Computers and Electrical Engineering

Dynamic Feature Weakening for Cross-Modality Person Re-Identification

--Manuscript Draft--

Manuscript Number:	COMPELECENG-D-23-00104R1
Article Type:	Research Paper
Keywords:	Person re-identification, Cross-modality, Dynamic feature weakening, Segmented learning network.
Abstract:	Cross-modality person re-identification (Re-ID) can be used to perform all-weather pedestrian monitoring tasks by matching the visual features captured in both visible and infrared images. However, the existing mainstream methods mainly rely on a few high response modality-invariant features, leading to problems such as easy loss of person cues and poor algorithm robustness in the recognition process. In this paper, we propose a cross-modality recognition method with dynamic feature weakening (DFW) and flexible segmentation learning as the core mechanisms, specifically: (1) DFW is introduced into the modality-shared network to weaken the model's over-reliance on high response regions of the image. As a result, it can globally collect and enrich modality-invariant features with diversity. (2) A segmented learning network (SLN) with overlap penalty constraint is used for fine-grained mining of each modality-invariant feature, while avoiding redundant focus on a region and maintaining complete learning of diversity information. The experimental results show that the method in this paper greatly enriches the feature representation, and significantly improves the recognition accuracy of the model. On the SYSU-MM01 dataset, compared with the traditional modality feature learning method, the Rank1 and mAP of this method are increased by 2.56% and 4.53%, respectively.

Dear editors and reviewers,

My paper titled ' Dynamic Feature Weakening for Cross-Modality Person Re-Identification', No. COMPLECENG-D-23-00104, received the reviewed modification comments on February 6, 2023. I would like to thank the editors and reviewers for their hard work in improving this paper. I have read the editors' and reviewers' comments carefully, adjusted the format and content of the paper, and made specific changes in response to the review comments (the relevant changes are marked in red in the " Revision of manuscript" and in the response below), as follows.

Reviewer #2: In this paper, the author proposes a module to prevent overfitting.

1. How to use F_{out}? The pipeline is not clear.We get the F_{out} and then how to use it?
The response is as follows.
This part has been relabeled in Figure2, and the weakened feature F_{out} can be obtained after dynamic weakening matrix processing, which is then fed into layer4, and then richer semantic information of pedestrians can be extracted.

2. I remember some papers like [a] Figure5, also say the loss can diversify the attention. What is the difference between your method and loss-based method?
[a] A Discriminatively Learned CNN Embedding for Person Reidentification. TOMM 2019.

The response is as follows.
[a] proposes a model based on identity loss and verification loss. Since the identity network and the verification network exhibit different activation patterns for pedestrians, the authors exploit the advantages of both networks in the proposed model to make the model activate more neurons for the purpose of attention diversification. In contrast, our model mainly adopts a dynamic feature weakening approach to make the model learn more diverse features by weakening its over-reliance on some salient features, and we also design an overlap penalty loss to constrain the sub-network learning distance to focus on the detailed features of different regions of the pedestrian.

3. It would be interesting to add re-ranking. Could you try gpu-based re-ranking at https://github.com/layumi/Person_reID_baseline_pytorch/tree/master/GPU-Re-Ranking if possible?

The response is as follows. Added in Tables 3 and 4.

4. The author only reported the results in limited papers. It would be good to include all published results on the leaderboard such as https://github.com/layumi/Person_reID_baseline_pytorch/tree/master/leaderboard-IR

The response is as follows.
The results on the leaderboard have been added to Tables 3 and 4. However, due to the limited space of the paper, we have mainly added the results of some recently published papers to the comparison experiments with mainstream methods in the paper.

5. This is not a weakness. I hope the author could add experiments or discussion on general image retrieval datasets, such as such as building retrieval [i] and drone-based geo-localization [j], which also focuses on multi-modality. I also look forward the author can add more of your ideas of person reid to future works. I think the paper is good to be accepted, and more your ideas will make it better.

The response is as follows.
This paper focuses on the cross-modal pedestrian re-identification task and validates the effectiveness of the proposed model by conducting experiments on a cross-modal pedestrian dataset. Special experiments and discussions for other multimodal datasets (e.g., the proposed building retrieval or drone-based geo-localization, etc.) will be considered in future work.

Reviewer #3: In this paper, the authors investigate dynamic feature weakening for cross-modality person re-identification. Although this work presents some interesting findings and it is well written, it has several concerns that needs to be alleviated. Here, there are some concerns of this reviewer:

1 The novelty/originality shall be justified by highlighting that the manuscript contains sufficient contributions to the new body of knowledge. The knowledge gap needs to be clearly addressed in Introduction.

The response is as follows.
We re-state in the introduction that, for example, the literature [9] [10] [11] relies mainly on high response regions of multi-modality images for recognition [12], while in practice high response regions may focus on only a small number of cues. In contrast, the dynamic feature weakening approach described in this paper adaptively weakens the reliance of model training on some of the cues in the high response regions to enable the network to focus more on non-significant regions and extract multi-modality invariant feature representations on a larger scale.

Some scholars often use hard spatial partitioning methods [14][15] to obtain fine-grained features of pedestrians. A direct hard spatial partitioning approach not only destroys the integrity of local features of pedestrians, but also affects inter-modality information matching. Some methods [16,17] introduce corresponding auxiliary alignment strategies to accommodate feature matching, but this will further increase the complexity of the model [18]. In contrast, this paper adopts a flexible segmentation method to mine the fine-grained features of pedestrians, and this mechanism can effectively avoid the risk of local integrity being destroyed due to hard segmentation without overburdening the model.

2 How scalable is the presented approach?

The response is as follows.
The dynamic feature weakening method proposed in this paper is a plug-and-play model that requires only a few interface parameters to be plugged into the model for application, and the algorithmic model can be multi-threaded for processing to improve operational efficiency.

3 The proposed method might be sensitive to the values of its main controlling parameters. How did you determine the parameters? Please elaborate on it.

The response is as follows.
The parameters involved in the method are mainly determined by experiments, for example, in the segmented learning network, we mentioned the separation parameter and the weight coefficient, we first fixed the weight coefficient as 1 and then set different separation parameters to determine the optimal value of the separation parameter by the results of the model on the two data sets. Then, we fixed the separation parameter in the same way to further determine the value of the weight coefficient.

4 Please specify details of the computing platform and programming language used in this study.

The response is as follows.
The model proposed in this paper is based on the NVIDIA 2080Ti GPU computing platform and 64-bit Ubuntu 18.04 operating system, and is implemented using the Pytorch 1.6 framework and the Python 3.6 programming language.

5 The computational cost of the proposed approach should be discussed in this work. The approach must be computationally efficient to be used in practical applications.

The response is as follows.
The computational costs of the different methods on the RegDB dataset are shown in Tab 2. We ran the AGW and MMG codes on the same experimental platform and environment, and to avoid experimental results being affected by parameters and number of images, we used the same batches and did not make changes to the hyperparameters of the model.

Since the model uses a segmented learning network consisting of two branches at layer4 stage, the metrics of parameters and FLOPs are increased compared to the baseline model and AGW which only uses a single-stream network. However, it can be seen from the experimental results that the accuracy of the proposed model in this paper has a significant improvement compared with the baseline model and AGW. In addition, the FLOPs and training time of MMG are as high as 36.633/G and 120/min, while our FLOPs and training time are only one-half of that method. Meanwhile, our Rank1 and mAP are higher by 1.58% and 5.73%, respectively. This validates the efficiency of our method.

