Logit and Feature Dual-level Alignment for Visible-Infrared Person Re-Identification

Qiong Wu[1], **Bohong Chen**[1], **Wenfeng liu**[1], **Shuman Fang**[1], **Yuexiao Ma**[1],

[1]Xiamen University, China

{qiong, bhchen}@stu.xmu.edu.cn, 1254022763lwf@gmail.com, {fangshuman, bobma}@stu.xmu.edu.cn

Abstract

Person re-identification (Re-ID) is an important task in video surveillance which automatically searches and identifies people across different cameras. Despite the extensive Re-ID progress in Visible cameras, few works have studied the Re-ID between infrared and Visible images, which is essentially a cross-modality problem and widely encountered in real-world scenarios. The key challenge lies in two folds, i.e., the lack of discriminative information to re-identify the same person between Visible and infrared modalities, and the difficulty to learn a robust metric for such a large-scale cross-modality retrieval. In this paper, we tackle the above two challenges by proposing a novel Dual-level Alignment Network (DANet). To handle the lack of insufficient discriminative information, we design a cutting-edge metric learning objective function to learn discriminative feature representation from different modalities. To handle the issue of modality discrepancy, we integrate a mutual learning manner, which minimize modality discrepancy by close the distribution of two modalities. The entire DANet can be trained in an end-to-end manner by using a standard deep neural network framework. We have quantized the performance of our work in the newly-released RegDB Visible-Infrared Re-ID benchmark, and have reported superior performance

Introduction

Person re-identification (Re-ID) (Gong et al. 2014) aims at matching individual pedestrian images in a query set to ones in a gallery set captured by different cameras. It is challenging due to the variations of viewpoints, body poses, illuminations, and backgrounds. Most existing person Re-ID methods focus on matching pedestrian images captured by visible cameras which can be formulated as a single-modality matching problem. However, such a setting is not workable for ever-increasing visible cameras in surveillance systems such as at night, which cannot capture discriminative information under poor illumination conditions.

Cutting-edge surveillance systems are able to automatically switch from visible to infrared mode, which has accumulated a significant amount of cross-modality data. Re-ID problem in such a cross-modality setting thereby becomes extremely challenging, which is essentially a cross-modality retrieval problem. Compared to conventional person Re-ID, a new challenge arises from the modality discrepancy by different spectrum cameras. To perform visible-infrared person Re-ID, many methods (Wu et al. 2017; Dai et al. 2018; Hao et al. 2019a; Wang et al. 2019b; Li et al. 2020) have been proposed, which aim to alleviate the modality discrepancy by aligning feature or pixel distributions. Despite the encouraging achievement, the existing approaches are still limited in learning discriminative features across different modalities. Besides the modality discrepancy, the limited information in infrared images also increases the difficulty of matching. As shown in Fig. 1, the visible-infrared person re-identification task need to face both the modality discrepancy and the discriminability intra the modality. Due to the large discrepancy between the features who have the same identity, the matching across the modality can be difficult. Essentially, existing visible-infrared person Re-ID methods typically consider the input image as a whole when alleviating the modality discrepancy, which neglects the discriminative information. It closes the features, especially the features of infrared images whose information are much more limited than the visible ones.

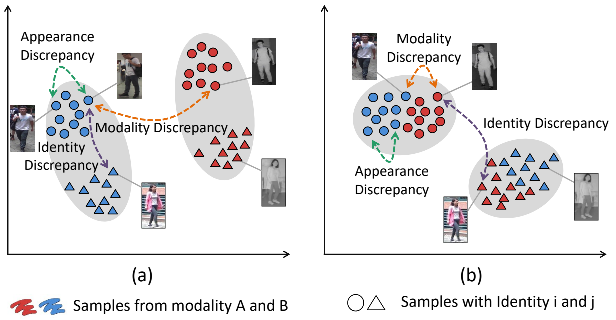

Figure 1: The modality discrepancy is more significant than the appearance variation in the cross-modality Re-ID problem. The distance of the same identity between different modalities is larger than that of different identities in the same modality in the feature space.

The proposed DANet uses a mutual learning manner to close the distribution of two modalities. Two modality-specific classifiers are applied to learning identity information from a certain modality. And we confuse the feature's prediction provided by two classifiers to enforce the backbone extract modality-irrelevant features. Furthermore, we proposed a cross-modality center loss to minimize inter-class ambiguity while maximizing cross-modality similarity among instances.

