Attributes based Visible-Infrared Person Re-identification

Aihua Zheng[1,2,3], Mengya Feng[2,4], Peng Pan[2,4], Bo Jiang[1,2,4], and Bin Luo[1,2,4,*]

[1] Information Materials and Intelligent Sensing Laboratory of Anhui Province
[2] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation
[3] School of Artificial Intelligence, Anhui University
[4] School of Computer Science and Technology, Anhui University
Emails:ahzheng214@foxmail.com,fmy5012@163.com,anlepanp@foxmail.com, zeyiabc@163.com,ahu_lb@163.com

Abstract. Visible-infrared person re-identification (VI-ReID) is a challenging cross-modality pedestrian retrieval problem. Although there is a huge gap between visible and infrared modality, the attributes of person are usually not changed across modalities, such as person's gender. Therefore, this paper proposes to use attribute labels as an auxiliary information to increase cross-modality similarity. In particular, we design the identity-based attention module to filter attribute noise. Then we propose the attributes-guided attention module to drive the model to focus on identity-related regions. In addition, we re-weight the attribute predictions considering the correlations among the attributes. Finally, we use the attention-align mechanism to align the attribute branch with the identity branch to ensure identity consistency. Extensive experiments demonstrate that proposed method achieves competitive performance compared with the state-of-the-art methods under various settings.

Keywords: Person Re-identification · Cross-modality · Attributes-based.

1 Introduction

Cross-modality visible-infrared person re-identification (VI-ReID)[1] aims to match images of people captured by visible and infrared (including near-[1] and far-infrared (thermal)[2]) cameras. VI-ReID is challenging due to large visual differences between the two modalities and changing camera environments, leading to large intra- and cross-modality variations. To address the above challenges, a series of approaches have been proposed[3–6].

As auxiliary information, attributes have been proved as an effective information to boost the vision tasks [7]. Introducing attributes in VI-ReID has the following advantages: First, the attribute information is modality-invariant. That is, the attributes of pedestrians generally do not change due to modality changes. Therefore, with the help of attribute information, intra-class cross-modality similarity can be increased. Second, detailed attribute labels explicitly guide the network to learn the person representation by designated human characteristics.

With only identity labels in datasets, it is hard for the VI-ReID networks to learn a robust semantic feature representation to infer the differences among pedestrians. With the attribute labels, the network is able to classify the pedestrians by explicitly focusing on some local semantic descriptions. Third, attributes can accelerate the retrieval process of VI-ReID by filtering out some gallery images without the same attributes as the query. Zhang *et al.*[8] propose a network to learn modality invariant and identity-specific local features with the joint supervision of attribute classification loss and identity classification loss. Simultaneously, they manually annotate attribute labels for SYSU-MM01 [1] dataset. However, they ignore the correlation between attribute and identity features, which will generate redundant information. Secondly, the correlations of attributes are not considered. Usually, a pedestrian presents multiple attributes at the same time, and correlations between attributes may help to re-weight the prediction of each attribute. For example "long hair" is highly correlated with gender being "female". In addition, they ignore the consistency of attribute and identity features. Two pedestrians with different identities may have the same attributes. At this time, pulling the distance between different identities through the loss function will impair the recognition ability of the network.

In order to solve the above problems and make full use of attribute information, we propose a novel attributes-based VI-ReID framework. It mainly consists of three key modules: identity-guided attention module (IA), attributes-guided attention module (AA) and attributes re-weighting module (RW). IA aims to obtain attention weights by computing the similarity between attribute features and identity features, and then weighting the attribute features to filter the attribute noise. AA uses attribute features and identity feature map to compute attention maps, with the aim of selecting regions of the feature map that are relevant to the intrinsic attributes, thus avoiding the network to focus on irrelevant information such as the background. Inspired by [9], an attributes re-weighting module (RW) is introduced to optimize attribute prediction by using the correlation between attributes. In addition, we propose an attention-align mechanism (ALG) to ensure identity consistency, which is achieved using attention alignment loss. Our main contributions are as follows:

- We propose an attributes-based VI-ReID, which improves the intra-class cross-modality similarity with attribute labels as auxiliary information.
- We propose an identity-guided attention module (IA), which aims to weight the attribute vectors using the correlation between attribute features and identity features.
- We propose an attributes-guided attention module (AA), which uses attention maps between attributes and identity feature map to drive the network more focused on attribute-related regions.
- We propose an attention-align mechanism (ALG), which uses attention alignment loss to ensure identity consistency.

