	95.06	88.95	93.65	87.30
PAENet	-	97.57	91.41	95.35	89.98

Table 4: Ablation experiment of different training methods.

Setting	R1	mAP	mINP
the performance at the 10th epoch	57.35	55.47	41.19
Optimizing PAE while fixing AAL	71.89	69.42	57.58
Optimizing ALL while fixing PAE	73.00	68.80	55.23
Jointly training PAE and AAL	74.22	73.90	64.29

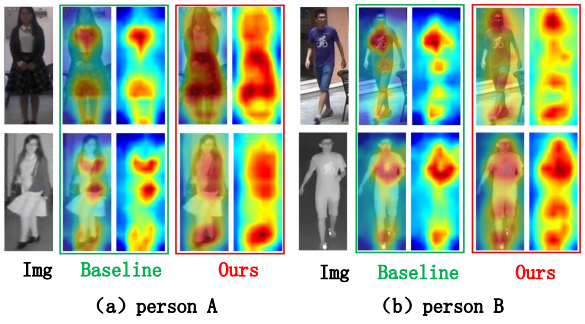


Figure 6: Attention map Comparison of person A and B.

5 CONCLUSIONS

This paper proposes a Progressive Attribute Embedding Net (PAENet) for cross-modality person ReID, which explores how to utilize auxiliary attributes to improve the performance of identifying pedestrians. It contains two main components in a unified framework: progressive attribute embedding (PAE) and attribute-based auxiliary learning (AAL). To mine the rich attribute semantic information,

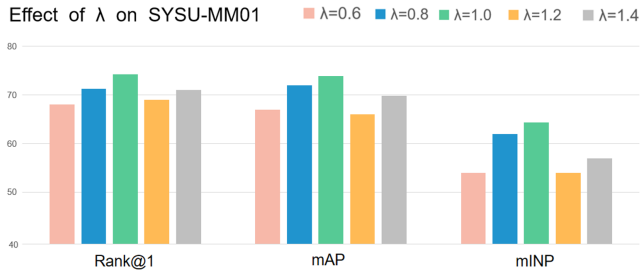


Figure 7: Performance analysis on parameters λ.

PAE fuses fine-grained information and image appearance features well. Guided by attributes, the network adaptively focuses on distinguished key regions and extracts fine-grained modality-sharing features. Moreover, in the training phase, AAL specifically makes full use of additional attribute classification branches to bridge the modal gap further and enhance the robustness of image features, which in turn helps to enhance the effectiveness of attribute recognition. Comprehensive experiments demonstrate the effectiveness of proposed method with superior improvement against the state-of-the-art methods. We believe our modules can also be applied in other tasks, which use auxiliary information.

Table 5: Evaluation on progressive attribute embedding.

	PAE			R1	R10	R20	mAP	mINP
	Emb1	*Emb2*	*Emb3*					
(a)	✗	✗	✗	69.32	96.83	99.11	66.55	53.43
(b)	✓	✗	✗	71.95	97.62	99.22	68.23	54.77
(c)	✓	✓	✗	72.73	97.60	99.45	69.55	56.73
(d)	✓	✓	✓	**74.22**	**99.03**	**99.97**	**73.90**	**64.29**

Table 6: Plugging PAE and AAL into five existing baselines.

		RegDB		SYSU-MM01			
		Visible to Infrared		All-search		Indoor-search	
Method	Venue	R1	mAP	R1	mAP	R1	mAP
AGW [38]	TPAMI 21	70.05	66.37	47.50	47.65	54.17	62.97
AGW+Ours	-	**92.06**	**87.26**	**69.80**	**68.75**	**74.76**	**81.55**
DDAG [37]	ECCV 20	69.34	63.46	54.75	53.02	61.02	67.98
DDAG+Ours	-	**91.03**	**85.70**	**68.13**	**65.21**	**74.10**	**79.58**
MPANet [32]	CVPR 21	82.80	87.70	70.58	68.24	76.64	80.95
MPANet+Ours	-	**90.56**	**90.43**	**73.93**	**74.08**	**79.10**	**82.19**
HCT [16]	TMM 20	91.05	83.28	61.68	57.51	63.41	68.17
HCT+Ours	-	**95.90**	**89.93**	**70.70**	**69.23**	**75.56**	**82.06**
MMD [12]	BMVC 21	95.06	88.95	66.75	62.25	71.64	75.95
MMD+Ours	-	**97.57**	**91.41**	**74.22**	**73.90**	**78.04**	**83.54**

ACKNOWLEDGMENTS

This research is supported in part by the National Natural Science Foundation of China (61976002 and 61860206004), and the Natural Science Foundation of Anhui Higher Education Institution of China (KJ2020A0033).

REFERENCES

[1] Yu-Tong Cao, Jingya Wang, and Dacheng Tao. 2020. Symbiotic Adversarial Learning for Attribute-based Person Search. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 230–247.

[2] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. 2022. Structure-Aware Positional Transformer for Visible-Infrared Person Re-Identification. *IEEE Transactions on Image Processing* 31 (2022), 2352–2364.

[3] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. 2021. Neural Feature Search for Rgb-infrared Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 587–597.

[4] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. 2020. Hi-CMD: Hierarchical Cross-modality Disentanglement for Visible-infrared Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10257–10266.

[5] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-modality Person Re-identification with Generative Adversarial Training. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 677–683.

[6] Chun-Mei Feng, Yunlu Yan, Geng Chen, Huazhu Fu, Yong Xu, and Ling Shao. 2021. Accelerated Multi-modal Mr Imaging with Transformers. *arXiv preprint arXiv:2106.14248* (2021).

