Visible-Infrared Person Re-Indentification via Feature Fusion and Deep Mutual Learning

Ziyang Lin 2112205250@mail2.gdut.edu.cn
and **Banghai Wang*** bhwang@gdut.edu.cn
School of Computer Science and Technology, Guangdong University of Technology

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Visible-Infrared Person Re-Identification (VI-ReID) aims to retrieve a set of person images captured from both visible and infrared camera views. Addressing the challenge of modal differences between visible and infrared images, we propose a VI-ReID network based on Feature Fusion and Deep Mutual Learning (DML). To enhance the model's robustness to color, we introduce a novel data augmentation method called Random Combination of Channels (RCC), which generates new images by randomly combining R, G, and B channels of visible images. Furthermore, to capture more informative features of individuals, we fuse the features from the middle layer of the network. To reduce the model's dependence on global features, we employ a fusion branch as an auxiliary branch, facilitating synchronous learning of global and fusion branches through Deep Mutual Learning . Extensive experiments on the SYSU-MM01 and RegDB datasets validate the superiority of our method, showcasing its excellent performance when compared to other state-of-the-art approaches.

Keywords: Visible-Infrared Person Re-Indentification, Cross-modality, Feature Fusion, Data Augmentation, Deep Mutual Learning.

1. Introduction

Person Re-identification (Re-ID) is directed towards retrieving images of the target person with the same identity from multiple unconnected cameras. The majority of existing Re-ID methods mainly concentrate on the matching among visible images which are typically collected under favorable lighting conditions. Nevertheless, in environments with inadequate lighting, like at night, visible images are unable to offer sufficient information. Due to this, VI-ReID has come into being as an alternative solution for the search between visible and infrared images.

However, there are large intra-class differences and modality discrepancies between visible and infrared images. The intra-class differences are the appearance differences within a certain mode caused by posture, clothing, angle and other factors, and the modality discrepancies are the internal differences between visible and infrared images caused by spectrum.

The researchers put forward a series of methods to narrow the modal gap between visible and infrared images. One approach is to extract representations from visible and infrared images and subsequently utilize feature-level constraints [20; 18] to align the feature

* Corresponding author.

distributions, thereby learning modal-invariant features. However, the aforesaid method has a tendency to overlook crucial discriminatory information, leading to a reduction in the accuracy of the model.

Another approach is to eliminate color differences. Several studies presently employ generative adversarial networks (GANs) [16; 3] to generate cross-modal images that can bridge the gap. However, the generation process frequently requires extra costs and creates inevitable noise. Other studies have directly used grayscale images for cross-modal matching, which eliminates color differences but also loses discriminant information in the color channel.

In order to solve the above problems, we proposes an end-to-end dual-flow cross-modal VI-ReID network to improve the cross-modality recognition performance. The main contributions can be summarized as follows.

1) A novel data augmentation method named Random Combination of Channels (RCC) is proposed to enhance the robustness of color and channel differences. The core concept is randomly combining R, G and B channels of visible images to generate new images to expand the dataset. Experiments have shown that this approach can improve performance without introducing huge amounts of computation.

2) The features of the middle layer of the network are fused in order to acquire more discriminatory features. Feature fusion effectively boosts the information of the person with discriminating appearance cues, thereby enhancing the performance.

3) In order to compensate for the loss of feature information in the global branch, the global and fusion branches are concatenated into a new branch, named joint branch. By synchronous learning of multiple branches through deep mutual learning, the network's ability to obtain more discriminative features is improved.

2. Related work

2.1. Visible-Infrared Person Re-Indentification

The use of infrared cameras in surveillance systems is common for capturing high-quality pedestrian imagery in low-light conditions. However, the challenge arises when applying single-modal Re-ID methods directly, as modality differences can lead to decreased recognition accuracy. VI-ReID plays a crucial role in addressing this issue by enabling identity matching between visible and infrared images, thereby improving recognition accuracy and overall system performance in surveillance applications.

In 2017, Wu et al. [15] publicly presented a large-scale VI-ReID dataset named SYSU-MM01 and proposed a zero-padding strategy. TSLFN [29] horizontally divides the global feature mapping into multiple stripes and employs local constraints for each stripe, , thereby enhancing recognition accuracy. Ye et al. [21] proposed a dynamic dual-attentive aggregation (DDAG) learning, which utilizes dynamic dual-attentive aggregation to improve model performance by leveraging attention-aware functions. Additionally, other studies have explored adversarial training strategies, such as D2RL[14] and cmGAN[3], to reduce cross-modal differences at the image level in VI-ReID tasks.

In recent research, Zhang et al.[26] introduced FMCNet, a network that addresses missing modal information at feature level, outperforming models that rely on image-level compensation. Neural feature search (NFS) [1] explores a method to automate feature selection

processes. Liu et al. [9] proposed the spectrum-aware feature augmentation network (SFA-Net), a dual-stream network that processes visible and infrared images simultaneously to enhance feature alignment and capture human cross-modal features effectively. Liu et al. [10] proposed AGMNet, which uses a CycleGAN [28] to normalize the grayscale of infrared images to generate a unified intermediate mode to overcome the modal differences.