2 Related Work

Visible-Infrared Person Re-ID. The visible-infrared cross-modality person re-identification (VI-ReID) [10] aims to match visible and infrared images of the same pedestrian under non-overlapping cameras. On the one hand, Wu *et al.* [1] first create the SYSU-MM01 dataset for the evaluation of VI-ReID. Ye *et al.* [3] propose to extract pedestrian features of different modalities using a two-stream network, and then further reduce the difference between the two modalities by constraining shared feature embedding. Subsequently, Ye *et al.* [11] solve the cross-modality discrepancy by a modality-aware collaborative learning approach. To make the shared features free of redundant information, Dai *et al.* [4] propose a GAN-based training method for shared feature learning.

All the above methods focus only on the learning of shared features and ignore the role of specific features. To address this problem, some methods based on modality-specific feature compensation have been proposed. Kniaz *et al.* [13] generate corresponding infrared images using visible images. Wang *et al.* [5] propose two-level difference reduction learning based on bidirectional loop generation to reduce the gap between different modalities. Lu *et al.* [6] use both shared and specific features for mutual transformation through a shared and specific feature transformation algorithm.

Attribute for Person Re-ID. Attributes, as an additional complimentary annotation, can provide higher-level semantic recognition information and have been introduced into pedestrian re-identification. Liu *et al.* [14] have annotated attributes for two datasets: Market-1501 and DukeMTMC-reID, and also designed a multi-task classification model using attribute labels to assist the person Re-ID task. Yang *et al.* [15] propose an HFE network based on cascaded feature embedding to explore the combination of attribute and ID information in attribute semantics for attribute recognition. Deep learning methods [16] use attributes to aid the supervision of joint training, thus improving the discrimination of identity features and enhancing the relevance of image pairs. To make full use of attribute information by dropping incorrectly labeled attributes, a feature aggregation strategy is proposed by Zhang *et al.* [17]. Tayet *et al.* [18] augment identity features with attribute attention graphs where class-sensitive activation regions for various attributes such as clothing color, hair, gender, etc. were highlighted.

The pedestrian attribute recognition task and the pedestrian re-identification task differ in their feature granularity approaches; the pedestrian re-identification task focuses on global features of pedestrian images, while the latter focuses on local features of pedestrian images. Most of the above approaches however ignore the differences between these two tasks.

3 Method

3.1 Architecture overview

The overview of the proposed method is illustrated in Fig. 1. For the visible branch, it contains an attribute classification branch and an identity classifica-

tion branch, which are used to extract attribute features and identity features, respectively. The infrared branch also follows this design. In identity classification branch, the input images including the visible images and infrared images are fed into the two-stream network to extra the image features. In attribute classification branch, we divide the feature map output from the fourth residual block of ResNet50 [20] into k overlapping horizontal sections (here k=8) to learn the attribute features of pedestrians respectively.

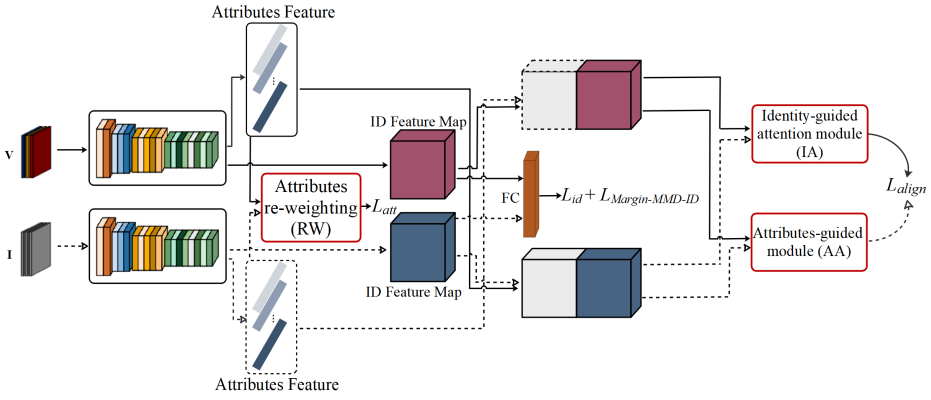

Fig. 1. Framework of attribute-based cross-modality person re-identification.