6 The novelty and contribution of the present work need further justifications. In practical applications, noises are inevitable. Please discuss how the noises would impact the results and conclusions of this study.

The response is as follows.
We have added some recently published paper results in our experiments comparing with mainstream methods, and added experiments and analysis such as re-ranking to further justify the work in this paper. In addition, since the samples in the dataset used in this paper are collected in

real scenes, the samples themselves have problems of background noise, occlusion, and pose differences, and our model is trained based on these samples, so the model has certain anti-interference capability. However, due to the limited dataset and insufficient training of the model, some noise may interfere with the pedestrian retrieval task, for example, different identity images with the same background can easily be recognized as the same pedestrian, which will make the recognition accuracy of the model lower.

7 Authors have not presented the limitations of this work. How this work can be extended in the future?

The response is as follows.
As we mentioned in our conclusion, the limitation of the work in this paper is that the network uses a uniform feature space learning approach to optimize the visible and infrared feature relationships. However, this approach limits the representation of effective information due to the large offset between heterogeneous data distributions. In the future, we consider building modality representation generators in front of the shared space to enhance the recognition capability of the model by adaptively generating intermediate representations with multi-modality characteristics.

8 The nomenclature should be included to help the reader to follow the paper conveniently.

The response is as follows.
In this paper, the abbreviations of nomenclature are explained when they first appear, e.g., Person re-identification (Re-ID), dynamic feature weakening (DFW), segmented learning network (SLN), etc.

Reviewer #4: This paper proposes a cross modality person re-identification approach using proposed dynamic feature weakening layer and segmented learning network with overlap penalty loss to increase the learned features diversity and avoid high overlap of model focusing area. This paper is well written. Here are some comments:

- What is the computational complexity of the proposed approach compared with SOTA methods listed in the paper?

The response is as follows.
We compare the proposed model with SOTA method in Table 2, where the FLOPs and training time of MMG are as high as 36.633/G and 120/min, while our FLOPs and training time are only one-half of that method, which shows that our method is simple and effective.

- In the experiment how many sub-networks (SLN) are used?

The response is as follows.
We use two sub-networks in our experiments.

- Please explain how the positive and negative are sampled for triplet loss $L_{\_hc\_trip}$ and $L_{\_re\_id}$, and why in equation (7), only the weight for L_op is considered for adjusting.

The response is as follows.

(1) L_{re_id} in ternary group loss in positive and negative sampling, first in a small batch, randomly selected a query image as a fixed sample, and then randomly selected a positive sample consistent with the identity of the query image and a negative sample with inconsistent identity to form a ternary group, repeat this step until finally generate a sufficient number of ternary group samples to build into a batch. And the heterogeneous central triad loss L_{hc_tri} is designed based on the triad loss, all of which is to randomly select a query image as a fixed sample, then randomly select a positive sample that is consistent with the identity of the query image but different modality, and then randomly select a negative sample that is inconsistent with the identity of the query image, which may be the same modality or different modality from the query image.

(2) The other two losses are regular losses in cross-modality person re-identification, which are mainly used to improve the similarity of cross-modality person within classes. And L_{op} is the loss designed in this paper for avoiding pedestrian detail features to be learned repeatedly. In order to discuss the impact of the degree of L_{op} constraint on the model, therefore, this paper focuses on the weight of this loss.

- In Fig 4 and 5, the performance does not have a clear pattern when S or lambda increasing. Is each sample in the plots based on several independent experiments? Similar for Fig 3

The response is as follows.

Figure 3 Figure 4 and Figure 5 are based on independent experiments. From Figure 3, it can be observed that when the threshold value is m, the optimal values of the weakening factor are 0.5 and 0.2 on both data sets, and the model accuracy decreases with the increase of the weakening factor. And when the weakening factor is e, the model accuracy shows the same trend on both data sets, which increases and then decreases, and reaches the best accuracy at the threshold value of 0.5. As can be seen from Figure 4, the separation parameter s, as an important parameter for the loss of overlap penalty, has an optimal value of 100 on both datasets, and the model accuracy decreases as s increases, and its performance does not show a clear pattern, which is mainly related to the samples, as the separation parameter is used to avoid the overlap of image areas of network concern, and the pedestrian size ratio varies in different sample images, so this performance does not present a clear pattern. The weight coefficients of the overlap penalty loss in Figure 5 are similar.

- In table 1, please list the result when only one of DFW, SLN, RE applied on top of baseline if possible

The response is as follows.

The results and analysis of this part of the experiment have been added and are shown in Table 1.

minors

- The last sentence of section 3.2 is not clear

The response is as follows.

The last sentence of section 3.2 reads “It is worth mentioning that in the process of weakening the high response features of the network, background noise, etc., which is not introduced to interfere with the model training.” . From the heat map in Figure 1, it can be seen that the model focuses on the location of the pedestrian's own features, rather than the surrounding objects and background information and other noise, which indicates that the model can learn richer pedestrian features by weakening the high response region, and at the same time the model does not learn some meaningless features to affect the model accuracy when it focuses on the non-significant region.

- In equation (1), please explain what GAP and GMP stand for when introduce these two pooling methods

The response is as follows.
We use global average pooling GAP and global maximum pooling GMP to compress the feature mapping channels, where GAP can obtain the contextual information of pedestrian images by calculating the average value of the whole region, while GMP obtains the discriminative information of pedestrians by focusing on the maximum response region of the image. By using a combination of GAP and GMP to downsample the features, it helps to improve the performance of the model.

Finally, all the authors would like to thank the reviewers for their very professional and detailed review comments, as well as the editor-in-chief and the editorial board for their many important suggestions and help in revising this paper!

Dynamic Feature Weakening for Cross-Modality Person Re-Identification

Abstract

Cross-modality person re-identification (Re-ID) can be used to perform all-weather pedestrian monitoring tasks by matching the visual features captured in both visible and infrared images. However, the existing mainstream methods mainly rely on a few high response modality-invariant features, leading to problems such as easy loss of person cues and poor algorithm robustness in the recognition process. In this paper, we propose a cross-modality recognition method with dynamic feature weakening (DFW) and flexible segmentation learning as the core mechanisms, specifically: (1) DFW is introduced into the modality-shared network to weaken the model's over-reliance on high response regions of the image. As a result, it can globally collect and enrich modality-invariant features with diversity. (2) A segmented learning network (SLN) with overlap penalty constraint is used for fine-grained mining of each modality-invariant feature, while avoiding redundant focus on a region and maintaining complete learning of diversity information. The experimental results show that the method in this paper greatly enriches the feature representation, and significantly improves the recognition accuracy of the model. On the SYSU-MM01 dataset, compared with the traditional modality feature learning method, the Rank1 and mAP of this method are increased by 2.56% and 4.53%, respectively.