The contributions of this work are summarized in the following:

- We propose an end-to-end Dual-level alignment network for RGB-IR person re-identification, where the cross-modality representations are learned in a mutual learning manner and a cross-modality center loss is proposed to ensure the discriminability of the features.
- To alleviate the modality discrepancy in the classifier level, we adopt a mutual learning manner to close the distribution of two modalities by confusing the prediction of two modality-specific classifiers.
- To alleviate the modality discrepancy in the feature level, we further propose an objective function called cross-modality center loss to learn discriminative representation while alleviating the modality discrepancy in the feature space.

Related Work

Visible-infrared Person Re-ID. Visible-infrared Person Re-ID has received increasing attention in recent years due to its effectiveness under the poor illumination conditions. To address the challenge caused by modality discrepancy, many cross-modality person Re-ID approaches have been proposed. Wu *et al.* (Wu et al. 2017) proposed a deep zero-padding network learning features in a common space and construct the first large-scale visible-infrared dataset named SYSU-MM01. To constrain the intra-modality and inter-modality variations, an end-to-end dual-stream hypersphere manifold embedding model is proposed in (Hao et al. 2019b). In (Ye et al. 2018b), a dual-path network with a bidirectional dual-constrained top-ranking loss is introduced to learn modality alignment feature representations. And *Ye et al.* also proposed a hierarchical cross-modality matching model that jointly optimizing the modality-specific and modality-shared metrics in (Ye et al. 2018a). DFE (Hao et al. 2019a) is proposed to align the information both in region and modality. Some works are GAN-based approaches, cmGAN (Dai et al. 2018), D^2RL (Wang et al. 2019b), AlignGAN (Wang et al. 2019a) and JSIA-ReID (Wang et al. 2020). cmGAN adopts generative adversarial training to map the features into a common space. D^2RL apply GANs to generate missing modality information extending the input of the feature extractor to four dimensions. Furthermore, AlignGAN and JSIA-ReID implement pixel and feature alignment in a unified GAN framework. Similar, Li *et al.* (Li et al. 2020) and cm-SSFT (Lu et al. 2020) generate a new modality between these two modalities to alleviate the modality discrepancy. Nevertheless, these approaches proposed to replenish the modality information or directly map the features into a common feature space. They mainly focus on alleviating the modality discrepancy while the robustness of the extracted features are ignored.

Metric Learning. Metric learning plays an important role in the deep learning method, due to its wide application. Center Loss (Iqbal et al. 2019) focus on increasing the similarity among the instances have the same ground-truth by closing them to their center. Consider the relationship to the negative instances, Triplet Loss (Schroff, Kalenichenko, and Philbin 2015) is proposed to minimize the distance between positive instance pair while maximizing the distance between negative instance pair. Recently, the metric learning is further consider the balance of grad between the positive pair and the negative pair to better optimizing the model. The proposed Circle Loss (Sun et al. 2020) combines the triplet loss with the Softmax loss to keep the balance between two types of pair. In this paper, we proposed a metric learning method that is designed to consider both the identity and modality for cross-modality retrieval.

Teacher-Student models. In semi-supervised learning methods and knowledge distillation methods, teacher-student models play an important role. The key idea of teacher-student models is to create consistent training supervision for each sample by collecting predictions from different models. It is an effective and widely used technique to transfer knowledge from a teacher to a student network. Temporal ensembling (Laine and Aila 2017) saves an average prediction in an exponential moving way for each sample as the supervisions of the unlabeled samples. To reduce the cost of saving predictions, Mean Teacher (Tarvainen and Valpola 2017) temporally averaged model weights at different training iterations to create the supervisions for unlabeled samples. Different from the one-way transfer between a teacher and a student, deep mutual learning (Zhang et al. 2018) is an ensemble of students who learn collaboratively and teach each other throughout the training process. Combine the mutual learning and mean teaching, MMT (Ge, Chen, and Li 2020) aims to reduce the impact of noise from the pseudo label by using two mean teachers to generate soft labels for another two networks. In these manners, they propose the distillation method to improve performance by teaching each other the specific knowledge. And we apply the mutual learning method to mine the common knowledge involved in the both two modalities closing the modality discrepancy.

