EasyChair Preprint

№ 8469

# A Knowledge-Driven Enhanced Module for Visible-Infrared Person Re-Identification

Shihao Shan, Enyuan Xiong, Xiang Yuan and Song Wu



July 12, 2022

# A Knowledge-driven Enhanced Module for Visible-Infrared Person Re-Identification *

Shihao Shan[1], Enyuan Xiong[2], Xiang Yuan[3], and Song Wu✉

1,2,3,✉ College of Computer and Information Science, Southwest University, Chongqing, China
songwuswu@swu.edu.cn

**Abstract.** Compared to traditional person re-identification, which only handles the intra-modality discrepancy, the Visible-Infrared Person Re-identification (VI-ReID) suffers from additional cross-modality discrepancy caused by the cross-domain inherent heterogeneity. However, most existing VI-ReID methods ignore the corresponding relationships of intrinsic property knowledge inside cross-modality. Inspired by the human brain's cognitive process of knowledge, in this paper, a novel Knowledge-driven Enhance Module (KDEM) method is designed to imitate the cognitive process of the human brain to achieve the effective matching of cross modalities. Our proposed KDEM aims to discover and integrate the significant semantic pattern from cross-modality representations into a new knowledge-enhanced modality and further enhance the matching accuracy of cross modalities. Meanwhile, a diversity loss is designed to exclude redundant knowledge and preserve the variety of semantic knowledge in the integrated knowledge-enhanced modality. Moreover, a consistency loss is designed to preserve the semantic correlation between the integrated knowledge-enhanced modality and the other two modalities. The evaluation results on two popular benchmark datasets demonstrated the effectiveness of the proposed KDEM, and it obtained competitive performance compared to state-of-the-art methods on the VI-ReID task. The source code of our KDEM is released at https://github.com/SWU-CS-MediaLab/KDEM.

**Keywords:** Cross-modality Retrieval · Person Re-identification · Deep Neural Network

## 1 Introduction

Person re-identification (Re-ID), which aims at associating the same pedestrian images across disjoint camera views, has attracted increasing attention from the computer vision community [22,20,13]. With the high semantic abstraction capability of deep convolutional neural networks (CNNs) [8], the recent CNNs based Re-ID methods have obtained encouraging performance in the visible spectrum images. However, the visible cameras capture the visible spectrum

* Supported by organization Southwest University.

images, which cannot provide sufficient discriminate information in the dark environment. Generally, the surveillance system will automatically be converted to infrared modal under poor lighting or dark conditions in real-life applications. Thus, the cross-modal matching task of Visible-Infrared Person Re-identification (VI-ReID) is proposed to match any one modality data with the corresponding person's another modality data [24]. Specifically, in the task of VI-ReID, since the images are captured by different wavelength ranges of visible and infrared cameras, as shown in Fig.1, the intra-modality variance involved in single-modality Re-ID and the cross-modal discrepancies resulting from the natural difference between the reflectivity of the visible modality and the emissivity of the infrared modality, both increase the challenge for VI-ReID task. In summary, the most challenging issue in VI-ReID is how to learn high discriminative features to reduce the intra-domain and cross-domain discrepancy between the infrared and visible for effective cross-modality matching.

(a) SYSU-MM01 (b) RegDB

**Fig. 1.** Difference between the visible modality and infrared modality, and images in each column are from the same identity. (a) examples from the SYSU-MM01 dataset. (b) examples from the RegDB dataset.

Common representation space learning and metric learning are usually used to overcome the cross-modal discrepancy [19,11]. The existing methods based on common representation space learning mainly focus on how to design a reasonable deep network architecture for extracting robust and discriminative features shared by cross modalities, minimizing the semantic gap between different modalities, and effectively comparing their semantic similarity. The methods based on metric learning aim to design a reasonable metric or loss function to effectively preserve the semantic similarity of different modalities during the deep network training and the loss function optimization. However, an insurmountable gap between visible and infrared modalities makes preserving the semantic similarity of different modalities exceedingly tricky.