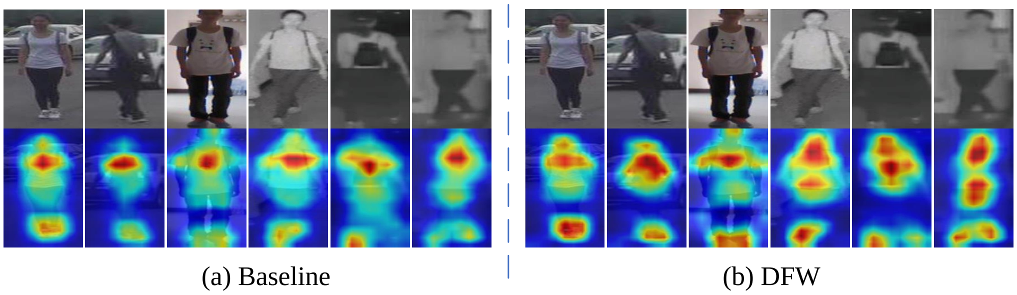
Keywords:
Person re-identification, Cross-modality, Dynamic feature weakening, Segmented learning network.

1. Introduction

Person re-identification (Re-ID) aims to retrieve a specific person from under multiple non-overlapping camera devices [1, 2], which has broad application prospects in the field of intelligent video surveillance. At present, great progress has been made in person Re-ID based on a single visible modality [3, 4]. Whereas, in low light or nighttime conditions, a surveillance system that integrates infrared and visible cameras is normally required to capture the visual information of a person [5]. The person Re-ID based on this monitoring system is essentially a cross-modality task. Compared with single-modality person Re-ID, the modality visual differences caused by different spectral cameras pose a huge challenge for cross-modality person retrieval [6, 7]. Furthermore, the problem of intra-class variation caused by different shooting angles, body occlusion, and poses [8] still exists in cross-modality.

Most of the existing cross-modality person Re-ID methods follow the idea of modality-shared feature (or modality-invariant feature) learning [9, 10]. The problem of low recognition accuracy caused by cross-modality differences is solved by extracting specific features of different modality images separately and then mapping them to the same public space to extract invariant information [11]. These methods primarily rely on the high response region of multi-modality images for recognition, Nevertheless, in practical situations, the high response region may centralize only a small number of cues [12]. As can be seen from Fig 1 (a), the high response regions of visible and infrared images are principally focused on the top logo, backpack straps, and shoes of the person. Nonetheless, other distinguishable modality-invariant features such as body contour and clothing type of the person are ignored. In cross-modality scenarios with complex conditions, any feature is at risk of being corrupted [13]. We hope the network will stop concentrating too much on one feature and instead follow as many modally-invariant features as possible to ensure that there are sufficient features available for cross-modality use, thereby enhancing the robustness of the model for modality transformation. Based on this, we propose a dynamic feature weakening (DFW) method to adaptively weaken the reliance of model training on certain cues in high response regions. This enables the network to focus more on non-significant regions and extract a broader range of multi-modality invariant feature representations.

Since DFW is a quantitative approach that sacrifices quality feature representation in a particular modality for a more effective cross-modality feature representation population, an extremely rich set of modality-invariant features can be learned by adopting DFW. To ensure that this information can be fully and effectively learned, we consider further mining the fine-grained features of pedestrians. Fine-grained features play an important role in distinguishing different modalities of persons with similar appearances but



(a) Baseline

(b) DFW

Figure 1: Visualization of the model feature heatmaps. (a) and (b) show the feature maps of the baseline model and our proposed DFW model, respectively.

different detailed information. Some scholars often use hard space partitioning methods [14, 15] to obtain fine-grained features of a person. Yet, due to the different imaging mechanisms and intra-class poses of multi-modality images, the visual features described in certain regions in visible and infrared images also differ. The direct adoption of hard spatial division will not only compromise the integrity of local features of a person but also affect inter-modality information matching. Some approaches [16, 17] have introduced corresponding auxiliary alignment strategies to accommodate feature matching, but they can increase the complexity of the model [18]. Therefore, we adopt a flexible segmentation mechanism, which can effectively avoid the risk of local integrity being destroyed due to hard segmentation without overburdening the model. Specifically, fine-grained cues of a person are mined using a segmentation learning network (SLN), and the overlap penalty loss is used to constrain the distance between regions of interest to different sub-networks, allowing more complementary fine-grained features to be learned. The segmented learning strategy is a flexible feature learning strategy adapted to cross-modality, which learns the complete detailed information of a person by constraining the corresponding loss without additional prior operations.

The main contributions of this paper can be summarized as follows.

1. In response to the problem that existing methods focus too much on high response regions cross-modality images, leading to a small number of features and poor algorithmic robustness, we propose DFW, which adaptively weakens the reliance on high response regions to enrich and diversify the modality-invariant features.
2. We employ SLN to learn modality-invariant features at a finer granularity, and design an overlap penalty loss to flexibly split segments and

avoid redundant attention to a region, which maintains the integrity and adequacy of the features.

3. Extensive experimental results on two cross-modality datasets demonstrate that our proposed framework outperforms some traditional techniques significantly.

The general structure of this paper is as follows. Chapter 2 introduces the related research work on cross-modality person Re-ID. Chapter 3 proposes a dynamic feature weakening algorithm and segmentation learning algorithm adapted to cross-modality scenarios. Chapter 4 validates the proposed method through experiments and compares it with some mainstream methods. The final chapter concludes the work presented in this paper.

2. Related work

In recent years, cross-modality person Re-ID has received increasing attention due to its effectiveness in low light conditions. Various cross-modality person Re-ID methods have been proposed to solve the challenge of visual disparity posed by different spectral sensors. Wu et al. [19] provided a large cross-modality dataset SYSU-MM01 and facilitated the learning of domain-specific nodes in the network by zero-padding the image channels. However, this method predominantly uses the identity information of a person for matching, and there is a serious deficiency in learning discriminative features. As a result, some scholars have started to turn their attention to the mining of multi-modality shared information. Ye et al. [8] proposed a dual-stream network structure based on AlexNet to extract multi-modality shared information by establishing connections between heterogeneous modalities in fully connected layers. However, since the fully connected layer deals with 1D vectors, there is a risk of losing the spatial information of a person when learning shared features in this layer. Liu et al.[20] obtained modality-invariant features by sharing partial convolution blocks, which improved the recognition ability of the model. Whereas, using only the convolutional layers to capture the invariant information of a person can result in missing feature usage. Huang et al. [21] captured modality-invariant features in modality-shared 2D and 3D space, respectively, which enhanced the discriminative power of modality feature representation. Nevertheless, the above methods rely on a small amount of information in the high response region for learning to mitigate the cross-modality variation problem, leading to insufficient learning of modality-invariant features. Ning et al. [12] proposed a feature weakening block to weaken the high response region features and let the model pay attention to other meaningful regions. This approach can learn richer person features, but it adopts a completely fixed weakening way and also does not go to account for the inter-modality variation problem. In cross-modality person Re-ID, the different focus of the model on multi-modality images leads to large differences in the high response regions of each modality image, and the uniform use of the set fixed way will affect the effectiveness of image weakening. Furthermore, when feature weakening is performed in shallow convolution, the model is prone to learn the specific information of each modality, which is not conducive to intra-class cross-modality person matching. Therefore, this paper designs a dynamic feature weakening (DFW) and introduces it into the shared deep feature space. This method can adaptively expand the range of high response regions of multi-modality images, enabling the modality-invariant features to be comprehensively learned.