Proposed Method

Problem Formulation

Let $\mathcal{V} = \{\mathbf{x}_v^{(i)}\}_{i=1}^{N_v}$ and $\mathcal{R} = \{\mathbf{x}_r^{(i)}\}_{i=1}^{N_r}$ respectively denote the visible images and infrared images in a cross-modality person Re-ID dataset, where N_v and N_r are the numbers of samples in these modalities. There are totally $N = N_v + N_r$ samples in the dataset with the corresponding identity label set $\mathcal{Y} = \{\mathbf{y}^{(i)}\}_{i=1}^{N_p}$, where N_p is the number of identities. Given a query of a certain pedestrian, the cross-modality person Re-ID aims to match the same person by finding a ranked list of images from another modality image set according to similarity.

As shown in Fig. 2, the Dual-level Alignment Network (DANet) learns cross-modality representations to perform visible-infrared person Re-ID. Firstly, DANet adopts a one-

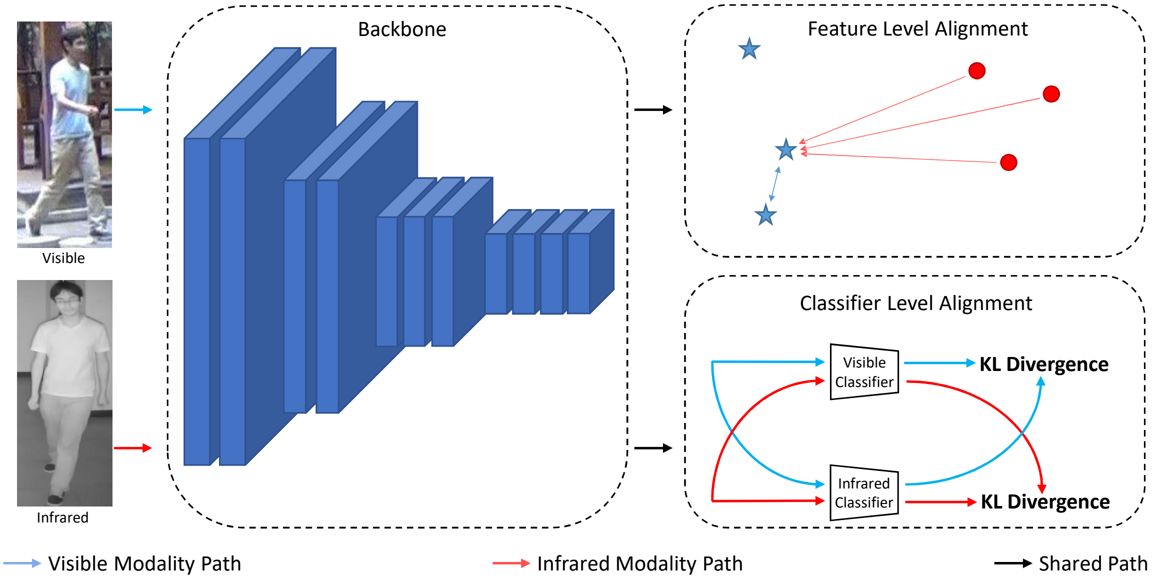

Figure 2: Framework of the proposed Dual-level Alignment Network (DANet). Firstly, the visible and infrared images are fed into the backbone to extract features. Then the features are aligned by reducing the distance to the center with the same identity and the different modality. Further the modality-specific classifiers are applied to predict the class of the features. To alleviate the modality discrepancy, the classifiers are used to predict the features from another modality and align the prediction distributions of a feature. In this way, the modality discrepancy is alleviated in both feature level and classifier level. And in test stage, the modality-irrelevant discriminative features extracted by the backbone will directly work as the representation of images.

stream convolutional neural network $E(\cdot)$ to extract feature maps from both visible and infrared modalities. Then DANet aligns both the modality and the identity in the feature level with the guide of cross-modality center loss. To further alleviate the modality discrepancy, we align the prediction of the same image generated by two different modality-specific classifiers.

Feature Level Alignment

For an input image x, no matter which modality it comes from, we first extract its feature map $\mathbf{Z} = E(\mathbf{x}) \in \mathbb{R}^{h \times w \times c}$, where h, w, c respectively denote the height, width and the dimension of the feature map.