2.2. Data augmentation

Data augmentation is a crucial technique widely used in various computer vision tasks to prevent overfitting of training data. Its primary purpose is to enhance the diversity of training data, thereby improving the model's generalization ability. By introducing variations and perturbations to the training data, data augmentation helps the model learn robust features and patterns.

Traditional data augmentation methods include random flipping, rotation, random erasing, cropping, warp scaling, adding noise, color jitter. Color jitter improves an image's colors by changing its brightness, contrast, and saturation, making the model less sensitive to color changes. Channel-level random erasing (CRE)[22] randomly select rectangular regions in the training image and replace their pixel values with random values from all three channels to simulate uncertain occlusion for enriching the variety of training samples.

Another research is based on image blending methods, including global image blending, such as MixUp [25], and local image blending, such as CutMix[24]. These methods synthesize virtual training samples by linearly interpolating the images and corresponding labels, resulting in smoother decision boundaries and reducing overfitting of the training data.

2.3. Deep Mutual Learning

Knowledge Distillation[6], first proposed by Hinton in 2014, involves transferring knowledge from a complex model (teacher) to a simpler model (student). However, traditional knowledge distillation faces challenges in VI-ReID due to the difficulty in creating a high-accuracy teacher model and the student model's limited learning from the teacher.

Zhang et al.[27] proposed Deep Mutual Learning, where the network serves as both teacher and student. This collaborative approach enables mutual learning and teaching between models during training.

3. Proposed method

In this section, we introduce the structure of the proposed VI-ReID network which is shown in Fig. 1. It mainly includes the following four modules: data augmentation module, feature extraction module, feature fusion module and deep mutual learning module. Next, we will introduce them in detail.

3.1. Data augmentation module

The Random Combination of Channels method is able to expand the data without introducing huge computational overhead. As shown in Fig. 2, its main idea is to disrupt the order of channels (R, G, or B), including R-G-B, R-B-G, B-G-R, B-R-G, G-B-R and G-R-B. In addition, the visible image is converted into a grayscale image. Randomly select

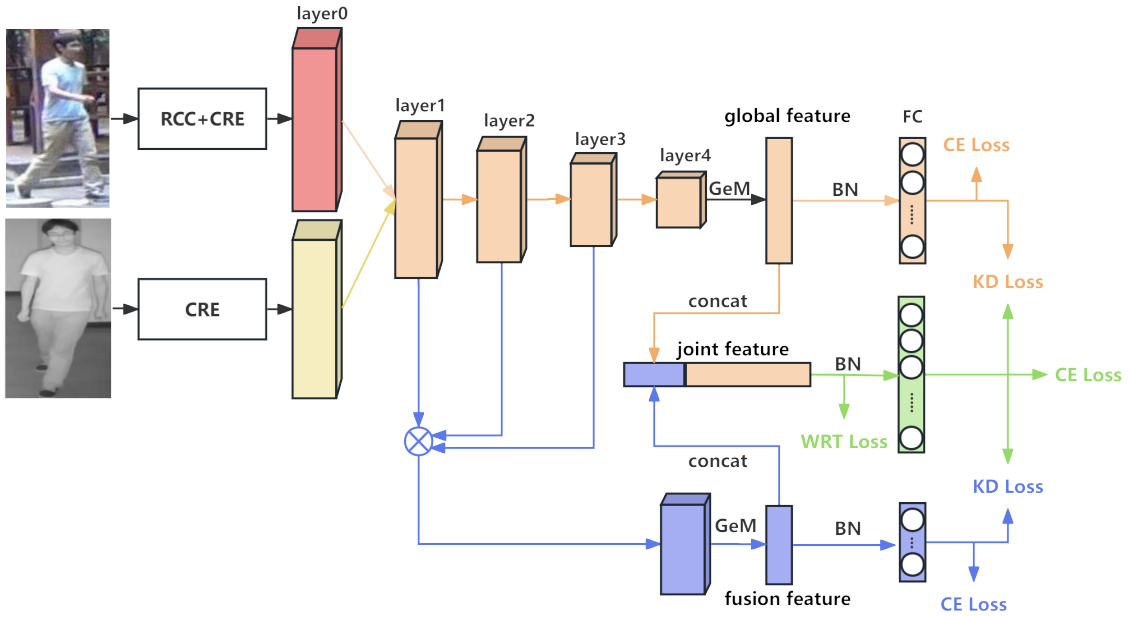

Figure 1: Framework of the proposed Network for Visible-Infrared Person Re-Identification

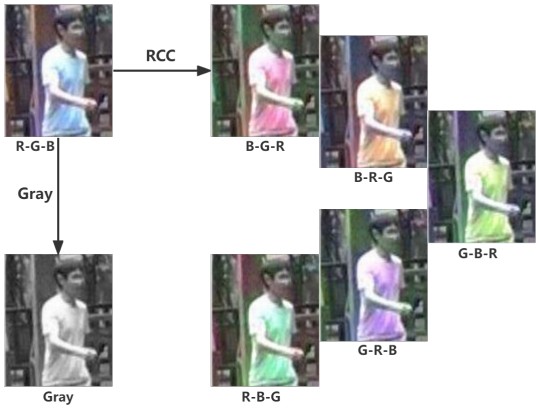

Figure 2: Illustration of the Random Combination of Channels(RCC)

one of the seven images mentioned above to replace the visible image. Integrating this method with other fundamental data augmentation techniques like random flipping, random erasing, and random cropping can enhance the model's performance with minimal additional computational cost. Compared with color jitter, RCC focuses more on the channels of image to enhance model's robustness of the differences between channels, which is beneficial for processing cross-modality images.