3.2 Attributes re-weighting module (RW)

For a pedestrian, there is usually multiple attribute information at the same time, and there is a certain correlation between different attributes. For example, "gender" is related to "hair length", and "skirt" is related to "gender". The attributes re-weighting module aims to exploit the correlation between attributes to optimize attribute prediction. For image x, a set of attribute predictions $\{\widetilde{a}^{(1)}, \widetilde{a}^{(2)},..., \widetilde{a}^{(k)}\}$ can be obtained through the attribute classification branch, where $\widetilde{a}^{(j)} \in [0,1]$ is the j-th attribute prediction score. Following the same design as [9], we concatenate the prediction scores as vector $\widetilde{\boldsymbol{a}} \in R^{1\times k}$. Then the confidence score $\boldsymbol{c}$ for its prediction $\widetilde{\boldsymbol{a}}$ is learned as,

$$\mathbf{c} = Sigmoid(\boldsymbol{w}\widetilde{\boldsymbol{a}}^{\boldsymbol{T}} + \boldsymbol{b}), \tag{1}$$

where $\boldsymbol{w} \in \mathrm{R}^{k\times k}$ and $\boldsymbol{b} \in \mathrm{R}^{k\times 1}$ are trainable parameters. In this way, the attributes re-weighting module converts the original predicted label $\widetilde{\boldsymbol{a}}$ into a new prediction score as,

$$\mathbf{a} = \boldsymbol{c} \cdot \widetilde{\boldsymbol{a}}^{\boldsymbol{T}}. \tag{2}$$

For instance, when the prediction scores of "long hair" and "dress" are higher, the network may tend to increase the prediction score of "female".

3.3 Attributes-guided attention module (AA)

Attribute features are usually associated with specific regions of an image. Therefore, we propose the attributes-guided attention module (AA) to select regions in identity feature map which are most relevant to intrinsic attributes. This can effectively avoid learning identity-independent features, such as background. As shown in Fig. 2. The input includes the identity feature map $\boldsymbol{V}$ and attribute embeddings $[\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, ..., \boldsymbol{a}^{(k)}]$, and produces an attention map for regional features. The i-th attribute-guided attention weights are given as,

$$\boldsymbol{m}^{(i)} = \sigma(\boldsymbol{V}^T \boldsymbol{a}^{(i)}), \tag{3}$$

where $\sigma(x)$ is sigmoid function. The generated attention weight mask $\boldsymbol{m}^{(i)} \in \mathrm{R}^L$ reflects the correlation between local region L and the i-th attribute. There are k attributes, so we can obtain k attention maps. Then, we merge them via maxpooling, $\boldsymbol{m}^{(region)} = \max(\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, ..., \boldsymbol{m}^{(k)})$ as the final attention map. The resulting attention map focuses on regions more associated with specific attributes, thus avoiding the attention to background information. The regional features are multiplied by the attention weights and summed to produce the identity representation $\boldsymbol{f}^{(region)} \in \mathrm{R}^d$,

$$\boldsymbol{f}^{(region)} = \frac{1}{L} \boldsymbol{V} \boldsymbol{m}^{(region)}. \tag{4}$$

3.4 Identity-guided attention module (IA)

IA aims to select attributes most related to identity as illustrated in Fig. 3. It takes the attribute embeddings $\boldsymbol{A} = [\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, ..., \boldsymbol{a}^{(k)}]$ and the identity embedding $\boldsymbol{v}^{(id)}$ as input. By calculating the similarity between attribute embeddings and identity embedding, the attention weights based on IA is obtained, and the calculation formula as,

$$s^{(attr)} = \sigma(\boldsymbol{A}^T \boldsymbol{v}^{(id)}). \tag{5}$$

The attribute features are fused via weighting. By this way, we can obtain attribute features $\boldsymbol{f}^{(attr)}$, which is most relevant to pedestrian identity.

$$\boldsymbol{f}^{(attr)} = \frac{1}{k} \boldsymbol{A} s^{(attr)}. \tag{6}$$

3.5 Attention-align mechanism (ALG)

To ensure that the final features have identity consistency, we design an attention-align mechanism between identity-guided branch and the attribute-guided branch, which is implemented by attention alignment loss. Assuming that the features learned by the two branches belong to the same identity, the features of identity and attribute should follow the same distribution. Therefore, the two should have a high similarity. Both $\boldsymbol{f}^{(attr)}$ and $\boldsymbol{f}^{(region)}$ are 256-dim feature vectors.