Generally, most existing VI-ReID methods ignore the corresponding relationships of intrinsic property knowledge inside different modality data. From the perspective of neuroscience, the human brain's cognitive process of knowledge

discovering and matching is firstly to parse the perceived raw multi-modality data. Then the specific expression pattern inside each modality data will be generated, and the high-level semantic knowledge will be abstracted for each specific expression pattern. The matching process will be completed by comparing the significant or noticeable semantic information in the high-level semantic knowledge. Inspired by this, in this paper, a novel Knowledge-driven Enhance Module (KDEM) is proposed, which is designed to imitate the cognitive process of the human brain, to achieve the effective matching of cross modalities. Specifically, the backbone deep neural network is first used to extract specific feature representations for visible and infrared modalities. The modality-specific feature representations of the visible and infrared modalities are then taken as the inputs of the KDEM. In the proposed KDEM, the high-level semantic knowledge representations will be generated by the followed several convolution operations. Meanwhile, the function of modality influence factors is designed in the KDEM to determine the high discriminative semantic information in the high-level semantic knowledge representations. The high discriminative semantic information extracted from each modality is further integrated as a new modality (knowledge-enhanced modality) to supervise common representation space learning. Thus, the generated knowledge-enhanced modality could learn the significant or noticeable semantic knowledge of both the positive and negative pair-wise samples from intra-modality and cross-modality, effectively reducing the semantic gap between the visible and infrared modalities.

Additionally, in the KDEM, a novel diversity loss is designed to make the generated knowledge-enhanced modality exclude redundant knowledge as much as possible and preserve the variety of semantic knowledge as much as possible. The diversity loss is designed by maximizing the standard deviation of the modality influence factors to enforce the KDEM better accumulate the diverse knowledge of visible and infrared modalities. Moreover, the consistency loss is designed to preserve the semantic correlation of the knowledge-enhanced modality similar to visible and infrared modalities. Benefiting from imitating the cognitive process of the human brain, the proposed KDEM, the designed diversity loss, and the proposed consistency loss could effectively discover and integrate a variety of high distinguish semantic knowledge with the consist of preserving the semantic correlation among different modalities. The main contributions of this paper are summarized as follows:

**Firstly**, a Knowledge-driven Enhance Module (KDEM) is proposed to imitate the cognitive process of the human brain. It could discover and integrate the significant semantic pattern from cross-modality data into a new knowledge-enhanced modality to supervise the learning of robust common representation space.

**Secondly**, the designed diversity loss could effectively exclude redundant knowledge and preserve the variety of semantic knowledge in the integrated knowledge-enhanced modality. Meanwhile, the consistency loss could also preserve the semantic correlation between the knowledge-enhanced modality and the visible and infrared modalities.

**Thirdly**, the evaluation results on two popular VI-ReID datasets show the effectiveness of our proposed KDEM. Meanwhile, compared with the existing state-of-the-art baseline, our KDEM achieves better gain in terms of Rank-k accuracy and mAP.

## 2 Related Work

Given an image of a specific query person in one modality, i.e., a visible image or an infrared image, VI-ReID is a person re-identification task whose goal is to retrieve the query image's counterpart from a gallery set of another modality [23]. The wavelength difference between visible and infrared lights results in a semantic gap in the cross-modality [16]. Thus, extracting cross-modal invariant discriminative knowledge in an advisable way contributes significantly to the task of VI-ReID. Most of the existing VI-ReID methods can be generally divided into two categories. The first category focuses on learning the modality-shared feature representations and aggregating the specific visible and infrared feature representations for better performance. For example, Wu et al.[19] proposed a deep zero-padding network learning feature in a common space and constructed the first large-scale visible-infrared dataset named SYSU-MM01. Ye et al.[24] proposed a dual-path network and a bi-directional dual-constrained top-ranking loss. Moreover, this network was introduced to learn modality alignment feature representations. Park et al.[11] use dense alignment to gain modality-shared discriminating local features. The secondary category methods mainly compensate for the lack of each modality's information, such as GAN-based approaches. Wang et al.[15] leverage GANs to transfer stylistic properties of infrared images to their visible counterparts. In essence, the methods mentioned above aim to overcome the semantic gap and improve the robustness of the model.