3.2. Feature extraction module

Global features are extracted using a ResNet-50[5] network that includes modal-specific and modal-shared modules. Specifically, the cross-modal datasets after data augmentation are input into the network separately. After passing through the modal-specific module

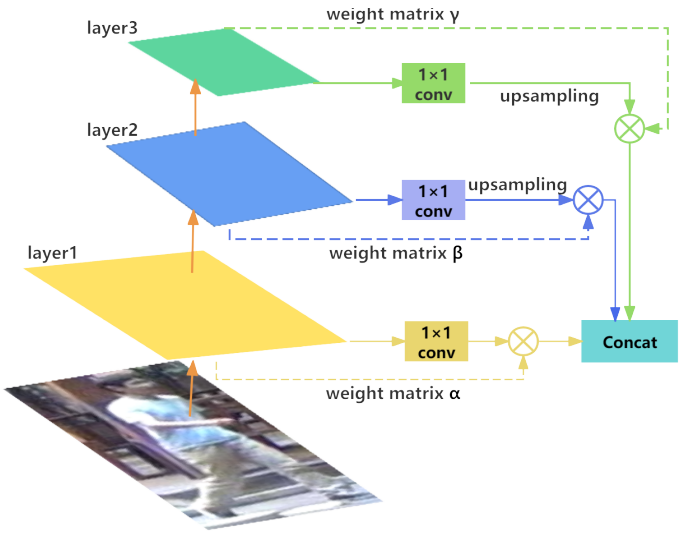

Figure 3: Framework of the Feature Fusion

$layer_0$, the extracted features are spliced separately to obtain f_0. The f_0 is then input into the modal sharing module,and the $layer_i$ represents the i-th layer of the Resnet-50. Therefore, the output of $layer_i$ can be represented as

$$f_i = layer_i(f_{i-1}), i = 1, 2, 3, 4. \tag{1}$$

the final output of resnet-50 is f_4, expressed in f_{global}.

To reduce the amount of computation, GeM pooling is used to transform a 3D feature f_{global} into a 1D embedding vector V^{global}.

3.3. Feature fusion module

Color and texture are often considered shallow features of an image. In contrast, deep features are advanced features learned from data through methods such as neural networks, usually including the shape, posture, angle, etc of objects. Feature fusion of shallow and deep features can improve the performance of the model since it can obtain more critical information. For the fusion of feature maps of different scales, directly using traditional feature fusion methods may result in the loss of some important information. Adaptively spatial feature fusion (ASFF)[11] is different from the previous multi-level feature fusion method based on elements. Its core idea is to adaptively learn the spatial weights of feature map fusion at various scales to achieve optimal feature fusion.

The feature f_1, f_2 and f_3 output from $layer_1$, $layer_2$ and $layer_3$ of the backbone network are fused. As shown in Fig. 3, since f_1, f_2 and f_3 have different resolutions and different channel counts, we make them the same through corresponding operations.

Specifically, channel numbers of f_1, f_2 and f_3 are reduced to 256 by 1×1 convolution and obtain $f_1^{'}$, $f_2^{'}$ and $f_3^{'}$. Afterwards, perform double upsampling on $f_2^{'}$ and quadruple upsampling on $f_3^{'}$ to obtain $f_2^{''}$ and $f_3^{''}$, so that they have the same resolutions and channel counts. The weight parameter matrix α, β and γ consistent with the resolution of the feature map $f_1^{'}$, $f_2^{''}$ and $f_3^{''}$ are multiplied point-by-point with them, respectively. The

values of each position in these weight matrixes are obtained by network training. Merge the multiplied features to obtain the fusion features f_{fusion}, calculated as

$$f_{fusion} = \left(\alpha \otimes f_1^{'}\right) \oplus \left(\beta \otimes f_2^{''}\right) \oplus \left(\gamma \otimes f_3^{''}\right). \tag{2}$$

where $\otimes$ represents point multiplication and $\oplus$ means concatenate method.

Similarly, GeM pooling is used to transform 3D feature f_{fusion} into 1D embedding vector V^{fusion}.