In addition, some traditional methods [20, 21] have mainly utilized spatial division methods for fine-grained feature extraction. Sun et al. [14] extracted the local information of a person by slicing the network output feature map. Wang et al. [15] proposed an MGN structure that evenly divided the feature map into different numbers of blocks to obtain multi-scale fine-grained information for person matching. Nonetheless, due to the presence of spatial misalignment in multi-modality images, direct hard division approaches for learning fine-grained cues lead to inconsistent matching of modality information, which in turn affects the recognition accuracy of the model. Yang et al. [18] abandoned the previous hard division and captured more effective local information by allowing each branch to activate different regions of the image. Inspired by this, we employ a segmented learning network (SLN) to guide the sub-networks in autonomously learning non-overlapping body regions in multi-modality images to flexibly mine fine-grained cues of a person.

3. Proposed method

3.1. Baseline network

Following the previous work [20], our baseline model adopts a dual-stream network structure. In the phase of modality-specific feature extraction, we extract modality-specific features of different modality images in a parameter independent manner by the shallow convolutional layers, which enhances

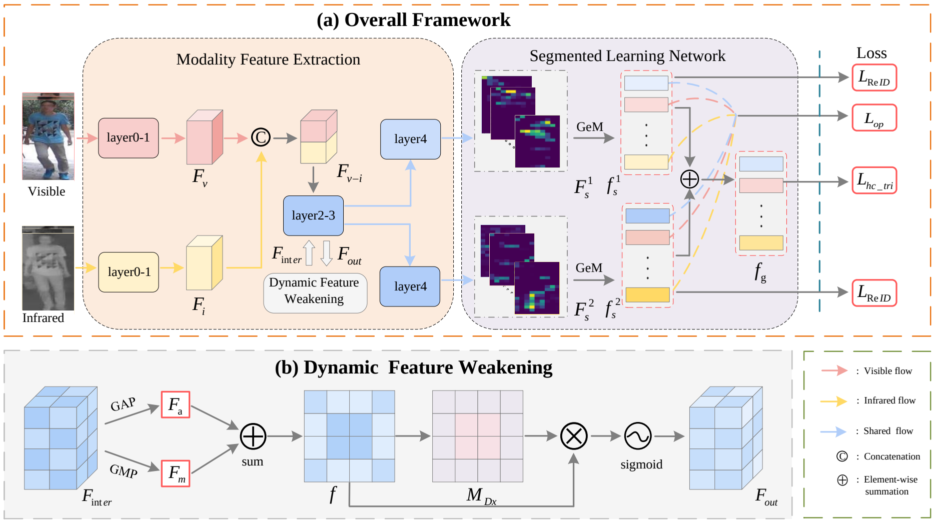


Figure 2: Illustration of the proposed model. (a) The visible and infrared images are respectively sent to the modality-specific network (layer0 and layer1) for modality-specific feature extraction, and then send to the designed modality-shared networks (layer2, layer3, and layer4) to extract the shared features. (b) Detailed structure diagram of the dynamic feature weakening. By performing a series of operations on the shared features of the intermediate layer, the weakened shared features are finally output.

the discriminative ability of the modality-specific features. In the phase of modality-shared feature extraction, we utilize the other three convolutional layers to learn modality-shared features in the way of parameter sharing for a cross-modality person matching.

3.2. Dynamic feature weakening

In cross-modality person retrieval, modality-invariant features tend to have better discriminative ability. Whereas, the model pays too much attention to the high response regions in the person image during the training process, which results in some modality-invariant features in the low response regions being ignored. Consequently, this study considers weakening the high response region features in the modality-shared network and focusing on other valuable visual cues. Specifically, we set a weakening factor α and a threshold β, where $\alpha, \beta \in [0, 1]$, to achieve significant region weakening by applying a weakening factor to the high response.

To simplify the feature value calculation, we utilize global average pooling (GAP) and global max pooling (GMP) in channel attention to compress the feature map $F_{inter} \in \mathbb{R}^{C\times H\times W}$ channels. GAP calculates the average value of the entire region, enabling it to obtain contextual information of the pedestrian image. GMP focuses on the maximum response region of the image, allowing it to obtain discriminative information of the pedestrian. By combining the use of GAP and GMP to downsample the features, it helps to improve the performance of the model. The calculation is as follows,

$$f = GAP(F_{inter}) + GMP(F_{inter}). \tag{1}$$

As the size of the high response region varies among different input images, using a fixed threshold and weakening factor can limit the effectiveness of delineating the high response region of the feature map and weakening it. Therefore, we propose a dynamic feature weakening (DFW) approach, illustrated in Fig 2 (b). The method involves first computing the mean value of all the feature values in the feature map f, which is then used as the threshold. The comparison of all feature values with mean value is employed to determine the feature region corresponding to the feature value above or equal to the mean value as the high response region, otherwise, it is a low response region. Subsequently, the proportion e of the area of the entire feature map occupied by the high response region is used as the weakening factor, and thus the dynamic weakening matrix M_{Dx} is constructed, where $x = \{1, 2, 3\}$. The matrix is represented as follows,

$$\begin{aligned} M_{D1}(i,j) &= \begin{cases} \alpha, & \text{if } f(i,j) \geq mean, \\ 1, & \text{otherwise}, \end{cases} \quad M_{D2}(i,j) = \begin{cases} e, & \text{if } f(i,j) \geq \beta, \\ 1, & \text{otherwise}, \end{cases} \\ M_{D3}(i,j) &= \begin{cases} e, & \text{if } f(i,j) \geq mean, \\ 1, & \text{otherwise}, \end{cases} \end{aligned} \tag{2}$$

Where $f(i,j)$ denotes the feature value, $M_{D1}(i,j)$ indicates the weakening matrix when the threshold is the mean value of the feature map and the weakening factor is a fixed value, $M_{D2}(i,j)$ represents the weakening matrix when the threshold is a fixed value and the weakening factor is the proportion of high response regions, and $M_{D3}(i,j)$ denotes the weakening matrix when the threshold is the mean value of the feature map and the weakening factor is the proportion of high response regions.

From the multiplication result of the feature map and the dynamic weakening matrix, the weakened shared features can be obtained, and the calculation is as follows,

$$F_{out} = sigmoid(f \otimes M_{Dx}). \tag{3}$$

The feature F_{out} processed by the dynamic weakening matrix is then fed into layer4, which in turn extracts richer semantic information of the pedestrian. The dynamic weakening matrix is constructed using the mean value and proportion of feature maps occupied by high response regions, thereby achieving the purpose of weakening significant regions. This approach offers the advantage that the model can always pay attention to the most useful cues, regardless of changing conditions, which significantly enhances the self-adaptability of the model.

Learning of overall cues of person images facilitates the representation of modality-invariant features to maximize the matching difficulties caused by information differences between heterogeneous data. We perform feature visualization on the baseline and DFW model for visible and infrared images, respectively, as shown in Fig 1. It is easy to observe that the high response region of the baseline model is mainly concentrated on the upper part of the person's body and the feet for both the visible and infrared images. In contrast, the DFW model significantly expands the response region from a few upper body parts to the head and body contour of the person, and strengthens the model's attention on distinguishable information, such as shoes. This indicates that the DFW model selects more modality-invariant features in the person images, which facilitates cross-modality person matching compared to the baseline model. It is worth mentioning that the model focuses on the location of the pedestrian itself, rather than the surrounding objects, background information, and other noise. This indicates that the model can learn richer pedestrian features by weakening the high response region. At the same time, the model does not learn some meaningless features to affect the model accuracy when focusing on the non-significant region.