## 3 Proposed Method

### 3.1 Overview

A dual-path deep neural network is designed for the VI-ReID, which learns the modality-specific feature representations and optimizes similarity metrics in an end-to-end manner [23]. The detailed architecture of our proposed method is shown in Fig.2, which includes a feature extractor $f(\cdot)$, a classifier $c(\cdot)$, and the knowledge-enhanced modality generator by the designed KDEM. The ResNet-50 [8] is utilized as the backbone for feature extracting, which includes a total of five stages of ResNet-50. The KDEM is designed to discover high discriminative semantic patterns from modality-specific feature representations, and integrate them into a new knowledge-enhanced modality to supervise the network optimization. The KDEM, as shown in Fig.2($b$), is plugged after stage-0 of ResNet-50. Moreover, in order to avoid modeling unknown redundant knowledge and preserve the variety of semantic knowledge during the knowledge accumulation process, meanwhile, to preserve the semantic correlation among different modalities, a diversity loss and a consistency loss are designed and detail is shown in Fig.2($c$).

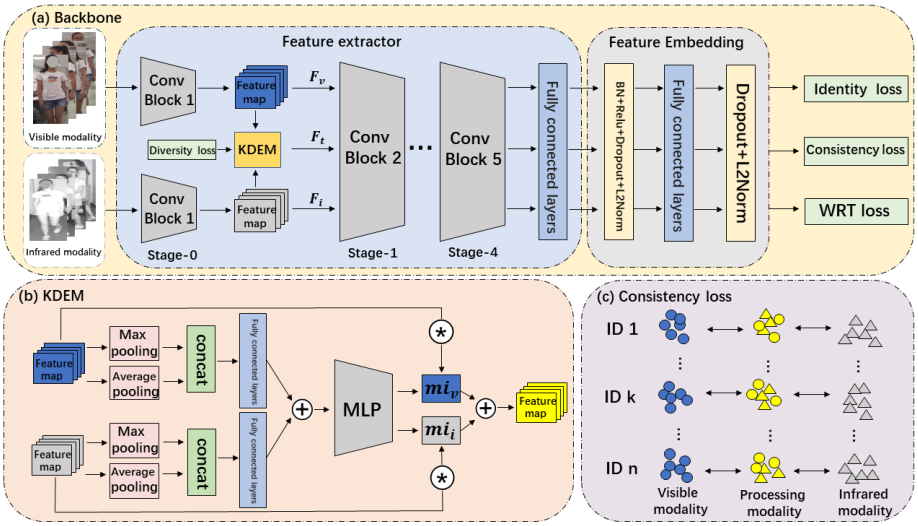


**Fig. 2.** (*a*) Illustration of our method. The visible and infrared modality features are mixed with KDEM to compose the feature of the knowledge-enhanced modality. The shared identity loss and WRT loss are enforced to help the model converge. (*b*) The concrete structure of KDEM. (*c*) The consistency loss is targeted at preserving the semantic correlation of the knowledge-enhanced modality which is similar to visible and infrared modalities. The circles and triangles represent the characteristic distribution of visible modality and infrared modality, respectively.

### 3.2 Feature extractor

The modality-specific features of the data samples are extracted through the stage-0 of the network, in which the weights are not shared. Moreover, the modality-shared features of the three modalities are extracted by the remaining stages of the weights-sharing network. In the architecture, the data samples of $n$ visible and $n$ infrared of the same identity are combined into $n$ pairs in a mini-batch and fed into the network. For the sample pair $(x_v, x_i)$, their modality-specific features is obtained by the stage-0 of feature extractor: $F_v$ and $F_i$,

$$F_v = f_0(x_v); F_i = f_0(x_i) \tag{1}$$

$f_0(\cdot)$ represents the stage-0 of ResNet-50. $x_v$ and $x_i$ are data samples from visible and infrared modalities. The $F_v$ and $F_i$ are mixed in KDEM (in section 3.3) for the purpose of obtaining the knowledge-enhanced modality feature $F_t$. Then, the feature representations of all the visible, infrared, and the knowledge-enhanced modality are fed into the following stage of the network. Furthermore, For each modality, the identity loss $\mathcal{L}_{id}$ and weight regularization triplets (WRT) loss $\mathcal{L}_{wrt}$[23] are applied to help the model converge effectively.