3.4. Deep Mutual Learning module

We merge the embedding vectors of the global and fusion branches into a new branch to obtain the joint feature, that is

$$V^{joint} = V^{global} \oplus V^{fusion}. \tag{3}$$

V^{joint} is used as the final representation for this task. Combine global and fusion information into high-dimensional information to generate prediction results. The predicted results are used as teacher signal, transmitted to specific branches to guide their learning process. Specific branch will adjust their learning objectives to align their outputs as closely as possible with teacher signal. As shown in Fig.1, without introducing additional pre-trained teacher model, the joint branch is used as an intermediary to conduct deep mutual learning with global branch and fusion branch respectively in order to achieve synchronous learning of the three branches.

Specifically, using softmax function for global, fusion and joint branches to obtain the posterior probabilities p_i^g, p_i^f and p_i^j which will be used as soft target to achieve mutual supervision and learning.

Formally, given N training samples defined as $\{x_i\}_i^N$ from M classes. x_i represents the person images. N represents the number of samples and i means the i-th sample. We define the corresponding label set as $\{y_i\}_i^N$ with $y_i \in \{1, 2, \ldots, M\}$. The posterior probability p_i^g can be obtained as follows

$$p_i^g = p\left(y_i \mid x_i\right) = \frac{\exp\left(W_i \times V_i^{global}\right)}{\sum_{m=1}^{M} \exp\left(W_m \times V_i^{global}\right)}. \tag{4}$$

where W_m is the weight parameter matrix of the last fully connected (FC) layer for m-th identity. p_i^f and p_i^j can be obtained through the same method.

Kullback-Leibler (KL) Divergence is a metric used to measure the similarity between two probability distributions. For global branch, using KL Divergence for p^j and p^g, as well as p^g and p^j to obtain the $D_{KL}\left(p_i^j \parallel p_i^g\right)$ and $D_{KL}\left(p_i^g \parallel p_i^j\right)$. For fusion branch, using the same method to obtain $D_{KL}\left(p_i^j \parallel p_i^f\right)$ and $D_{KL}\left(p_i^f \parallel p_i^j\right)$. We use L_{KD_1} and L_{KD_2} to reduce the distance between two predicted distributions, thereby achieving synchronous

learning of multi branches. The formulas for L_{KD_1} and L_{KD_2} are as follows

$$\begin{aligned} L_{KD_1} &= \frac{1}{2}\left[D_{KL}\left(p_i^j \parallel p_i^g\right) + D_{KL}\left(p_i^g \parallel p_i^j\right)\right] \\ &= \frac{1}{2}\left[\sum_{i=1}^{N} p_i^j\left(\ln p_i^j - ln p_i^g\right) + \sum_{i=1}^{N} p_i^g\left(\ln p_i^g - ln p_i^j\right)\right]. \end{aligned} \tag{5}$$

$$\begin{aligned} L_{KD_2} &= \frac{1}{2}\left[D_{KL}\left(p_i^j \parallel p_i^f\right) + D_{KL}\left(p_i^f \parallel p_i^j\right)\right] \\ &= \frac{1}{2}\left[\sum_{i=1}^{N} p_i^j\left(\ln p_i^j - ln p_i^f\right) + \sum_{i=1}^{N} p_i^f\left(\ln p_i^f - ln p_i^j\right)\right]. \end{aligned} \tag{6}$$

3.5. Loss function

Identity loss. Cross Entropy (CE) loss is widely used in classification tasks to measure the difference between the true value and the model's predicted value.

The identity loss of global branch is computed as

$$L_{id}^{global} = -\frac{1}{N}\sum_{i=1}^{N} \log p_i^g. \tag{7}$$

where N is the total number of training samples.

Similarly, the expression for the identity loss of the fusion branch and joint feature branch are as follows

$$L_{id}^{fusion} = -\frac{1}{N}\sum_{i=1}^{N} \log p_i^f, L_{id}^{joint} = -\frac{1}{N}\sum_{i=1}^{N} \log p_i^j. \tag{8}$$

The L_{KD_1} and L_{KD_2} are taken as label smoothing regularization terms for L_{id}^{global} and L_{id}^{fusion}, respectively. The final identity loss of the global and fusion branches are as follows

$$\tilde{L}_{id}^{global} = L_{id}^{global} + \theta_1 L_{KD_1}. \tag{9}$$

$$\tilde{L}_{id}^{fusion} = L_{id}^{fusion} + \theta_2 L_{KD_2}. \tag{10}$$

where θ_1 and θ_2 is the predefined weight coefficient, which will be analyzed in 4.4.

Weighted Regularization Triplet (WRT) loss. The joint feature are trained with the WRT loss which inherits the advantages of optimizing the relative distance between positive and negative pairs and avoids introducing additional margin parameters. WRT loss is computed as

$$L_{WRT}(i,j,k) = \ln\left(1 + \exp(w_i^p d_{ij}^p - w_i^n d_{ik}^n)\right). \tag{11}$$

$$w_i^p = \frac{\exp(d_{ij}^p)}{\sum_{d^p \in P}(d^p)}, w_i^n = \frac{\exp(d_{ik}^n)}{\sum_{d^n \in N}(d^n)}. \tag{12}$$

where N is the total number of training samples, (i, j, k) represent the triplets within each training batch, P and N are the positive and negative sample set, d_{ij}^{p} and d_{ik}^{n} represent the Euclidean distance between positive and negative samples.