A conventional triplet loss with hard sample mining intra the mini-batch can be represent as:

$$\mathcal{L}_{tri} = \frac{1}{n+m} \sum_{i=1}^{n+m} [\rho - \min_{\substack{k=1,\ldots,n+m \\ y^{(i)} \neq y^{(k)}}} \{||\mathbf{f}^{(i)} - \mathbf{f}^{(k)}||_2\} + \max_{\substack{j=1,\ldots,n+m \\ y^{(i)} = y^{(j)}}} \{||\mathbf{f}^{(i)} - \mathbf{f}^{(j)}||_2\}]_+, \tag{1}$$

where n and m are the numbers of visible images and infrared images in the current batch and ρ is the margin represents the least distance between negative pairs and positive pairs. The $y^{(i)}$ is the identity of feature $\mathbf{f}^{(i)}$ and $[\cdot]_+$ represent function $\max\{\cdot, 0\}$. However, this loss do not consider the modality information. In the test stage, the query set and the gallery set are from different modalities. Thus, for a query sample, we just need to ensure the most similar sample in another modality has the same identity.

To consider the identity and modality information at the same time, we define the cross-modality triplet loss as follow:

$$\mathcal{L}_{c-tri} = \frac{1}{n+m} \sum_{i=1}^{n+m} [\rho - \min_{\substack{k=1,\ldots,n+m \\ y^{(i)} \neq y^{(k)}, s^{(i)} \neq s^{(k)}}} \{||\mathbf{f}^{(i)} - \mathbf{f}^{(k)}||_2\} + \max_{\substack{j=1,\ldots,n+m \\ y^{(i)} = y^{(j)}, s^{(i)} \neq s^{(j)}}} \{||\mathbf{f}^{(i)} - \mathbf{f}^{(j)}||_2\}]_+, \tag{2}$$

where $s^{(i)}$ is the modality label of feature $\mathbf{f}^{(i)}$. Compare with the conventional triplet loss, the cross-modality loss only sample images from different modalities to form the pairs and the optimization process will learn identity and modality information at the same time.

However, the modality discrepancy is increases with the increasing distance to the negative sample. To avoid the identity learning affect modality learning in a bad manner, we propose the cross-modality center loss:

$$\mathcal{L}_{cc} = \frac{1}{n+m} \sum_{i=1}^{n+m} \Big(||\mathbf{f}^{(i)} - \mathbf{c}^{y^{(i)}}_{s\hat{(i)}}||_2 + [\rho - \min_{\substack{k=1,\ldots,n+m \\ y^{(i)} \neq y^{(k)}, s^{(i)} = s^{(k)}}} \{||\mathbf{c}^{y^{(i)}}_{s^{(i)}} - \mathbf{f}^{(k)}||_2\}]_+ \Big), \tag{3}$$

where $\mathbf{c}^{y^{(i)}}_{s\hat{(i)}}$ is the center of features have identity $y^{(i)}$ from different modality to feature $\mathbf{f}^{(i)}$. The Eq. 3 avoid increasing the distance between features from different modalities which increases the modality discrepancy at the same time.

Classifer Level Alignment

Given features $\mathbf{f}_v$ from the visible modality and $\mathbf{f}_r$ from the infrared modality, the modality-specific classifiers provide their predictions. These classifiers are trained with the following cross-entropy loss in a supervised manner:

$$\begin{aligned}\mathcal{L}_{sid} = &-\frac{1}{n}\sum_{i=1}^{n}\log P(y_v^{(i)}|\mathrm{C}_v(\mathbf{f}_v^{(i)}|\theta_v)) \\ &-\frac{1}{m}\sum_{j=1}^{m}\log P(y_r^{(j)}|\mathrm{C}_r(\mathbf{f}_r^{(j)}|\theta_r)),\end{aligned} \tag{4}$$

where n and m respectively denote the numbers of visible and infrared images in the current batch, $\mathbf{y}_v^{(i)}$ and $\mathbf{y}_r^{(j)}$ respectively denote the corresponding label of $\mathbf{f}_v^{(i)}$ and $\mathbf{f}_r^{(j)}$, and $\mathrm{C}_v(\mathbf{f}_v^{(i)}|\theta_v)$ and $\mathrm{C}_r(\mathbf{f}_r^{(j)}|\theta_r)$ are predictions of the two classifiers with parameter θ_v and θ_r, respectively.

As the training images fed to each classifier come from a certain modality, the classifier learns the knowledge only from its corresponding modality. Thus, given a feature $\mathbf{f}$, no matter which modality it comes from, if two modality-specific classifiers provide the same prediction, it means this feature can be regarded as from both two modalities. In other words, the modality discrepancy is eliminated.