$$\mathcal{L}_{id} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p\left(y_i \mid C(f(x_i))\right)\right) \tag{2}$$

where $n$ represents the number of training samples in a mini-batch, $f(\cdot)$ and $c(\cdot)$ are the feature extractor and a classifier. Given an input image $x_i$ with a label $y_i$, $p(y_i|C(f(x_i)))$ represents the probability that a sample $x_i$ is correctly classified into labeled class $y_i$. Besides, we employ a WRT loss $\mathcal{L}_{wrt}$ on three-modality-shared features.

$$\mathcal{L}_{wrt}(i,j,k) = \log\left(1+\exp\left(w_i^p d_{ij}^p - w_i^n d_{ik}^n\right)\right) \tag{3}$$

$$w_{ij}^p = \frac{\exp\left(d_{ij}^p\right)}{\sum_{d_{ij}^p\in\mathcal{P}_i}\exp\left(d_{ij}^p\right)}, w_{ik}^n = \frac{\exp\left(-d_{ik}^n\right)}{\sum_{d_{ik}^n\in\mathcal{N}_i}\exp\left(-d_{ik}^n\right)} \tag{4}$$

The tuple $(i,j,k)$ represents the hard sample in each batch size. $\dot{i}_y$, $\mathcal{P}_i$ and $\mathcal{N}_i$ denote anchor, positive set and negative set respectively. $d_{ij}^p$ and $d_{ik}^n$ represent the distance between anchor and the positive and negative samples, respectively.

### 3.3 Knowledge-driven Enhance Module

In this section, we exceedingly elaborate on how the KDEM works. The modality-specific features $F_v$ and $F_i$ from stage-0 of the network are used as inputs for the KDEM, and synthesize the feature $F_p$ of knowledge-enhanced modality with the assistance of modality influence factors. As shown in Fig.2$(b)$, the modality influence factors are obtained by the following formula:

$$\alpha = \delta\left(MLP\left(\sum_{m\in\{v,i\}} FC\left([F_m^{avg};F_m^{max}]\right)\right)\right) \tag{5}$$

$\delta$ is the binary softmax function. After the Average-pooling and Max-pooling operations on the $F_v$ and $F_i$, the features are concatenated for each modality. $[F_v^{avg};F_v^{max}]$ and $[F_i^{avg};F_i^{max}]$ are defined for the visible modality and infrared modality, respectively. The aforementioned two pooling operations are combined to exclude redundant knowledge. After the following fully connected operations, these vectors are sent to MLP to calculate their influence factors.

Aforementioned procedures are designed to obtain the modality influence factors $\alpha = (mi_v,mi_i)$, where $mi_v + mi_i = 1$. $mi_v$ and $mi_i$ are the modality influence factors for the visible modality and the infrared modality, respectively. The feature of the knowledge-enhanced modality is obtained by mixing the visible and infrared modality-specific features with two modality influence factors. The feature of knowledge-enhanced modality $F_t$ is formulated by:

$$F_t = mi_v \cdot F_v + mi_i \cdot F_i \tag{6}$$

Here $F_v$ and $F_i$ are the modality-specific features from stage-0 of ResNet-50, which are obtained by Eq.1. The synthesis of $F_t$ can represent the newly generated knowledge-enhanced modality. By using adaptive modality influence factors, a suitable knowledge-enhanced modality can be generated, which can help the model decrease the pressure of the modal-shared feature learning and accumulate high distinguish semantic knowledge of two completely different modalities. The knowledge-enhanced modality only exists in the training stage and is deleted in the test stage.

Additionally, the characteristic of the knowledge-enhanced modality should be diversified as much as possible so that the network can better accumulate knowledge from the two modalities. To this end, diversity loss is designed to enlarge the standard deviation of the two modality influence factors in a mini-batch. The diversity loss $L_{div}$ is defined as follows:

$$L_{div} = -\left[\phi\left(mi_v\right)_{i=1}^{n} + \phi\left(mi_i\right)_{i=1}^{n}\right] \tag{7}$$

$n$ represents the number of training samples in a mini-batch. $\phi$ means calculating the standard deviation of the modality influence factors in a mini-batch. By minimizing $L_{div}$, the feature of the knowledge-enhanced modality becomes as diverse as possible, which can be more conducive to the knowledge-enhanced modality and modal-shared feature learning.