Total loss. Combining these individual losses, we finally define the total loss for the overall network as follows

$$L_{total} = \tilde{L}_{id}^{global} + \tilde{L}_{id}^{fusion} + L_{id}^{joint} + L_{WRT}. \tag{13}$$

4. Experiments and analysis

4.1. Dataset and Evaluation Metric

Dataset. In this work, we use two public datasets, SYSU-MM01 [15] and RegDB [13] to evaluate the performance of proposed model.

SYSU-MM01 dataset is a large cross-modality dataset collected by Sun Yat-sen University. It contains images taken by six cameras (two infrared cameras and four visible cameras), including indoor and outdoor environments. The dataset contains a total of 30,071 visible images and 15,792 infrared images of 491 individual identities. In this work, We conducted experiments in two different evaluation modes, namely All-search and Indoor-search modes. For All-search mode, 3803 IR images from cameras 3 and 6 are used for querying. For Indoor-search, images taken by only two indoor cameras are used.

RegDB dataset consists of 8240 images with 412 identities, of which 206 identities are used for training and the rest for testing. For each person, there are 10 visible and infrared images. The dataset has two modes: visible images query infrared images and infrared images query visible images.

Evaluation Metric. Following the previous works [19], Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) are used as evaluation metrics. In addition, we also introduced mean Inverse Negative Penalty (mINP) [23] metric to measure the retrieval performance of the model. CMC is a metric used to evaluate the performance of retrieval systems, which measures the performance by considering the correct match rate within the top n positions (rank-n) of the retrieval results.

4.2. Experimental Settings

The experiments are deployed on an NVIDIA GeForce 3080 GPU with Pytorch. All input images are adjusted to 288×144. These images are data augmented by RCC and CRE[22].

The total number of training epochs is 80, the batch size is set to 64. For each mini-batch, 8 identities are randomly sampled, each identity contains 4 visible images and 4 infrared images. When training on SYSU-MM01 and RegDB datasets, the initial learning rate is 0.1 and reduce by 0.1 and 0.01 times at 20, 50 epochs. An SGD optimizer with a weight attenuation of 5×10^{-4} and a momentum of 0.9 are used to update the parameters of the network. The weight parameters θ_1 and θ_2 are set to 0.6 and 1.8.

Table 1: Evaluate our methods under All-search mode on SYSU-MM01 dataset

Methods					All-search				
Baseline	RCC	CRE	Fusion	DML	r=1	r=10	r=20	mAP	mINP
✓	-	-	-	-	49.86	86.98	93.70	46.55	33.10
✓	-	-	✓	-	54.12	89.90	95.15	52.99	39.80
✓	✓	-	-	-	56.87	91.64	96.70	54.30	39.89
✓	-	✓	-	-	59.48	92.29	96.43	59.16	43.82
✓	✓	✓	-	-	65.45	94.55	97.96	60.78	44.92
✓	✓	✓	✓	-	69.68	96.28	98.87	65.66	50.86
✓	✓	✓	✓	✓	72.57	96.60	98.89	68.61	54.64

Table 2: Ablation experimental of feature fusion at different layers.

Settings				All-search				
layer1	layer2	layer3	layer4	r=1	r=10	r=20	mAP	mINP
✓	✓	-	-	66.01	94.49	98.19	62.30	47.68
✓	-	✓	-	71.58	95.90	98.72	67.19	52.55
-	✓	✓	-	72.39	96.64	99.02	68.10	53.47
✓	✓	✓	-	**72.57**	**96.60**	**98.89**	**68.61**	**54.64**
✓	✓	-	✓	67.59	95.09	98.31	65.39	52.52
✓	-	✓	✓	70.21	95.82	98.66	66.83	53.66
-	✓	✓	✓	71.08	96.07	98.84	67.43	53.96
✓	✓	✓	✓	71.69	96.15	98.87	67.52	54.39

Table 3: Evaluation for global and fusion branches on SYSU-MM01 dataset using different classification loss. L_{id} means L_{id}^{global} and L_{id}^{fusion}. L_{KD} means L_{KD_1} and L_{KD_2}. LSR denotes label smoothing regularization.

Settings			All-search				
L_{id}	L_{KD}	LSR	r=1	r=10	r=20	mAP	mINP
-	-	-	69.59	96.06	98.74	65.36	50.40
✓	-	-	69.68	96.28	98.87	65.66	50.86
-	✓	-	68.37	95.75	98.56	65.32	50.35
✓	✓	-	71.24	96.02	98.82	67.38	53.21
✓	✓	✓	**72.57**	**96.60**	**98.89**	**68.61**	**54.64**

Table 4: Evaluation the application of Deep Mutual Learning across different branches. None means not using DML. Global-to-Fusion represents the knowledge transfer from the global branch to the fusion branch.