To this end, we impose a modality constraint based on Kullback-Leibler divergence as

$$\begin{aligned}\mathcal{L}_M = &\frac{1}{n}\sum_{i=1}^{n}\mathrm{C}_r(\mathbf{f}_v^{(i)}|\theta_r)\log\frac{\mathrm{C}_r(\mathbf{f}_v^{(i)}|\theta_r)}{\mathrm{C}_v(\mathbf{f}_v^{(i)}|\theta_v)} \\ &+\frac{1}{m}\sum_{j=1}^{m}\mathrm{C}_v(\mathbf{f}_r^{(j)}|\theta_v)\log\frac{\mathrm{C}_v(\mathbf{f}_r^{(j)}|\theta_v)}{\mathrm{C}_r(\mathbf{f}_r^{(j)}|\theta_r)}.\end{aligned} \tag{5}$$

This loss encourages the modality-specific classifiers to provide consistent predictions for the same-identity feature, no matter what modalities it comes from.

Optimization

The total loss $\mathcal{L}$ of DANet is defined as

$$\mathcal{L} = \mathcal{L}_{id} + \lambda_1\mathcal{L}_{cc} + \lambda_2\mathcal{L}_M, \tag{6}$$

where $\mathcal{L}_{id}$ is the cross-entropy loss which can be seen as the baseline in retrieval tasks . λ_1 and λ_2 are hype-parameters to balance the contributions of individual loss terms.

Experiments

Datasets and Experimental Setting

Datasets. We evaluate our method on two public datasets **SYSU-MM01** (Wu et al. 2017) and **RegDB** (Nguyen et al. 2017).

- **SYSU-MM01** is a large-scale dataset collected by four visible cameras and two near-infrared ones, including both indoor and outdoor environments. The dataset consists of $30,071$ visible images and $15,792$ infrared images of 491 identities, where the images of each identity are captured by one visible camera and one near-infrared camera at least. The training set contains $22,258$ visible images and $11,909$ infrared ones involving 395 identities, while the query set and the gallery set contain $3,803$ infrared images and 301 $(3,010)$ randomly sampled visible images from 96 identities for *single-shot* (*multi-shot*).
- **RegDB** is constructed by a pair of aligned cameras (one visible and one thermal). It contains $8,240$ images of 412 identities, each having 10 images from a visible camera and 10 images from a thermal one. The dataset is randomly split into two halves: the images of 206 identities for training and the rest also involving 206 identities for testing.

Evaluation metrics. To perform a fair comparison with existing methods, all experiments follow the common evaluation settings in existing cross-modality Re-ID methods. **SYSU-MM01** has two different evaluation settings: the *all-search* mode and *indoor-search* mode. In the *all-search* mode, the gallery set contains images from all the visible cameras, while in the *indoor-search* mode, the gallery set only contains images from the indoor visible cameras. Following (Ye et al. 2018b), **RegDB** contains two test modes: use infrared images as query set and visible images as gallery set, and vice versa. For both datasets, the Cumulative Matching Characteristic (CMC) and mean Average Precision (*m*AP) metrics are adopted to evaluate the performance.

Implementation details. We implement our DANet with PyTorch and train it on a single GTX1080Ti GPU. The mini-batch size is set to 128. For each mini-batch, we randomly sample 16 identities and 8 images for each identity. The model is optimized by using Adam with an initial learning rate of 3.5×10^{-4}, which decays at the $80th$ and $120th$ epoch with a decay factor of 0.1, and the weight decay is set to 5×10^{-4}. The total number of training epochs is set to 140. The cross-modality center loss margin ρ is set to 0.7. The hype-parameters λ_1 and λ_2 are set to 1.0 and 2.5, respectively.

The ResNet-50 (He et al. 2016) pre-trained on ImageNet is employed as the backbone, where the stride size of the last convolutional layer is set to 1. The classifier C_v, C_r, and C are implemented by a single FC layer without bias. Consider the aspect ratio of raw images, the input images are re-scaled to a fixed size of 384×128. In the training stage, the input images are randomly flipped and erased with 50% probability.

Comparison with State-of-the-art Methods

We compare our DANet with state-of-the-art (SOTA) visible-infrared cross-modality person Re-ID approaches. The compared SOTAs include three base methods (Two-stream, One-stream and Zero-Padding) (Wu et al. 2017), three GAN-based methods (cmGAN (Dai et al. 2018), AlignGAN (Wang et al. 2019a) and JSIA-ReID (Wang et al. 2020)), two methods by aligning the modality on a middle modality (XIV-ReID (Li et al. 2020) and cm-SSFT (Lu et al. 2020)). one similarity-based method (SIM (Jia et al. 2020)), and two dual-level alignment methods (DFE (Hao et al. 2019a) and $\mathrm{D}^2\mathrm{RL}$ (Wang et al. 2019b)).