### 3.4 Consistency loss

Features of the modeled knowledge-enhanced modality should maintain semantic similarity between visible and infrared modalities to avoid the accumulation of redundant knowledge. Thus, a consistency loss is designed, as shown in Fig.2(c). For the features of the same identity, the feature distributional similarity between the knowledge-enhanced modality and the other two modalities should be preserved. The consistency loss is represented as follows:

$$L_{con} = \frac{1}{n}\sum_{i=1}^{n}\sum_{m\in\{v,i\}} mi_m \cdot \left\|f\left(F_m\right) - f\left(F_t\right)\right\|_2 \tag{8}$$

The modality influence factors $mi_m$ are the weights of the consistency loss function. $\|:\|_2$ is a L2-norm which is used to measure the distance of features. $F_m$ is obtained by Eq.1. $F_t$ is the feature of knowledge-enhanced modality and obtained in Eq.6. The $f(\cdot)$ is the mapping from stage-1 to stage-4 of the backbone. The gradients of $L_{con}$ for feature $F_m$ and $F_p$ can be directly concluded as follows.

*Proof.* This subsection proves that $L_{con}$ can find the derivative.

$$\begin{aligned}\frac{\partial L_{con}}{\partial F_t} &= \frac{\partial L_{con}}{\partial f\left(F_p\right)} \cdot \frac{\partial f\left(F_t\right)}{\partial F_t} \\ &= \left\{-2 \cdot mi_v \cdot \left[f\left(F_v\right) - f\left(F_t\right)\right] - 2 \cdot mi_i \cdot \left[f\left(F_i\right) - f\left(F_t\right)\right]\right\} \cdot f^{'}\left(F_t\right) \\ &= \left\{-2\left[mi_v \cdot f\left(F_v\right) + mi_i \cdot f\left(F_i\right) - f\left(F_t\right)\right]\right\} \cdot f^{'}\left(F_t\right)\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial L_{con}}{\partial F_m} &= \frac{\partial L_{con}}{\partial f\left(F_m\right)} \cdot \frac{\partial f\left(F_m\right)}{\partial F_m} \\ &= \left\{2 \cdot mi_m \cdot \left[f\left(F_m\right) - f\left(F_t\right)\right]\right\} \cdot f^{'}\left(F_m\right)\end{aligned} \tag{10}$$

### 3.5 Overall training

The overall training loss $L_{all}$ is represented as follows:

$$\mathcal{L}_{all} = \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{wrt} + \lambda_3 \mathcal{L}_{div} + \lambda_4 \mathcal{L}_{con} \tag{11}$$

where $\lambda_1, \lambda_2$ , $\lambda_3$ and $\lambda_4$ are the weights of the overall loss function. The algorithm flow of KDEM are given in Algorithm.1

**Algorithm 1** algorithm of KDEM.

**Input:** Visible sample $x_v$ and infrared sample $x_i$ with their labels;
**Output:** The trained feature extractor $f(\cdot)$ and classifier $c(\cdot)$;
**Initialization:** Initialize the network $f(\cdot)$ in ImageNet-pretrained ResNet-50, initialize iteration number *epoch*, learning rate and other hyper-parameters.
**for** $i = 1$ to *epoch* **do**
- ·Use stage-0 of ResNet-50 to extract features $F_v$ and $F_i$ for the visible and infrared modalities;
- ·Calculate the knowledge-enhanced modality feature $F_p$ by Eq.6;
- ·Feed forward the batch into the following network
- ·Calculate the overall loss by Eq.11 .
- ·Update the network $f(\cdot)$ and classifier $c(\cdot)$ by SGD to descending gradients of Eq.11.

**end**
**Until** model convergence or the fixed *epoch*;

## 4 Experiments

### 4.1 Datasets and Settings

Two available datasets are used to evaluate the performance of our proposed KDEM: RegDB[4] and SYSU-MM01[19]. The experiments follow the evaluation protocol as described in Ye et.al[23]. The RegDB dataset is divided into 206 training identities and 206 test identities, and the number of both visible and infrared images is ten for each identity. SYSU-MM01 is split into 395 identities and 96 identities. The former of each dataset is used for training, while the latter is used for testing.