Settings	All-search				
	r=1	r=10	r=20	mAP	mINP
None	69.68	96.28	98.87	65.66	50.86
Global-to-Fusion	68.67	94.76	97.84	64.31	49.44
Joint-to-Global	70.18	95.59	98.48	65.95	51.31
Joint-to-Fusion	70.87	95.68	98.49	66.78	52.38
Joint-to-Global&Fusion	72.57	96.60	98.89	68.61	54.64

4.3. Ablation experiment

In order to verify the effectiveness of the proposed method, ablation experiments are performed under All-search mode on SYSU-MM01 dataset, and the experimental results are shown in Table 1. The baseline method uses ResNet-50 as the backbone network followed by the GeM pooling, batch normalization and fully connected layer and trained with WRT loss and CE loss in this setting.

Effectiveness of the Random Combination of Channels (RCC). Comparing the results of the first and third rows, it can be seen that rank-1, mAP and mINP increased by 7.01%, 7.75% and 6.79% after using RCC, which fully proved the effectiveness of this method.

Effectiveness of the Feature Fusion. Comparing the experimental results in the first and second rows, it can be seen that the rank-1, mAP and mINP are 4.26%, 6.44% and 6.70% higher than baseline after using feature fusion, which indicates that feature fusion can capture feature information at different layers. We conducted ablation experiments on feature fusion between different layers. As shown in Table 2. When layer1, layer2, and layer3 are fused, the best result can be obtained. The fusion of the layer4 will result in reduced performance and increased computational complexity.

Effectiveness of the Deep Mutual Learning (DML). Comparing the fifth and sixth rows, DML improves the rank-1, mAP and mINP by 2.89%, 2.95% and 3.78%. This proves that using fusion branch as an auxiliary branch to reduce network's dependence on global features is effective.

To further explore the role of DML, we experimented with different classification functions on the global and fusion branches. As shown in Table 3. When using the L_{id} or L_{KD} alone, the improvement in accuracy is marginal. When L_{id} and L_{KD} are employed concurrently, the performance the performance is superior to either used alone, owing to the interplay between the L_{id} (hard target) and the L_{KD} (soft target). In addition, the introduction of L_{KD} as a label smoothing regularization term improved the performance of the model, indicating its contribution to enhancing the model's generalization ability.

Furthermore, as shown in Table 4, we employ DML between different branches. DML between global and fusion branches resulted in a decrease of 1.01%/1.35%/1.42% in rank-1/mAP/mINP because the global and fusion branches are asynchronous. If the prediction from one branch is used as the target distribution, the learned asynchronous knowledge will result in performance degradation. The model's performance is improved through the DML between global or fusion branches and joint branch because joint feature are high-dimensional information that includes both global and local features. Therefore, joint branch enables synchronized learning with the global and fusion branches. When these three branches learn synchronously, the model achieves its peak performance.

4.4. Parameter analysis

In this section, the weight parameters θ_1 and θ_2 introduced in 3.5 are analyzed. Fixing one parameter and then adjusting the other. Specifically, when evaluating the parameter θ_1, we first assign θ_2 a fixed value of 1.0, and then adjust $\theta_1 \in [0, 2]$ to observe the change in performance. We conducted experiments under the All-search mode of SYSU-MM01 dataset. The experimental results are shown in the Fig. 4.

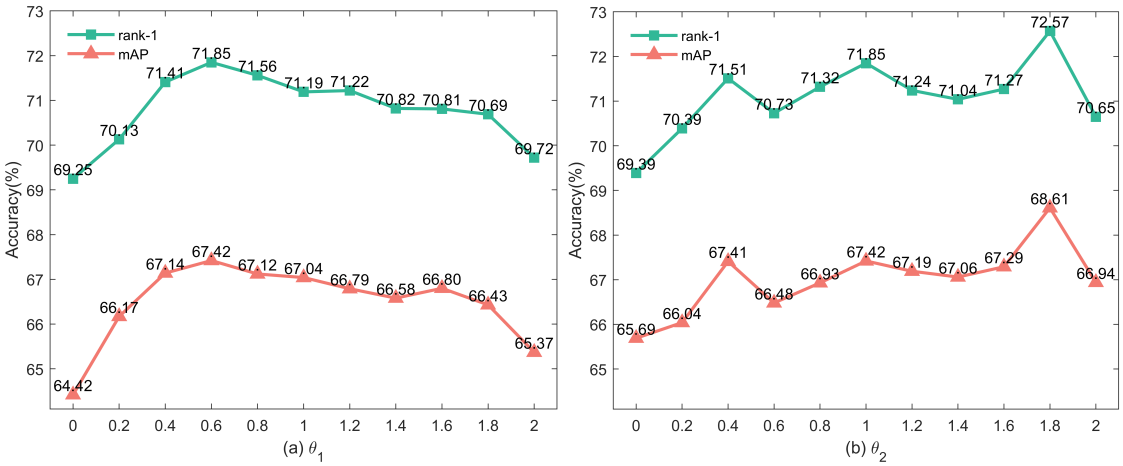

Figure 4: Evaluate the effect of parameters θ_1 and θ_2 under the All-search mode on SYSU-MM01 dataset.