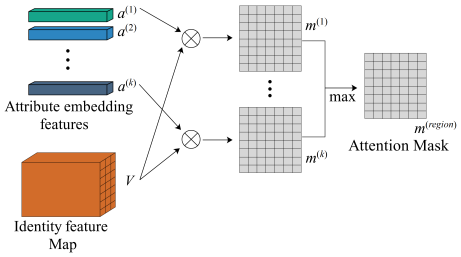

Fig. 2. Framework of attributes-guided attention module.

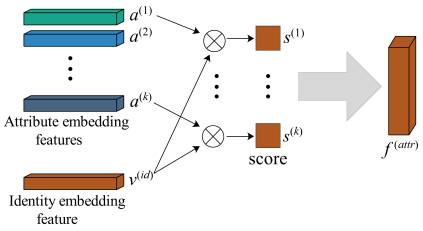

Fig. 3. Framework of identity-guided attention module.

We regard each dimensional as a sample point in the 256-dim space and $\boldsymbol{f}^{(attr)} \sim \mathrm{N}_a(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, $\boldsymbol{f}^{(region)} \sim \mathrm{N}_r(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, where $\boldsymbol{\mu}$ is 256-dim mean vector and $\boldsymbol{\Sigma}$ is 256×256 covariance matrix. Inspired by [21], we adopt Jensen-Shannon (JS) divergence [22] to compute the similarity between N_a and N_r. The JS divergence between N_a and N_r as,

$$JS(N_a, N_r) = D_{KL}(N_a \parallel N) + D_{KL}(N_r \parallel N), \tag{7}$$

where N is the mixture $(N_a + N_r)/2$, and D_{KL} means the KullbackLeibler (KL) divergence. Since the $\boldsymbol{f}^{(attr)}$ and $\boldsymbol{f}^{(region)}$ are constrained by identity information, the feature space is compact and we can use $JS_1(N_a, N_r)$ to measure the similarity.

$$JS_1(N_a, N_r) = D_{KL}(N_a \parallel N_r) + D_{KL}(N_r \parallel N_a). \tag{8}$$

Given two distributions $N_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $N_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ with the same dimension d, the KL divergence is as,

$$D_{KL}(N_0 \parallel N_1) = \frac{1}{2}(tr(\sum_1^{-1} \boldsymbol{\Sigma}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \sum_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - d + \ln(\frac{det\boldsymbol{\Sigma}_1}{det\boldsymbol{\Sigma}_0})). \tag{9}$$

In this way, the similarity JS_1 can be rewritten as,

$$\begin{aligned} JS_1(N_a, N_r) = &\frac{1}{2}(tr(\sum_1^{-1} \boldsymbol{\Sigma}_a) + (\boldsymbol{\mu}_r - \boldsymbol{\mu}_a)^T \sum_1^{-1} (\boldsymbol{\mu}_r - \boldsymbol{\mu}_a) - d + \ln(\frac{det\boldsymbol{\Sigma}_r}{det\boldsymbol{\Sigma}_a})) \\ &+ \frac{1}{2}(tr(\sum_1^{-1} \boldsymbol{\Sigma}_r) + (\boldsymbol{\mu}_a - \boldsymbol{\mu}_r)^T \sum_1^{-1} (\boldsymbol{\mu}_a - \boldsymbol{\mu}_r) - d + \ln(\frac{det\boldsymbol{\Sigma}_a}{det\boldsymbol{\Sigma}_r})). \end{aligned} \tag{10}$$

Each channel of $\boldsymbol{f}^{(attr)}$ are relatively independent since they are extracted by d individual convolution filters, analogously $\boldsymbol{f}^{(region)}$. Therefore, we only consider the diagonal elements of the covariance $\boldsymbol{\Sigma}$ and the other elements are zero. Identity consistency is guaranteed by minimizing the $JS_1(N_a, N_r)$ feature distribution. The final alignment loss calculation formula is as,

$$L_{align} = \frac{1}{2}[\parallel \boldsymbol{\mu}_a - \boldsymbol{\mu}_r \parallel_2^2 + \parallel \boldsymbol{\sigma}_a - \boldsymbol{\sigma}_r \parallel_2^2], \tag{11}$$

where $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_r$ represent the mean vectors of $\boldsymbol{f}^{(attr)}$ and $\boldsymbol{f}^{(region)}$, respectively, and $\boldsymbol{\sigma}_a$, $\boldsymbol{\sigma}_r$ are vectors consisting of the diagonal elements of the covariance matrix.