The proposed method is implemented with PyTorch. Furthermore, the initial learning rate and optimization method are 0.1 and SGD. The $\lambda_1, \lambda_2$ and $\lambda_3$ from Eq.11 are set to be 1 and $\lambda_4$ is set to be 0.1. The batch size for each modality is set to be eight on one single TITAN Xp GPU. The training epoch is set to be

80. The evaluation protocol adopts the Rank-1, 10, 20 accuracy, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP).

### 4.2 Comparison with State-of-the-art Methods

As shown in Table.1 and Table.2, we made an objective comparison of our method with the state-of-the-art. The Rank-1, 10, 20 accuracy(%) and mean average precision (mAP)(%), and mean Inverse Negative Penalty (mINP)(%) are reported in each table.

**Table 1.** Comparison with the state-of-the-art methods on RegDB dataset.

| Method | Venue | Visible to Infrared | | | | | Infrared to Visible | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rank-1 | Rank-10 | Rank-20 | mAP | mINp | Rank-1 | Rank-10 | Rank-20 | mAP | mINp |
| BDTR[22] | IJCAI(2018) | 33.56 | 58.61 | 67.43 | 32.76 | - | 32.92 | 58.46 | 68.43 | 31.96 | - |
| D$^2$*RL*[17] | CVPR(2019) | 43.4 | 66.1 | 76.3 | 44.1 | - | - | - | - | - | - |
| AlignGAN[16] | ICCV(2019) | 57.9 | - | - | 53.6 | - | 56.3 | - | - | 53.4 | - |
| Xmodal[9] | AAAI(2020) | 62.21 | 83.13 | 91.72 | 60.18 | - | - | - | - | - | - |
| DDAG[24] | ECCV(2020) | 69.34 | 86.19 | 91.49 | 63.46 | 49.24 | 68.06 | 85.15 | 90.31 | 61.80 | 48.62 |
| Hi-CMD[2] | CVPR(2020) | 70.93 | 86.39 | - | 66.04 | - | - | - | - | - | - |
| AGW[23] | TAPMI(2021) | 70.05 | 86.21 | 91.55 | 66.37 | 50.19 | 70.04 | 87.12 | 91.84 | 65.90 | 51.24 |
| FBP-AL[18] | TNNLS(2021) | 73.98 | **89.71** | **93.69** | 68.24 | - | 70.05 | **89.22** | **93.88** | 66.61 | - |
| cmAlign[11] | ICCV(2021) | 74.17 | - | - | 67.64 | - | 72.43 | - | - | 65.46 | - |
| **KDEM(Ours)** | - | **77.33** | 88.25 | 91.70 | **70.32** | **56.08** | **76.26** | 87.62 | 90.92 | **67.77** | **52.38** |

**Performance Analysis on RegDB.** The evaluation results on RegDB show that KDEM achieves the most advanced performance in terms of Rank-1 accuracy, mAP, and mINP. Although the FBP-AL[18] shows better results in terms of Rank-10, 20 accuracy, the FBP-AL[18] needs to segment the body structure additionally, which is not efficient for model training. Compared to cmAlign[11], under the test mode of Visible to Infrared, the performance of KDEM is improved by 3.16% and 2.68% in terms of Rank-1 and mAP, respectively. As for the test mode of Infrared to Visible, our KDEM can also increase by 3.83% and 2.31% in performance. The performance improvement indicates that our model can effectively reduce the semantic gap in cross modalities by the generated knowledge-enhanced modality.