As can be seen from Fig. 4(a), when θ_1 is less than 0.6, rank-1 and mAP also increase and reach the highest value when $\theta_1 = 0.6$. After that, with the increase of θ_1, rank-1 and mAP showed a downward trend. Therefore, we choose 0.6 as the value of θ_1 for the later experiment.

Fix θ_1 and adjust θ_2, the experimental results are shown in Fig. 4(b), θ_2 meets three peak points at 0.4, 1 and 1.8. When $\theta_2 = 1.8$, rank-1 and mAP are the highest so that we select 1.8 as the value of θ_2. In summary, the weight parameters θ_1 and θ_2 are set to 0.6 and 1.8.

Comparing the performance curves in Fig. 4(a) and Fig. 4(b), we can see that the performance curve of θ_1 is relatively stable, while the performance curve of θ_2 fluctuates significantly. This indicates that our model is more sensitive to θ_2, which means it is necessary to balance the contributions of global and fusion branches.

4.5. Visualization

Attentive feature maps. GradCam is employed to produce attentive feature maps for visualizing the features learned by the global and fusion branches. In an attentive feature map, regions closer to red indicate higher values. As shown in Fig. 5, the results reveal that the red area in the output of the fusion branch is more extensive compared to the global branch output. This suggests that the fusion branch can extract more feature information, compensating to some extent for the crucial person-related information overlooked by the global branch.

Retrieval result. Fig. 6 shows the top 10 search results of the proposed method on the SYSU-MM01 dataset are displayed, with the order of images indicating the similarity ranking. In the results, a green box signifies a match between the retrieved image and the query image, while a red box indicates a mismatch. These results demonstrate that the methods introduced in this paper enhance the accuracy of image retrieval, showcasing the effectiveness of the proposed approach in VI-ReID tasks.

Figure 5: Example of the attentive feature maps. The first row is the original image, the second and third rows are the outputs of the global and fusion branches.

Figure 6: Example of retrieval results on SYSU-MM01 dataset.

4.6. Comparison with state-of-the-arts

The method is compared with various state-of-the-art VI-ReID approaches on the SYSU-MM01 and RegDB datasets. These comparisons include the earlier methods Zero-Pad[15] and Hi-CMD[2], the methods based on feature-level constraints such as BDTR[19] and AGW[23] , the methods based on generative adversarial networks like cmGAN[3] and D2RL[14], the intermediate mode-based methods including X-model[8] and cm-SSFT[12], as well as advanced methods like SFA-Net[9], NFS[1], FMCNet[26], MID[7], MCLNet[4], AGMNet[10] and MFCS[17]. Benchmarking the proposed approach against established

Table 5: Comparison with other advanced methods under Visible-Thermal mode on RegDB dataset.

Methods	Publish	Visible-Thermal					Thermal-Visible				
		r=1	r=10	r=20	mAP	mINP	r=1	r=10	r=20	mAP	mINP
Zero-Pad[15]	ICCV17	17.75	34.21	44.35	18.90	-	16.63	34.68	44.25	17.82	-
BDRT[19]	IEEE19	34.62	58.96	68.72	33.46	-	34.21	58.74	68.64	32.49	-
D2RL[14]	CVPR19	43.40	66.10	76.30	44.10	-	-	-	-	-	-
X-model[8]	AAAI20	62.21	83.13	91.72	60.18	-	-	-	-	-	-
DDAG[21]	ECCV20	69.34	86.19	91.49	63.46	49.24	68.06	85.15	90.31	61.80	48.62
AGW[23]	TPAMI21	70.05	86.21	94.55	66.37	50.19	70.49	87.21	91.84	65.90	51.24
Hi-CMD[2]	CVPR20	70.93	86.39	-	66.04	-	-	-	-	-	-
cm-SSFT[12]	CVPR20	72.30	-	-	72.90	-	71.00	-	-	71.70	-
SFA-Net[9]	IEEE21	76.31	91.02	94.27	68.00	-	70.15	85.24	89.27	63.77	-
MCLNet[4]	ICCV21	80.31	92.70	96.03	73.07	57.39	75.93	90.93	94.59	69.49	52.63
NFS[1]	CVPR21	80.54	94.96	95.07	72.10	-	77.95	90.45	93.62	69.79	-
MFCS[17]	IEEE24	85.34	-	-	76.39	-	83.88	-	-	75.16	-
MID[7]	AAAI22	87.45	-	-	84.85	-	84.29	93.44	-	81.41	-
AGMNet[10]	IEEE23	88.40	95.10	96.94	81.45	68.51	85.34	94.56	97.48	81.19	65.76
FMCNet[26]	CVPR22	89.12	-	-	84.43	-	88.38	-	-	83.86	-
Ours		**92.00**	**97.34**	**98.38**	**88.01**	**79.77**	**90.01**	**96.90**	**98.30**	**86.15**	**75.31**

methods allows for a comprehensive evaluation of its performance and effectiveness in the VI-ReID tasks.