[7] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. 2019. Learning Modality-specific Representations for Visible-infrared Person Re-identification. *IEEE Transactions on Image Processing* 29 (2019), 579–590.

[8] Sixue Gong, Xiaoming Liu, and Anil K Jain. 2020. Jointly De-biasing Face Recognition and Demographic Attribute Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 330–347.

[9] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. 2019. HSME: Hypersphere Manifold Embedding for Visible-thermal Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33. 8385–8392.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[11] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. 2022. Modality-Adaptive Mixup and Invariant Decomposition for RGB-Infrared Person Re-Identification. *arXiv preprint arXiv:2203.01735* (2022).

[12] Chaitra Jambigi, Ruchit Rawal, and Anirban Chakraborty. 2021. MMD-ReID: A Simple but Effective Solution for Visible-thermal Person ReID. *arXiv preprint arXiv:2111.05059* (2021).

[13] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. 2020. Infrared-visible Cross-modal Person Re-identification with an X Modality. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 4610–4617.

[14] Huafeng Li, Shuanglin Yan, Zhengtao Yu, and Dapeng Tao. 2019. Attribute-identity Embedding and Self-supervised Learning for Scalable Person Re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 10 (2019), 3472–3485.

[15] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, and Nicu Sebe. 2020. Class-aware Modality Mix and Center-guided Metric Learning for Visible-thermal Person Re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*. 889–897.

[16] Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. 2020. Parameter Sharing Exploration and Hetero-center Triplet Loss for Visible-thermal Person Re-identification. *IEEE Transactions on Multimedia* 23 (2020), 4414–4425.

[17] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. 2017. Hydraplus-net: Attentive Deep Features for Pedestrian Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 350–359.

[18] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. 2020. Cross-modality Person Re-identification with Shared-specific Feature Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13379–13389.

[19] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. 2017. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors* 17, 3 (2017), 605.

[20] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. 2016. Unsupervised Cross-dataset Transfer Learning for Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1306–1315.

[21] Wanru Song, Jieying Zheng, Yahong Wu, Changhong Chen, and Feng Liu. 2019. A Two-stage Attribute-constraint Network for Video-based Person Re-identification. *IEEE Access* 7 (2019), 8508–8518.

[22] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. 2019. Pamtri: Pose-aware Multi-task Learning for Vehicle Re-identification Using Highly Randomized Synthetic Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 211–220.

[23] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using T-SNE. *Journal of machine learning research* 9, 11 (2008).

[24] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. 2019. Rgb-infrared Cross-modality Person Re-identification Via Joint Pixel and Feature Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3623–3632.

[25] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. 2020. Cross-modality Paired-images Generation for RGB-infrared Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 12144–12151.

[26] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 2018. Transferable Joint Attribute-identity Deep Learning for Unsupervised Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2275–2284.

[27] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. 2007. Shape and Appearance Context Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1–8.

[28] Zheng Wang, Junjun Jiang, Yang Wu, Mang Ye, Xiang Bai, and Shin'ichi Satoh. 2019. Learning Sparse and Identity-preserved Hidden Attributes for Person Re-identification. *IEEE Transactions on Image Processing* 29 (2019), 2013–2025.

[29] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. 2019. Learning to Reduce Dual-level Discrepancy for Infrared-visible Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 618–626.

[30] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. 2020. RGB-IR Person Re-identification by Cross-modality Similarity Preservation. *International Journal of Computer Vision* 128, 6 (2020), 1765–1785.

[31] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-infrared Cross-modality Person Re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 5380–5389.

[32] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. 2021. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4330–4339.

[33] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. 2020. Simulating Content Consistent Vehicle Datasets with Attribute Descent. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 775–791.

[34] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. 2020. Cross-modality Person Re-identification via Modality-aware Collaborative Ensemble Learning. *IEEE Transactions on Image Processing* 29 (2020), 9387–9399.

[35] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. 2018. Hierarchical Discriminative Learning for Visible Thermal Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 7501–7508.

[36] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. 2019. Bi-directional Center-constrained Top-ranking for Visible thermal Person Re-identification. *IEEE Transactions on Information Forensics and Security* 15 (2019), 407–419.

[37] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. 2020. Dynamic Dual-attentive Aggregation Learning for Visible-infrared Person Re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 229–247.

[38] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1.

[39] Mang Ye, Jianbing Shen, and Ling Shao. 2020. Visible-infrared Person Re-identification via Homogeneous Augmented Tri-modal Learning. *IEEE Transactions on Information Forensics and Security* 16 (2020), 728–739.

[40] Jianfu Zhang, Li Niu, and Liqing Zhang. 2020. Person Re-identification with Reinforced Attribute Attention Selection. *IEEE Transactions on Image Processing* 30 (2020), 603–616.

[41] Liyan Zhang, Guodong Du, Fan Liu, Huawei Tu, and Xiangbo Shu. 2021. Global-local Multiple Granularity Learning for Cross-modality Visible-infrared Person Re-identification. *IEEE Transactions on Neural Networks and Learning Systems* (2021), 1–11.

[42] Shikun Zhang, Changhong Chen, Wanru Song, and Zongliang Gan. 2020. Deep Feature Learning with Attributes for Cross-modality Person Re-identification. *Journal of Electronic Imaging* 29, 3 (2020), 033017.

[43] Zhiwei Zhao, Bin Liu, Qi Chu, Yan Lu, and Nenghai Yu. 2021. Joint Color-irrelevant Consistency Learning and Identity-aware Modality Adaptation for Visible-infrared Cross Modality Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 35. 3520–3528.

[44] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1116–1124.