3.6 Optimization

Attribute classification branch. In order to better learn attribute features, we set a classifier for each attribute to classify it through the constraints of attribute labels. In our model, the binary cross-entropy loss is used for optimization, we take the sum of all the suffered losses for k attribute predictions on the input image x_i as the loss for the i-th sample, and the loss calculation formula is as,

$$L_{att} = -\sum_{i=1}^{n}\sum_{j=1}^{k}[y_i \log(p_i) + (1-y_i)\log(1-p_i)], \tag{12}$$

where y_i represents the category of the i-th attribute, p_i the predicted value of the i-th attribute classifier, and k represents the number of attributes ($k = 8$).
Identity classification branch. For the identity classification branch, the identity loss is used for optimization, and the loss calculation formula is as,

$$L_{id} = -\sum_{i=1}^{n} \log \frac{e^{\boldsymbol{W}_{y_i}^T \boldsymbol{x}_i + \boldsymbol{b}_{y_i}}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j^T \boldsymbol{x}_i + \boldsymbol{b}_j}}, \tag{13}$$

where n represents the number of identitys, $\boldsymbol{W}_j$ represents the parameters for the j-th column, $\boldsymbol{b}$ represents the bias term, $\boldsymbol{x}_i$ denotes the features extracted by i-th sample belonging to the y_i class. Inspired by [23], we use MMD loss to minimize the intra-class distance. The overall objective function is,

$$L = L_{align} + L_{id} + \lambda_1 L_{att} + \lambda_2 L_{Margin-MMD-ID}. \tag{14}$$

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

SYSU-MM01 [1] is a large-scale dataset public by Wu et al. in 2017, which consists of visible and near-infrared images. The training set consists of 395 pedestrians, including 22,258 visible images and 11,909 infrared images. The test set consists of 96 pedestrians and contains 3,803 infrared images as a query set. It contains two different testing settings, all-search and indoor-search mode. Besides, we use the attribute labels marked by Zhang *et al.* [8], including eight attributes: gender (male, female), hair length (long, short), wearing glasses or not (yes, no), sleeve length (long, short), type of lower-body clothing (dress, pants), length of lower-body clothing (long, short), carrying backpack or not (yes, no), and carrying satchel or not (yes, no). For one certain attribute, the value of positive example is 1 and the value of negative example is 0.

RegBD [2] is collected by a dual-camera system, including 412 pedestrians and 8,240 images in total. Each identity has 10 different thermal images and 10 different visible images. Besides, we manually annotate the same eight attribute labels for this dataset according to the Zhang *et al.* [8].

All experimental settings follow the standard evaluation protocol of existing VI-ReID methods: the Cumulative Matching Characteristics (CMC) curve and the mean Average Precision (mAP). We adopt the CMC at rank-1, rank-10 and rank-20. Besides, all the experimental results are based on the average of 10 random trials.

4.2 Implementation details

Following existing VI-ReID works, we adopt ResNet50 [20] as our backbone network for fair comparison. The last residual block is shared for each modality while the other blocks are specific. 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 [24]. 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. The hyperparameter λ_1 is set to 0.15, and the hyperparameter λ_2 is set to 0.25 following the setting in [23].