3.3. Segmented learning network

To further facilitate modality-invariant feature learning, we employ a segmented learning network (SLN) that ensures the integrity and complementarity of a person's detailed information by allowing the network to flexibly mine fine-grained features in different areas of a multi-modality image. The SLN is shown in Fig 2 (a). Our specific approach is to design layer4 as a segmented feature extraction structure to learn the fine-grained features of person images separately. To avoid problems such as overlearning salient area features and ignoring non-salient area features, which can be caused by each sub-network focusing only on salient areas of the image, we design an overlap penalty loss (described in the next section) to constrain the learning regions of the sub-network. By making the learning areas of the sub-networks non-overlapping, the model captures a greater diversity of spatially fine-grained features.

3.4. Loss function

Overlap penalty loss. We aim to extract complementary fine-grained features in the person image by constraining the distances between image regions that each sub-network focuses on. To achieve this, we adopt GeM pooling for the feature maps $F_s^1 \in \mathbb{R}^{C\times H\times W}$ and $F_s^2 \in \mathbb{R}^{C\times H\times W}$ respectively, which are calculated as follows,

$$\begin{aligned} f_s^1 &= GeM(F_s^1), \\ f_s^2 &= GeM(F_s^2), \end{aligned} \tag{4}$$

The position of the maximum response in $f_s^1 \in \mathbb{R}^{H\times W}$ and $f_s^2 \in \mathbb{R}^{H\times W}$ is taken as the center of the region of interest in the sub-network, and the center distance of the two regions is calculated. The overlap penalty loss function can be defined as follows,

$$L_{op} = \frac{1}{N} \cdot max[0, S - d(f_s^1, f_s^2)], \tag{5}$$

where N denotes the number of input images in the batch, S denotes the separation parameter, and $d(f_s^1, f_s^2)$ denotes the center distance between the concern regions of the sub-network.

Hetero-center triplet loss. In the cross-modality person retrieval problem, we hope to narrow the distance between intra-class cross-modality samples. Accordingly, we exploit the hetero-center triplet loss to constrain the total feature learning, and by limiting the distance from the center of the anchor sample to the center of positive and negative samples of other modalities. The hetero-center triplet loss is expressed as follows,

$$\begin{aligned} L_{hc\_tri} = & \sum_{i=1}^{P} \left[\rho + \left\| c_v^i - c_I^i \right\|_2 - \min_{\substack{n \in \{v,I\} \\ j \neq i}} \left\| c_v^i - c_n^j \right\|_2 \right]_+ \\ & + \sum_{i=1}^{P} \left[\rho + \left\| c_I^i - c_v^i \right\|_2 - \min_{\substack{n \in \{v,I\} \\ j \neq i}} \left\| c_I^i - c_n^j \right\|_2 \right]_+ , \end{aligned} \tag{6}$$

where $\left\| c_v - c_I \right\|_2$ denotes the Euclidean distance of the center of visible image c_v and the center of infrared image c_I, $\left\| c_v - c_n \right\|_2$ denotes the Euclidean distance of the center of visible image c_v and the center of negative sample c_n, and ρ is the margin.

The overall loss. We utilize L_{op} to constrain the center distance between the sub-network attention regions. For the human part features obtained by the each sub-network, we exploit L_{ReID} (including the classification loss and triplet loss) for supervised learning. We employ L_{hc_tri} to calculate the final features. The total loss is expressed as follows,

$$L_{all} = \lambda L_{op} + L_{hc\_tri} + L_{ReID}, \tag{7}$$

where λ is the weight coefficient to balance loss.

3.5. Overall training process

The overall training process of our model is shown in Algorithm 1.

4. Experiments

4.1. Datasets description

The SYSU-MM01 [19] dataset consists of data collected by four visible cameras and two infrared cameras. The training set includes 395 person identities, with 22258 visible images and 11909 infrared images. The testing set includes 96 person identities, with 301 visible images and 3803 infrared images. During the testing phase, the infrared images in the testing set form the query set, while the visible images form the gallery set. The SYSU-MM01 has all-search and indoor-search test modes. Because of the complex outdoor environment, the all-search mode tends to be more challenging than the indoor-search mode.

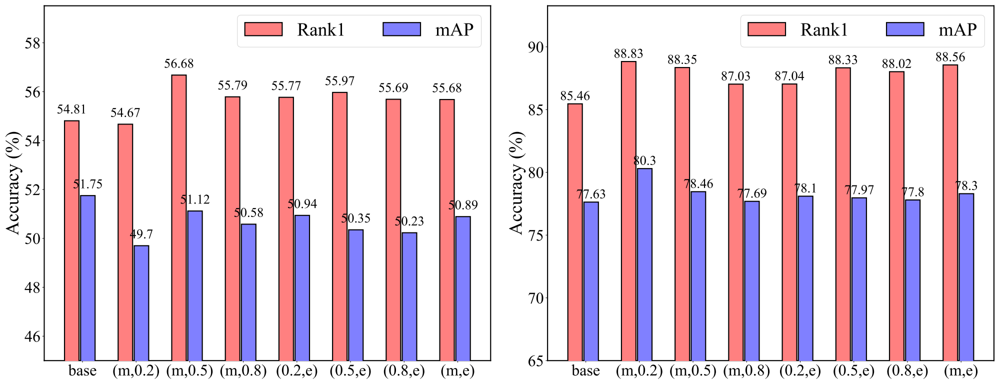
Algorithm 1 The overall training process

```
Input: Visible image set $V = \{v_1, \dots v_n\}$, Infrared image set $I = \{i_1, \dots i_n\}$;
Output: The trained cross-modality person Re-ID model;
Initialize the backbone network with the ImageNet pre-trained ResNet-50;
Preprocess images using random erasure;
Insert the DFW in the first two layers in the backbone shared network and initialize it randomly;
for $epoch = 1$ to $MaxEpochs$ do
    Select $2PK$ person images, and extract the modality-shared feature $F_{in}$ in the backbone shared network;
    Perform global average pooling and global max pooling on $F_{inter}$ respectively, and then sum up to obtain the feature map $f$;
    Obtain the weakened feature $F_{out}$ by Eq (3);
    Construct SLN to learn the non-overlapping area features $F_s^1$ and $F_s^2$ of the image respectively;
    Calculate the total training loss by Eq (5);
    Update the DFW and SLN together by back-propagating the gradient of Eq (7);
end for
```

The RegDB [22] dataset is collected by one visible camera and one thermal camera. The dataset contains 412 person identities, with 10 visible and thermal images for each person, respectively, for a total of 8240 images. According to the evaluation protocol of [23], the dataset is equally divided into a training set and a test set, each containing 206 person identities, 2060 visible images, and 2060 thermal images. Furthermore, RegDB includes visible-thermal and thermal-visible retrieval modes.

4.2. Evaluation metrics

We employ the cumulative matching characteristic (CMC) and average accuracy mean (mAP) to evaluate the effectiveness of this method. The CMC metric is typically represented by the value of Rank-k, which represents the probability that the first k retrieved images with the highest similarity between the query set and the gallery set are matched by the correct person. For instance, Rank1 indicates that the first retrieved image is matched by the correct person's probability. The mAP represents the mean value of the



(a) SYSU-MM01 (b) RegDB

Figure 3: Exploring the effects of different feature thresholds and weakening factors on model accuracy on SYSU-MM01 and RegDB datasets.