Evaluations on RegDB. As shown in Table 5, under the visible to thermal mode of RegDB dataset, our method achieves the Rank-1 accuracy of 92.00%, mAP of 88.01% and mINP of 79.77% . The rank-1 and mAP are increased by 2.88% and 3.58% compared with FMCNet[26]. Compared with AGMNet[10], the rank-1/mAP/mINP have improved by 3.60%/6.56%/10.60%.

Evaluations on SYSU-MM01. As shown in Table 6, our method performs better than most existing SOTAs. Specifically, our method achieves the Rank-1 accuracy of 72.57%, mAP of 68.61% and mINP of 54.64% in all-search mode. Compared with the latest method MFCS[17], the rank-1/mAP of our work have increased 1.98%/1.12% and 2.03%/1.15% under All-search and Indoor-search modes, respectively.

According to the above comparative experiments, it can be seen that the proposed method has achieved good results on both SYSU-MM01 and RegDB datasets.

5. Conclution

To address the limitations posed by the small dataset size, we introduce a data augmentation method named RCC, aimed at mitigating modal differences and enhancing model performance without adding substantial computational overhead. By fusing middle layer features of the network, we tackle the issue of shallow and deep features containing disparate information, ensuring that key features are not overlooked. Through the implementation of deep mutual learning, we enable synchronous learning across three branches. Experimental

Table 6: Comparison with other advanced methods under All-search and Indoor-search modes on SYSU-MM01 dataset

Methods	Publish	All-search					Indoor-search				
		r=1	r=10	r=20	mAP	mINP	r=1	r=10	r=20	mAP	mINP
Zero-Pad[15]	ICCV17	14.80	54.21	71.33	15.95	-	20.58	63.38	85.79	26.92	-
cmGAN[3]	IJCAI18	26.97	67.51	80.56	27.80	-	31.63	77.23	89.18	42.19	-
BDTR[19]	IEEE19	27.32	66.96	81.07	27.32	-	32.46	77.42	89.62	42.46	-
D2RL[14]	CVPR19	28.90	70.60	82.40	29.20	-	-	-	-	-	-
Hi-CMD[2]	CVPR20	34.94	77.58	-	35.94	-	-	-	-	-	-
AGW[23]	TPAMI21	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.20
X-model[8]	AAAI20	49.92	89.79	95.96	50.73	-	-	-	-	-	-
DDAG[21]	ECCV20	54.75	90.39	95.81	53.02	39.62	61.02	94.06	98.41	67.98	62.61
NFS[1]	CVPR21	56.91	91.34	96.52	55.45	-	62.79	96.53	99.07	69.79	-
MID[7]	AAAI22	60.27	92.90	-	59.40	-	64.86	96.12	-	70.12	-
cm-SSFT[12]	CVPR20	61.60	89.20	93.90	63.20	-	70.50	94.90	97.70	72.60	-
MCLNet[4]	ICCV21	65.40	93.33	97.14	61.98	47.39	72.56	96.98	99.20	72.10	-
FMCNet[26]	CVPR22	66.34	-	-	62.51	-	68.15	-	-	74.09	-
AGMNet[10]	IEEE23	69.63	96.27	98.82	66.11	52.24	74.68	97.51	99.14	78.30	74.00
MFCS[17]	IEEE24	70.59	96.22	98.77	67.49	-	75.98	98.12	99.62	80.24	-
Ours		**72.57**	**96.60**	**98.89**	**68.61**	**54.64**	**78.01**	**98.13**	**99.64**	**81.39**	**77.54**

results demonstrate the effectiveness of this approach compared to models solely relying on global features, showcasing its capability to improve performance in VI-ReID tasks.

References

[1] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 587–597, 2021.

[2] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10257–10266, 2020.

[3] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 6, 2018.

[4] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 16403–16412, 2021.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[7] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1034–1042, 2022.

[8] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4610–4617, 2020.

[9] Haojie Liu, Shun Ma, Daoxun Xia, and Shaozi Li. Sfanet: A spectrum-aware feature augmentation network for visible-infrared person reidentification. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[10] Haojie Liu, Daoxun Xia, and Wei Jiang. Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification. *IEEE Journal of Selected Topics in Signal Processing*, 2023.

[11] Songtao Liu, Di Huang, and Yunhong Wang. Learning spatial fusion for single-shot object detection. *arXiv preprint arXiv:1911.09516*, 2019.

[12] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020.

[13] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[14] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–626, 2019.

[15] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017.

[16] Daoxun Xia, Haojie Liu, Lili Xu, and Linna Wang. Visible-infrared person re-identification with data augmentation via cycle-consistent adversarial network. *Neurocomputing*, 443:35–46, 2021.

[17] Xi Yang, Wenjiao Dong, Meijie Li, Ziyu Wei, Nannan Wang, and Xinbo Gao. Co-operative separation of modality shared-specific features for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 2024.

[18] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[19] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019.

[20] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29:9387–9399, 2020.

[21] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 229–247. Springer, 2020.

[22] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021.

[23] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[24] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[25] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[26] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7349–7358, 2022.

[27] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[29] Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386:97–109, 2020.