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
Zero-Pad [1]	ICCV 17	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
TONE[3]	AAAI 18	12.52	50.72	68.60	14.42	20.82	68.86	84.46	26.38
HCML[3]	AAAI 18	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN[4]	IJCAI 18	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
BDTR[25]	IJCAI 18	27.32	66.96	81.07	27.32	31.92	77.18	89.28	41.86
eBDTR[25]	TIFS 19	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
D^2RL[5]	CVPR 19	28.9	70.6	82.4	29.2	-	-	-	-
AlignGAN[10]	ICCV 19	42.40	85.00	93.70	40.70	45.90	87.60	94.40	54.30
AGW[26]	TPAMI 21	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
ATTR[8]	JEI 20	47.14	87.93	94.45	47.08	48.03	88.13	95.14	56.84
XIV-ReID [27]	AAAI 20	49.92	89.79	95.96	50.73	-	-	-	-
DDAG[28]	ECCV 20	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
cm-ssFT [6]	CVPR 20	61.60	89.20	93.90	63.20	70.50	94.90	97.70	72.60
NFS[29]	CVPR 21	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
CICL[30]	AAAI 21	57.20	94.30	98.40	59.30	66.60	98.80	99.70	74.70
HCT[31]	TMM 20	61.68	93.10	97.17	57.51	63.41	91.69	95.28	68.17
MID[33]	AAAI 22	60.27	92.90	-	59.40	64.86	96.12	-	70.12
GLMC[34]	TNNLS 21	64.37	93.90	97.53	63.43	67.35	98.10	99.77	74.02
SPOT[35]	TIP 22	65.34	92.73	97.04	62.25	69.42	96.22	99.12	74.63
MMD[23]	BMVC 21	66.75	94.16	97.38	62.25	71.64	97.75	99.52	75.95
AB-ReID	-	69.91	97.65	99.49	66.55	72.57	98.53	99.76	78.27

4.3 Comparison with State-of-the-Art Methods

Results on SYSU-MM01 Dataset. The results of SYSU-MM01 dataset are shown in Table 1. Our method significantly outperforms the existing methods under the challenging all-search mode. Although the rank-10 and rank-20 of our proposed method have a slight disadvantage in the indoor-search mode, but with significantly higher mAP as well as rank-1. The ATTR [8] first uses attributes

however works modestly in VI-ReID. The main reason is that it simply embeds attribute information into the network without fully considering the relationship between attributes and identity features.

Results on RegDB Dataset. Table 2 shows the experimental results on the RegDB dataset. It can be seen that our proposed method has obvious advantages. In visible to thermal mode, our method improves 1.11% and 1.74% in rank-1 and mAP, respectively. Moreover, in thermal to visible mode, the our proposed method is close to the highest accuracy on rank-20, only 0.19% lower than it.

Table 2. Comparison with the state-of-the-arts on RegDB 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		*Visible to Thermal*				*Thermal to Visible*			
Method	Venue	$r=1$	$r=10$	$r=20$	mAP	$r=1$	$r=10$	$r=20$	mAP
Zero-Pad [1]	ICCV 17	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
HCML[3]	AAAI 18	24.44	47.53	56.78	20.08	21.70	45.02	55.58	22.24
BDTR[25]	IJCAI 18	33.56	58.61	67.43	32.76	32.92	58.46	68.43	31.96
eBDTR[25]	TIFS 19	34.62	58.96	68.72	33.46	34.21	58.74	68.64	32.49
D^2RL[5]	CVPR 19	43.40	66.10	76.30	44.1	-	-	-	-
AlignGAN[10]	ICCV 19	57.90	-	-	53.60	56.30	-	-	53.40
XIV-ReID [27]	AAAI 20	62.21	83.13	91.72	60.18	-	-	-	-
DDAG[28]	ECCV 20	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
cm-ssFT [6]	CVPR 20	72.30	-	-	72.90	71.00	-	-	71.70
NFS[29]	CVPR 21	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
CICL[30]	AAAI 21	78.80	-	-	69.40	77.90	-	-	69.40
HCT[31]	TMM 20	91.05	97.16	98.57	83.28	89.30	96.41	98.16	81.46
HAT[32]	TIFS 20	55.29	92.14	97.36	53.89	62.10	95.75	99.20	69.37
MID[33]	AAAI 22	87.45	95.73	-	84.85	84.29	93.44	-	81.41
MMD[23]	BMVC 21	95.06	98.67	99.31	88.95	93.65	97.55	98.38	87.30
AB-ReID	-	96.17	98.79	99.84	90.69	94.83	98.07	99.01	89.42

4.4 Ablation Study

Table 3 evaluates the effectiveness of four components including the attributes-based attention module (AA), the identity-based attention module (IA), the attributes re-weighting module (RW), and the attention-align mechanism (ALG) on the SYSU-MM01 dataset under all-search mode. Specifically, "B" indicates the baseline without the four components. By progressively introducing the four components, both rank-1 and mAP increase, which evidences the contribution of each component. Integrating all the four components reach the best performance, which verifies the mutual benefits of the components.