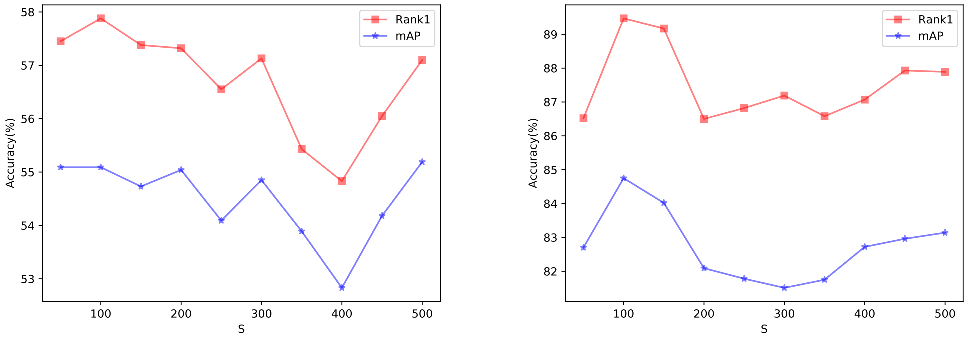
average accuracy of all query images, which reflects the ranking degree of correctly matched images in the retrieved list.

4.3. Implementation details

The model proposed in this paper is based on NVIDIA 2080Ti GPU computing platform and 64-bit Ubuntu 18.04 operating system, and is implemented using Pytorch1.6 framework and Python3.6 programming language. The ResNet50 pre-trained on ImageNet is used as the backbone network of the model. We adopt a batch sampling strategy to randomly select P identities in the dataset, and then K visible images and K infrared images are randomly selected for each identity. We set $P = 6$, $K = 6$ for the SYSU-MM01 dataset, and set $P = 8$, $K = 4$ for the RegDB dataset respectively. To boost the generalization ability of the model, in addition to common strategies such as random image cropping and horizontal flipping, we also exploit random erasure for data enhancement. During the training process, we take advantage of the SGD optimizer to optimize the network parameters. At the meantime, the initial learning rate is set to 0.1 and the training epoch is set to 60. For the hyper-parameter λ in Eq (7), we set λ to 0.5 and 1 for SYSU-MM01 and RegDB datasets, respectively.

4.4. Parameters analysis

The effect of the weaken parameter. We introduce DFW based on the baseline model and set up experiments according to the dynamic weaken-



(a) SYSU-MM01

(b) RegDB

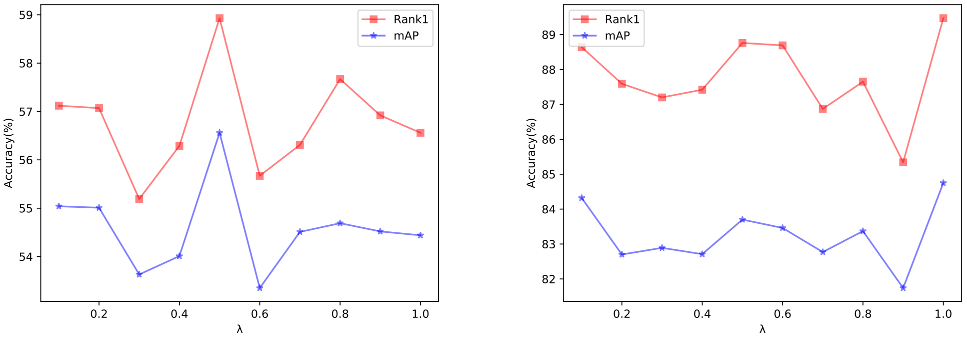
Figure 4: Exploring the effect of different separation parameters S on model accuracy on SYSU-MM01 and RegDB datasets.

ing matrix M_{Dx}, where $\alpha, \beta = 0.2, 0.5, 0.8$, to explore the impact of dynamic feature thresholds and weakening factors on model accuracy. As shown in Figure 3, m denotes the mean value of the feature map, and e denotes the weakening factor. From Fig 3 (a), it is apparent that the model performs best on the SYSU-MM01 dataset when the threshold is m and the weakening factor is 0.5. Compared with the baseline model, its Rank1 accuracy is improved by 1.87%, while mAP is reduced by 0.63%. Moreover, the visual retrieval results in Fig 6 demonstrate that the top-ranked person images have the highest similarity to the appearance features of the query images. This is because the main representation of modality-invariant features is the appearance features of a person, and the use of DFW notably enhances the learning of person appearance features. Therefore, the Rank1 accuracy of person images is significantly improved when cross-modality person matching is performed. Nevertheless, the mAP does not show the same enhancement phenomenon since the randomly selected query images in this dataset are not always clear, and the person appearance features extracted by the model are limited for some poorly lit person images. When the model introduces DFW, the discriminative ability of the person's appearance features is weakened to some extent, causing reduced feature similarity between intra-class images and some person matching errors. From Fig 3 (b), it is overt that the model works better on the RegDB dataset when the threshold is m and the weakening factor is 0.2. The Rank1 accuracy is improved by 3.37%, and mAP is improved by 2.67% compared to the baseline model. The accuracy of Rank1 and mAP in other groups of experiments have also improved. Thus, it can be concluded that the model performs best when the threshold is the mean value of the feature map.

For different datasets, the optimal value of the weakening factor is also different. When the weakness factor is too small, the network does not sufficiently learn the person's body features. If the weakening factor is too large, the network will extract some meaningless background features, which can affect the recognition accuracy of the model. Overall, DFW has a certain improvement effect on the accuracy of the model, but for the problem of decreasing discriminability of modality-invariant features, it is necessary to further enhance the learning of important fine-grained cues in person images.

The effect of the separation parameter. The role of the separation parameter S is to constrain the distance between the image areas concerned by the sub-network segments to avoid the problem of decreasing spatial feature diversity owing to repeated learning. We explored the effect of S on the model recognition accuracy under the condition of weight coefficient $\lambda = 1$. From Fig 4 (a), it is evident that the model achieves the best performance when $S = 100$ on the SYSU-MM01 dataset, with Rank1 accuracy at 57.88% and mAP at 55.09%. Compared with the baseline model (as show in Fig 3), the Rank1 accuracy and mAP have increased by 3.07% and 3.34%, respectively. Similarly, in Fig 4 (b), the accuracy of Rank1 is 89.47% and mAP is 84.75% when $S = 100$ on the RegDB dataset, which is an improvement of 4.01% and 7.12%, respectively, compared to the baseline model. From Fig 4, it can be observed that the model performs better on both datasets when the separation parameter S is set to 100. When the separation parameter is set to other values, the accuracy of the model decreases. This suggests that when the separation parameter is too small, the focus areas of the network are too close or even overlap, leading to the learned features becoming similar and thereby decreasing feature diversity. Conversely, when the separation parameter is too large, some valuable fine-grained cues are ignored by each sub-network, resulting in a decrease in the model's accuracy.