Method	All-Search								Indoor-Search							
	Single-Shot				Multi-Shot				Single-Shot				Multi-Shot			
	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP
Two-stream (Wu et al. 2017)	11.65	47.99	65.50	12.85	16.33	58.35	74.46	8.03	15.60	61.18	81.02	21.49	22.49	72.22	88.61	13.92
One-stream (Wu et al. 2017)	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
Zero-Padding (Wu et al. 2017)	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.86
cmGAN (Dai et al. 2018)	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
$\mathrm{D}^2\mathrm{RL}$ (Wang et al. 2019b)	28.90	70.60	82.40	29.20	-	-	-	-	-	-	-	-	-	-	-	-
JSIA-ReID (Wang et al. 2020)	38.10	80.70	89.90	36.90	45.10	85.70	93.80	29.50	43.80	86.20	94.20	52.90	52.70	91.10	96.40	42.70
AlignGAN (Wang et al. 2019a)	42.40	85.00	93.70	40.70	51.50	89.40	95.70	33.90	45.90	87.60	94.40	54.30	57.10	92.70	97.40	45.30
cm-SSFT(sq) (Lu et al. 2020)	47.70	-	-	54.10	-	-	-	-	57.40	-	-	59.10	-	-	-	-
DFE (Hao et al. 2019a)	48.71	88.86	95.27	48.59	54.63	91.62	96.83	42.14	52.25	89.86	95.85	59.68	59.62	94.45	98.07	50.60
XIV-ReID (Li et al. 2020)	49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
SIM (Jia et al. 2020)	56.93	-	-	60.88	-	-	-	-	-	-	-	-	-	-	-	-
cm-SSFT (Lu et al. 2020)	61.60	89.20	93.90	63.20	63.40	91.20	95.70	**62.00**	70.50	94.90	97.70	72.60	73.00	96.30	99.10	**72.40**
DANet (Ours)	**66.89**	**95.06**	**98.09**	**64.69**	**72.53**	**97.08**	**99.08**	59.53	**73.21**	**97.90**	**99.38**	**78.11**	**80.30**	**99.15**	**99.82**	72.15

Table 1: Comparison of CMC (%) and *m*AP (%) performances with the state-of-the-art methods on **SYSU-MM01**

Method	infrared2visible		visible2infrared	
	Rank-1	*m*AP	Rank-1	*m*AP
Zero-Padding (Wu et al. 2017)	16.7	17.9	17.8	18.9
$\mathrm{D}^2\mathrm{RL}$ (Wang et al. 2019b)	-	-	43.4	44.1
JSIA-ReID (Wang et al. 2020)	48.1	48.9	48.5	49.3
AlignGAN (Wang et al. 2019a)	56.3	53.4	57.9	53.6
XIV-ReID (Li et al. 2020)	62.3	60.2	-	-
DFE (Hao et al. 2019a)	68.0	66.7	70.2	69.2
cm-SSFT (Lu et al. 2020)	71.0	71.7	72.3	72.9
SIM (Jia et al. 2020)	75.2	**78.3**	74.7	**75.2**
DANet(Ours)	**78.3**	73.8	**79.4**	74.3

Table 2: Comparison of the CMC (%) and *m*AP (%) performances with SOTAs on **RegDB**

Method	SYSU-MM01			
	single-shot all-search			
	Rank-1	Rank-10	Rank-20	*m*AP
Baseline	54.50	88.55	94.69	51.84
B + CC	58.57	91.69	97.24	57.78
B + ML	63.49	92.50	96.71	60.17
B + ML + CC	66.89	95.06	98.09	64.69

Table 3: Ablation study in terms of CMC (%) and *m*AP (%) on **SYSU-MM01**

Comparisons on SYSU-MM01. The comparison results on **SYSU-MM01** are shown in Table 1. The proposed DANet outperforms existing SOTAs by large margins. Specifically, DANet achieves the Rank-1 accuracy of 66.89% and *m*AP of 64.69% in the *all-search* and *single-shot* mode, significantly improving the Rank-1 accuracy by 5.29% and *m*AP by 5.04% over the best SOTA cm-SSFT.