4.5 Other Analysis

Experiments on different networks. To further prove that our modules are plug-and-play, we experimented on three different networks. As shown in Table 4, our proposed method can significantly boost the performance by easily integrating into the existing networks.

Hyperparameters analysis. We evaluate the effect of hyperparameter λ_1 on SYSU-MM01 dataset under the all-search and indoor-search modes, as shown in Fig. 4 and Fig. 5. Clearly, the highest recognition accuracy is achieved when λ_1 takes the value of 0.15 in both the all-search and indoor-search modes. Therefore, the value of the hyperparameter λ_1 in the Eq. (14) is set to 0.15.

Table 3. The effectiveness of modules we proposed. The rank-1 accuracy(%) and mAP (%) are reported.

Index	B	AA	IA	RW	ALG	rank-1	mAP
(1)	✓	✗	✗	✗	✗	60.74	55.97
(2)	✓	✓	✗	✗	✗	64.84	61.71
(3)	✓	✓	✓	✗	✗	66.58	64.89
(4)	✓	✓	✓	✓	✗	68.11	65.73
(5)	✓	✓	✓	✓	✓	**69.91**	**66.55**

Table 4. The effectiveness of modules we proposed. The rank-1 accuracy(%) and mAP (%) are reported.

Method	rank-1	mAP
AGW	47.50	47.65
AGW+Ours	59.47	58.94
(TSLFN+HC)	56.96	54.95
(TSLFN+HC+Ours)	63.38	61.73
(MMD)	66.75	62.25
(MMD+Ours)	69.05	65.50

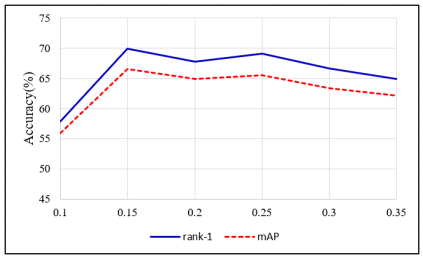

Fig. 4. Effect of hyperparameter λ_1 in all-search mode.

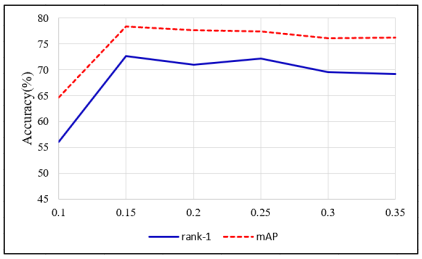

Fig. 5. Effect of hyperparameter λ_1 in indoor-search mode.

5 Conclusions

In this paper, we proposed attributes-based VI-ReID, which increases intra-class cross-modality similarity and mitigates heterogeneity with the help of auxiliary attribute labels. Specifically, attribute noise is filtered by the identity-based guided attention module. The model is prompted to focus on identity-related regions and filter irrelevant information such as background by the attributes-based guided attention module. At the same time, the attributes re-weighting module is designed to fully explore the correlation between attributes. Finally, we propose the attention-align mechanism to align the attribute branches and identity branches to ensure the consistency of pedestrian identity. Extensive experiments validate the effectiveness of our proposed approach.

References

1. Wu, A.C., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380-5389(2017)
2. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition systembased on a combination of body images from visible light and thermal cameras. Sensors 17(3), 605 (2017)