The effect of the weight coefficient. The role of the weight coefficient λ is to constrain the output of L_{op} and ensure the stability of the network. As can be seen from Fig 4, the model shows better results when the separation parameter is set to 100, which indicates that the distance between the areas concerned by the sub-network is more appropriate at this time. Therefore, we explore the effect of the weight coefficient λ on the recognition accuracy



(a) SYSU-MM01

(b) RegDB

Figure 5: Exploring the effect of different weight coefficients λ on model accuracy on SYSU-MM01 and RegDB datasets.

of the model by fixing the separation parameter $S = 100$. From Fig 5 (a), it can be seen obviously that the model effect reaches the best when $\lambda = 0.5$ on the SYSU-MM01 dataset, the accuracy of Rank1 is 58.93%, and mAP is 56.56%. Compared with the baseline model, its Rank1 accuracy and mAP have increased by 4.12% and 4.81%, respectively. When $\lambda \in [0.1, 1]$ and $\lambda \neq 0.5$, the accuracy of Rank1 and mAP decreases, which indicates that when weight coefficient are too small, the constraints on L_{op} is more relaxed, increasing the instability of the algorithm. And too large weight coefficient make the constraint overly strict, leading to poor generalization ablity of the model. As can be seen from Fig 5 (b) that on the RegDB dataset, the model achieves the best effect when $\lambda = 1$, and its Rank1 accuracy and mAP are improved by 4.01% and 7.12%, respectively. Compared with the baseline model, while the model accuracy decreases when λ take other values. The experimental analysis combined with Fig 4 (b) separation parameter S shows that for the RegDB dataset, the model can learn more comprehensive features when $\lambda = 1$ and $S = 100$, and no additional constraints on the overlap penalty loss output are needed at this time.

4.5. Ablation study

To verify the effectiveness of the proposed components in our model, we record the changes in model accuracy after adding each component in Table 1. Among them, 'DFW' indicates dynamic feature weakening, 'SLN' indicates a segmented learning network, and 'RE' indicates a random erasure strategy.

Table 1: Ablation Study of CMC (%) and mAP (%) performances on the SYSU-MM01 and RegDB datasets.

DFW	SLN	RE	SYSU-MM01				RegDB			
			R1	R10	R20	mAP	R1	R10	R20	mAP
✗	✗	✗	54.81	88.08	94.53	51.75	85.46	93.59	95.72	77.63
✓	✗	✗	56.68	88.82	94.55	51.12	88.83	95.70	97.27	80.30
✗	✓	✗	58.93	90.94	96.39	56.56	89.47	95.59	97.43	84.75
✗	✗	✓	58.01	90.11	95.96	55.72	89.05	94.85	98.26	84.08
✓	✗	✓	58.20	90.42	95.71	52.10	89.33	95.62	96.86	82.39
✓	✓	✗	59.48	92.03	97.18	57.69	89.82	96.01	97.69	85.28
✗	✓	✓	60.59	93.36	98.14	59.34	91.73	97.02	98.36	88.19
✓	✓	✓	64.24	93.83	98.20	62.04	93.18	97.71	98.73	89.83

Note that the reported results are in the all-search mode for the SYSU-MM01 dataset while in the visible-thermal mode for the RegDB dataset.

Firstly, the baseline model is trained under the supervision of person Re-ID loss L_{ReID}, which consists of modality-specific module, modality-shared module, and person Re-ID branch. In the second to fourth rows of the table, DFW, SLN, and RE are introduced into the model, respectively. Compared with the baseline model, the model with the introduction of each method has a significant improvement in recognition rate on both datasets. Additionally, we also combined the two methods arbitrarily in turn, and their model accuracy improved in all cases. Considering the significant differences between multi-modality images, although the network relies on DFW to capture more comprehensive modality-invariant features, some discriminative detail information may not be fully learned. Furthermore, it can be seen from the experimental results that the model only captures local information of people for recognition, which does not achieve the best results. Finally, we incorporate DFW, SLN and RE into the last row of the model. By weakening the high response region, we capture the modality-invariant features of multi-modality person images as much as possible, and further enhance the learning of modality-invariant features by mining the fine-grained informa-

Table 2: Comparison of computational costs of different methods on RegDB dataset.

Methods	Params/M	FLOPs/G	Train time/min	R1	mAP
Baseline	23.74	10.301	≈ 45	85.46	77.63
AGW[24]	23.55	10.318	≈ 47	70.05	66.37
MMG[4]	23.53	36.633	≈ 120	91.60	84.10
Ours	26.71	15.150	≈ 58	93.18	89.83

tion of the person. Meanwhile, RE is used to erase part of the input images, which greatly improves the recognition accuracy of the model.

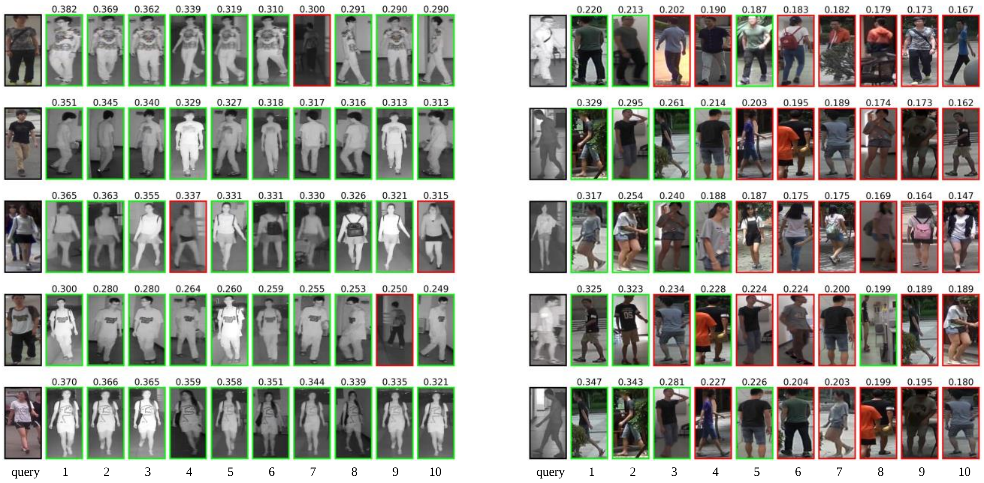
4.6. Calculated cost analysis

The computational costs of the different methods on the RegDB dataset are shown in Tab 2. We ran the AGW and MMG codes on the same experimental platform and environment, and to avoid experimental results being affected by parameters and number of images, we used the same batches and did not make changes to the hyperparameters of the model.

Since the model uses a segmented learning network consisting of two branches at layer4 stage, the metrics of parameters and FLOPs are increased compared to the baseline model and AGW which only uses a single-stream network. However, it can be seen from the experimental results that the accuracy of the proposed model in this paper has a significant improvement compared with the baseline model and AGW. In addition, the FLOPs and training time of MMG are as high as 36.633/G and 120/min, while our FLOPs and training time are only one-half of that method. Meanwhile, our Rank1 and mAP are higher by 1.58% and 5.73%, respectively. This validates the efficiency of our method.