Comparisons on RegDB. We also evaluate DANet on a small-scale dataset, **RegDB**, as shown in Table 2. Similar to the results on **SYSU-MM01**, DANet consistently outperforms current SOTAs. Specifically, we achieve Rank-1 accuracy of 78.3% and *m*AP of 73.8% in *infrared2visible* mode, and Rank-1 accuracy of 79.4% and *m*AP of 74.3% in *visible2infrared* mode, significantly improving the Rank-1 by 3.1% and 4.7% in two test modes over the best SOTA SIM.

The above results demonstrate the outstanding performance of MPANet thanks to its ability in cross-modality nuances discovery for visible-infrared person Re-ID.

Ablation Study

In this section, we conduct an ablation experiment to evaluate the contribution of each module. The baseline method uses ResNet-50 as the backbone network followed by the BN neck and an FC layer as the classifier and trained with $\mathcal{L}_{id}$ in the same setting. The ablation experiment is conducted on **SYSU-MM01** in the *all-search single-shot* mode. To illustrate the contribution of each module or objective function, we add them into the model one by one.

As shown in Table 3, the effectiveness of each component is revealed. Compare with baseline, the cross-modality center (CC) loss improve the Rank-1 accuracy and *m*AP accuracy by 4.07% and 5.94%, respectively. And mutual learning (ML) respectively improve the Rank-1 accuracy and *m*AP accuracy by 8.99% and 8.33%. When these two objective functions work together to learn identity and alleviate modality discrepancy, the Rank-1 accuracy and *m*AP accuracy are significantly improved by 12.39% and 12.85%. The results demonstrate that each module plays a role effectively in alleviating modality discrepancy or improving discriminability.

Conclusion

In this paper, we proposed the joint Dual-level Alignment network, termed DANet, to learn the identity information while alleviating the modality discrepancy for visible-infrared person Re-ID. To this end, the proposed DANet focuses on extracting modality-irrelevant features that particularly attend to the identity information. Specifically, DANet first align the modality discrepancy in the feature level by the cross-modality center loss which alleviate the modality discrepancy by reducing the distance between the centers from different modalities of features with a certain identity. Then, in the classifier level, the classifiers learn from each other to increase the similarity between distributions alleviating the modality discrepancy. Finally, We optimize the DANet in an end-to-end manner. Experiment results on two public datasets SYSU-MM01 and RegDB amply proves essential to discover cross-modality nuances in cross-modality retrieval problem and demonstrate the effectiveness of MPANet for visible-infrared person Re-ID.

References

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 Twenty-Seventh International Joint Conference on Artificial Intelligence*, 677–683.

Ge, Y.; Chen, D.; and Li, H. 2020. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Gong, S.; Cristani, M.; Loy, C. C.; and Hospedales, T. M. 2014. The Re-identification Challenge. In *Person Re-Identification*, 1–20.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence*, 8385–8392.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Iqbal, M.; Sameem, M. S. I.; Naqvi, N.; Kanwal, S.; and Ye, Z. 2019. A deep learning approach for face recognition based on angularly discriminative features. *Pattern Recognit. Lett.* 128: 414–419.

Jia, M.; Zhai, Y.; Lu, S.; Ma, S.; and Zhang, J. 2020. A Similarity Inference Metric for RGB-Infrared Cross-Modality Person Re-identification. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 1026–1032.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

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 Thirty-Second AAAI Conference on Artificial Intelligence*, 4610–4617.

Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; and Yu, N. 2020. Cross-Modality Person Re-Identification With Shared-Specific Feature Transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 13376–13386.

Nguyen, D. T.; Hong, H. G.; Kim, K.; and Park, K. R. 2017. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors* 17(3): 605.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition*, 815–823. IEEE Computer Society.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle Loss: A Unified Perspective of Pair Similarity Optimization. In *Conference on Computer Vision and Pattern Recognition*, 6397–6406. IEEE.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019a. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. In *IEEE International Conference on Computer Vision*, 3622–3631.

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

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.; and Satoh, S. 2019b. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 618–626.

Wu, A.; Zheng, W.; Yu, H.; Gong, S.; and Lai, J. 2017. RGB-Infrared Cross-Modality Person Re-identification. In *IEEE International Conference on Computer Vision*, 5390–5399.

Ye, M.; Lan, X.; Li, J.; and Yuen, P. C. 2018a. Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 7501–7508.

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

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep Mutual Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4320–4328.