3. Ye,M., Lan,X., Li,J., and Yuen,P.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence, pp. 7501-7508(2018)
4. Dai, P., Ji, R., Wang, H., Wu, Q., and Huang, Y.: Cross-modality person reidentification with generative adversarial training. In: Proceedings of International Joint Conference on Artificial In-telligence, pp. 677–683(2018)
5. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y. Y., and Satoh, S. I.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 618-626(2019)
6. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., and Yu, N.: Cross-modality Person re-identification with Shared-Specific Feature Transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 13379-13389(2020)
7. Cao, Y. T., Wang, J., and Tao, D.: Symbiotic adversarial learning for attribute-based person search. In: European Conference on Computer Vision, pp. 230-247(2020)
8. Zhang, S., Chen, C., Song, W. and Gan, Z.: Deep feature learning with attributes for cross-modality person re-identification. Journal of Electronic Imaging, **29**(3), 033017(2020)
9. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., and Yang, Y.: Improving person re-identification by attribute and identity learning. Pattern Recognition. **95**, 151-161(2019)
10. Wang, G. A., Zhang, T., Cheng, J., Liu, S., Yang, Y.,and Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3623-3632(2019)
11. Ye, M., Lan, X., and Leng, Q.: Modality-aware collaborative learning for visible thermal person re-identification. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 347-355(2019)
12. Hao, Y., Wang, N., Li, J., and Gao, X.: HSME: Hypersphere manifold embedding for visible thermal person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, **33**(1), pp. 8385-8392(2019).
13. Kniaz, V. V., Knyaz, V. A., Hladuvka, J., Kropatsch, W. G., and Mizginov, V.: ThermalGAN: Mul-timodal Color-to-Thermal Image Translation for Person Re-identification in Multispectral Dataset. In: Proceedings of the European Conference on Computer Vision, pp. 606-624(2018)
14. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., and Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision, pp. 350-359(2017)
15. Yang, J., Fan, J., Wang, Y., Wang, Y., Gan, W., Liu, L., and Wu, W.: Hierarchical feature embedding for attribute recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13055-13064(2020)
16. Li, H., Yan, S., Yu, Z., and Tao, D.: Attribute-identity embedding and self-supervised learning for scalable person re-identification. In: IEEE Transactions on Circuits and Systems for Video Technology, **30**(10), 3472-3485(2019)
17. Zhang, J., Niu, L., and Zhang, L.: Person re-identification with reinforced attribute attention selection. In: IEEE Transactions on Image Processing, **30**, 603-616(2020)
18. Tay, C. P., Roy, S., and Yap, K. H.: Aanet: Attribute attention network for person re-identifications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7134-7143(2019)

19. Wang, Z., Jiang, J., Wu, Y., Ye, M., Bai, X., and Satoh, S. I.: Learning sparse and identity-preserved hidden attributes for person re-identification. In: IEEE Transactions on Image Processing, **29**, pp. 2013-2025(2019)
20. He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778(2016)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S.,and Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems, **27**(2014)
22. Kullback, S.: Information theory and statistics. Courier Corporation.(1997)
23. Jambigi, C., Rawal, R., and Chakraborty, A.: MMD-ReID: A Simple but Effective Solution for Visible-Thermal Person ReID. arXiv preprint arXiv:2111.05059(2021)
24. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. arXiv preprint arXiv:1906.08332(2019)
25. Ye, M., Lan, X., Wang, Z., and Yuen, P. C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. In: IEEE Transactions on Information Forensics and Security, **15**, pp. 407-419(2019)
26. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C.: Deep learning for person re-identification: A survey and outlook. In: IEEE Transactions on Pattern Analysis and Machine Intelligence,(2020)
27. Li, D., Wei, X., Hong, X., and Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In : Proceedings of the AAAI Conference on Artificial Intelligence, **34**(4), pp. 4610-4617(2020)
28. Ye, M., Shen, J., J Crandall, D., Shao, L., and Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: European Conference on Computer Vision, pp. 229-247(2020)
29. Chen, Y., Wan, L., Li, Z., Jing, Q., and Sun, Z.: Neural feature search for rgb-infrared person re-identification. In : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 587-597(2021)
30. Zhao, Z., Liu, B., Chu, Q., Lu, Y., and Yu, N.: 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, **35**(4), pp. 3520-3528(2021)
31. Liu, H., Tan, X., and Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. In: IEEE Transactions on Multimedia, pp. 4414-4425(2020)
32. Ye, M., Shen, J., and Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. In: IEEE Transactions on Information Forensics and Security, 728-739(2020)
33. Huang, Z., Liu, J., Li, L., Zheng, K., and Zha, Z. J.: Modality-Adaptive Mixup and Invariant Decomposition for RGB-Infrared Person Re-Identification. arXiv preprint arXiv:2203.01735.
34. Zhang, L., Du, G., Liu, F., Tu, H., and Shu, X.: Global-local multiple granularity learning for cross-modality visible-infrared person reidentification. In: IEEE Transactions on Neural Networks and Learning Systems.(2021)
35. Chen, C., Ye, M., Qi, M., Wu, J., Jiang, J., and Lin, C. W. : Structure-Aware Positional Transformer for Visible-Infrared Person Re-Identification. In: IEEE Transactions on Image Processing, pp. 2352-2364(2022)
36. Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., and Tao, D.: Hetero-center loss for cross-modality person re-identification. Neurocomputing, **389**, pp. 97-109(2020)