**Table 2.** Comparison with the state-of-the-art methods on SYSU-MM01 dataset.

| Method | Venue | All search | | | | | Indoor search | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rank-1 | Rank-10 | Rank-20 | mAP | mINp | Rank-1 | Rank-10 | Rank-20 | mAP | mINp |
| cmGAN[12] | IJCAI(2018) | 27.0 | 67.5 | 80.6 | 27.8 | - | 31.6 | 77.2 | 89.2 | 42.2 | - |
| D$^2$*RL*[17] | CVPR(2019) | 28.9 | 70.6 | 82.4 | 29.2 | - | - | - | - | - | - |
| AlignGAN[16] | ICCV(2019) | 42.40 | 85.0 | 93.7 | 40.7 | - | 45.9 | 87.6 | 94.4 | 54.3 | - |
| Xmodal[9] | AAAI(2020) | 49.92 | 89.79 | 95.96 | 50.73 | - | - | - | - | - | - |
| DDAG[24] | ECCV(2020) | 54.75 | 90.39 | 95.81 | 53.02 | 39.62 | 61.02 | 94.06 | 98.40 | 67.98 | 62.61 |
| Hi-CMD[2] | CVPR(2020) | 34.94 | 77.58 | - | 35.94 | - | - | - | - | - | - |
| AGW[23] | TAPMI(2021) | 47.50 | 84.39 | 62.14 | 47.65 | 35.30 | 54.17 | 91.14 | 95.98 | 62.97 | 59.23 |
| FBP-AL[18] | TNNLS(2021) | 54.14 | 86.04 | 93.03 | 50.20 | - | - | - | - | - | - |
| cmAlign[11] | ICCV(2021) | 55.41 | - | - | 54.14 | - | 58.46 | - | - | 66.33 | - |
| **KDEM(Ours)** | - | **58.09** | **91.19** | **96.63** | **55.52** | **40.69** | **62.18** | **94.38** | **98.64** | **68.30** | **64.11** |

**Performance Analysis on SYSU-MM01.** It can be seen from the Table.2 that KDEM achieves a new state-of-the-art performance on SYSU-MM01 in both all search and indoor search modes. Compared to cmAlign[11] in All search mode, our KDEM gains 2.68% and 1.38% in Rank-1 and mAP. As for the Indoor search, the performance of KDEM can also be improved by 3.72% and 1.97% on Rank-1 and mAP. The class activation mapping [14] of our model is shown in Fig.3. We can notice that our model mainly focuses on cross-modal invariant features, such as the face, clothing logos, backpacks, and gait, which are significant or noticeable indicators of one pedestrian's identity.

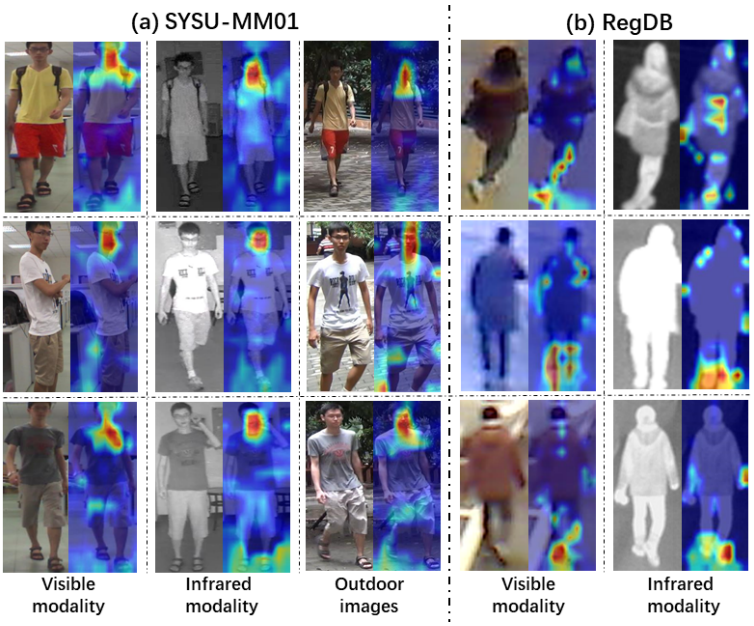


**Fig. 3.** Class Activation Mapping of our model.

### 4.3 Ablation Study

Ablation experiments are performed on the RegDB dataset to verify the effectiveness of KDEM. As shown in Table.3, superior results have been achieved by the proposed KDEM, diversity loss, and consistency loss. A remarkable improvement in performance is obtained by KDEM, which shows that the KDEM can effectively overcome the cross-modal semantic gap by accumulating the high distinguish knowledge from cross modalities. Moreover, the improvement of the loss functions is prominent. The diversity loss can help the model accumulate diverse knowledge and exclude redundant knowledge of visible and infrared modalities. The consistency loss aims to preserve the distribution and semantic similarity between the knowledge-enhanced modality and the other two modalities. Both of $L_{div}$ and $L_{con}$ achieved good results.

**Table 3.** Ablation Study on RegDB dataset.

| Baseline | KDEM | $L_{div}$ | $L_{con}$ | Rank-1 | Rank-10 | Rank-20 | mAP | mINP |
|---|---|---|---|---|---|---|---|---|
| √ | - | - | - | 70.05 | 86.21 | 91.55 | 66.37 | 50.19 |
| √ | √ | - | - | 73.69 | 86.84 | 91.61 | 68.36 | 54.59 |
| √ | √ | √ | - | 76.41 | 87.55 | 91.68 | 69.40 | 54.85 |
| √ | √ | - | √ | 75.29 | 87.33 | 91.65 | 69.08 | 55.19 |
| √ | √ | √ | √ | **77.33** | **88.25** | **91.70** | **70.32** | **56.08** |

## 5 Conclusion

A Knowledge-driven Enhance Module (KDEM), which imitates the cognitive process of the human brain, is designed to tackle the difficulties of VI-ReID task. Extensive experiments are performed on two popular datasets to evaluate the performance of our KDEM, and KDEM obtained competitive performance compared to state-of-the-art methods. Meanwhile, ablation studies demonstrate that the architecture of KDEM can discover significant semantic knowledge from cross modalities and integrate them into a knowledge-enhanced modality to robust the supervision of feature representations learning. Moreover, the diversity loss can effectively improve the variety of semantic knowledge in the knowledge-enhanced modality, and the consistency loss can also preserve the semantic correlation between the knowledge-enhanced modality and the other two modalities.

## References


1. Chen, S., Wu, S., Wang, L.: Hierarchical semantic interaction-based deep hashing network for cross-modal retrieval. PeerJ Computer Science **7**(2), e552 (2021)
2. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 10254–10263. IEEE (2020) 9
3. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. pp. 677–683. International Joint Conferences on Artificial Intelligence Organization (2018)
4. Dat, N., Hong, H., Ki, K., Kang, P.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017) 8
5. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE Trans. Image Process. **29**, 579–590 (2020)
6. Gao, S., Wang, J., Lu, H., Liu, Z.: Pose-guided visible part matching for occluded person reid. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
7. Han, X.F., Laga, H., Bennamoun, M.: Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. IEEE (2016) 1, 4

9. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20). pp. 4610–4617 (2020) 9
10. Mao, X., Li, Q., Xie, H.: Aligngan: Learning to align cross-domain images with conditional generative adversarial networks (2017)
11. Park, H., Lee, S., Lee, J., Ham, B.: Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences (2021) 2, 4, 9, 10
12. Pingyang, D., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. pp. 677–683 (07 2018). 9
13. Pu, N., Chen, W., Liu, Y., Bakker, E.M., Lew, M.S.: Dual Gaussian-Based Variational Subspace Disentanglement for Visible-Infrared Person Re-Identification, p. 2149–2158. Association for Computing Machinery, New York, NY, USA (2020) 1
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision (2017) 10
15. Wang, G.A., Yang, T., Cheng, J., Chang, J., Liang, X., Hou, Z.: Cross-modality paired-images generation for rgb-infrared person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence (2020) 4
16. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 3622–3631. IEEE (2019) 4, 9
17. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 9
18. Wei, Z., Yang, X., Wang, N., Gao, X.: Flexible body partition-based adversarial learning for visible infrared person re-identification. IEEE Transactions on Neural Networks and Learning Systems **PP**(99) (2021) 9
19. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2, 4, 8
20. Xu, X., Wu, S., Liu, S., Xiao, G.: Cross-modal based person re-identification via channel exchange and adversarial learning. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds.) Neural Information Processing - 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8-12, 2021, Proceedings, Part I. Lecture Notes in Computer Science, vol. 13108, pp. 500–511. Springer (2021) 1
21. Xzab, D., Xw, A., Emb, C., Song, W.A.: Multi-label semantics preserving based deep cross-modal hashing. Signal Processing: Image Communication (2021)
22. Ye, M., Lan, X., Wang, Z., Yuen, P.C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE transactions on information forensics and security **15**, 407–419 (2020) 1, 9
23. Ye, M., Shen, J., Lin, G., Xiang, T., Hoi, S.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2021) 4, 5, 8, 9
24. Ye, M., Shen, J., Crandall, D.J., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII. Lecture Notes in Computer Science, vol. 12362, pp. 229–247. Springer (2020) 2, 4, 9