4.7. Visualization

We randomly selected 5 visible images and 5 infrared images from the SYSU-MM01 dataset as query samples and visualized the retrieval results in all-search mode. The retrieved images are sorted from left to right according to the similarity magnitude, and the similarity score is displayed at the top of each image, as Fig 6 shows. From the retrieval results of the visible-infrared image in Fig 6(a), it can be observed that the top-ranked images are



(a) Visible to infrared retrieval results

(b) Infrared to visible retrieval results

Figure 6: Visualization of top-10 retrieval results of some query samples in the SYSU-MM01 dataset with the model in this paper.

more similar to their corresponding query samples, in terms of clothing type, body contour, and other cues. This indicates that a person's appearance information is crucial for the model's recognition. However, the retrieval lists in the first, third, and fourth rows show that some person images are incorrectly matched, mainly due to the overall darkness of the image and the distance between the shooting device and the person. The top-ranked images in Fig 6 (b) show that the retrieved images have different colors as a result of the lack of important color information in infrared images. Moreover, some matching error images suggest that similar person postures increase the difficulty of infrared-visible image retrieval, resulting in persons of different genders being considered as having the same identity. In general, the retrieval of infrared-visible images is much more challenging than visible-infrared image retrieval, primarily due to the low resolution, poor visual effect, and lack of obvious color information of the infrared image. The use of infrared images as query samples can lead to lower model recognition rates due to the limitation of available information.

4.8. Comparison with the state-of-the-art methods

In this section, we compare the proposed model with some state-of-the-art (SOTA) methods and use re-ranking with k-reciprocal encoding. The

Table 3: Comparison of CMC (%) and mAP (%) performances with the state-of-the-art methods on the SYSU-MM01 dataset.

	All-Search				Indoor-Search			
Methods	R1	R10	R20	mAP	R1	R10	R20	mAP
Two-stream[5]	11.65	47.99	65.50	12.85	15.60	61.18	81.02	21.49
Zero-padding[5]	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
BDTR[8]	27.32	66.96	81.07	27.32	31.92	77.18	89.28	41.86
D2RL[25]	28.90	70.60	82.40	29.20	-	-	-	-
CC-S[26]	33.40	78.60	89.40	37.20	-	-	-	-
DGD_MSR[9]	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
Hi-CMD[26]	34.94	77.58	-	35.94	-	-	-	-
EDFL[7]	36.90	84.50	93.20	40.70	-	-	-	-
BEAT[23]	38.57	76.64	86.39	38.61	-	-	-	-
CMPG[5]	43.40	81.30	90.50	38.00	46.80	88.20	94.70	54.70
AGW[24]	47.50	-	-	47.65	54.17	-	-	62.97
LbA[27]	55.41	-	-	54.14	61.02	-	-	66.33
MBN[3]	56.07	88.59	94.75	54.28	-	-	-	-
HC[11]	56.96	91.50	96.82	54.95	59.74	92.07	96.22	64.91
WIT[28]	59.20	91.70	96.50	57.30	60.70	94.10	98.40	67.10
HCT[20]	61.68	93.10	97.17	57.51	63.41	91.69	95.28	68.17
MTMFE[21]	62.56	**93.85**	97.63	60.57	65.06	95.17	98.17	73.86
Ours	**64.24**	93.83	**98.20**	**62.04**	**67.97**	**97.68**	**99.72**	**74.09**
Ours(RK)	74.13	95.77	98.84	72.95	81.35	96.32	98.61	83.96

experimental results of the model on the SYSU-MM01 dataset and RegDB dataset are shown in Table 3 and Table 4, respectively.

As can be seen from Table 3, our proposed model outperforms the SOTA model significantly on the SYSU-MM01 dataset. Among them, in the all-search mode, compared with HCT, the accuracy of Rank1 is increased by 2.56%, and mAP is increased by 4.53%. Compared with MTMFE, the accuracy of Rank1 and mAP are improved by 1.68% and 1.47%, respectively. In the indoor-search mode, compared with HCT, the accuracy of Rank1 of the model is increased by 4.56%, and mAP is increased by 5.92%. Compared with MTMFE, the accuracy of Rank1 and mAP are increased by 2.91% and 0.23%, respectively. From the comparison results of Table 4 in RegDB, it can be seen that in the visible-thermal retrieval mode, compared with HCT, the

Table 4: Comparison of CMC (%) and mAP (%) performances with the state-of-the-art methods on the RegDB dataset.

	Visible to Thermal				Thermal to Visible			
Methods	R1	R10	R20	mAP	R1	R10	R20	mAP
Two-stream[5]	12.40	30.4	41.00	13.40	-	-	-	-
Zero-padding[5]	17.80	34.2	44.40	18.90	16.63	34.68	44.25	17.82
BDTR[8]	30.56	54.62	65.42	32.45	-	-	-	-
D2RL[25]	43.40	66.10	76.30	44.10	-	-	-	-
DGD_MSR[9]	48.43	70.32	79.59	48.67	-	-	-	-
Hi-CMD[13]	70.93	86.39	-	66.04	-	-	-	-
EDFL[7]	52.58	72.10	81.47	52.98	51.89	72.09	81.04	52.13
CMPG[5]	52.10	-	-	51.90	51.30	-	-	52.00
BEAT[23]	67.45	-	-	66.51	66.48	-	-	67.31
MBN[3]	67.80	-	-	65.50	66.20	-	-	64.20
AGW[24]	70.05	-	-	66.37	-	-	-	-
LbA[27]	74.17	-	-	67.64	72.43	-	-	65.64
WIT[28]	85.00	96.90	98.80	75.90	-	-	-	-
MMN[4]	91.60	97.70	97.70	84.10	87.50	96.00	98.10	80.50
HCT[20]	91.05	97.16	98.57	83.28	89.30	96.41	**98.16**	81.46
MTMFE[21]	76.10	88.86	92.41	74.39	72.18	87.06	92.38	71.04
Ours	**93.18**	**97.71**	**98.73**	**89.83**	**92.86**	**97.27**	98.15	**89.47**
Ours(RK)	94.31	97.84	98.22	93.64	93.46	94.64	98.53	93.27

accuracy of Rank1 is increased by 2.13%, and mAP is increased by 6.55%. Compared with MTMFE, the accuracy of Rank1 and mAP are improved by 17.08% and 15.44% respectively. Similarly, our model also shows certain advantages in the thermal-visible retrieval mode. Compared with HCT, the accuracy of Rank1 is increased by 3.56%, and mAP is increased by 8.01%. Compared with MTMFE, the accuracy of Rank1 and mAP are increased by 20.68% and 18.43%, respectively. With the addition of re-ranking, the accuracy of the model was significantly improved, in which the model obtained 74.13% accuracy for Rank1 on the SYSU-MM01 dataset, which verified the effectiveness of the method in this paper.

HCT and MTMFE mainly rely on partial cues in the shared space for recognition, leading to the underutilization of multi-modality features, and the spatially cut approach further affects the retrieval accuracy of a person. In contrast, our model uses DFW to enrich and diversify the modality-invariant information. SLN is utilized to autonomously focus on and learn fine-grained features in different regions of a person to ensure the integral learning of diverse features, effectively improving the model's recognition capability

5. Conclusion

In this paper, we design a dynamic feature weakening (DFW) method to train a deep person Re-ID model by providing more comprehensive modality-invariant information, which reduces inter-modality differences. Furthermore, we also utilize a segmented learning network (SLN) to flexibly mine fine-grained cues from different parts of the person, which helps alleviate interference due to similar person appearance under heterogeneous modalities. Experiments at SYSU-MM01 and RegDB show that our proposed method outperforms some traditional cross-modality person Re-ID methods. The method is easy to understand without involving tedious steps, and the DFW can be easily integrated into other models as a plug-and-play solution. Due to the large offset between heterogeneous data distributions, the uniform feature space learning approach limits the expression of useful information. In the future, we plan to build a modality representation generator in front of the shared space to enhance the recognition capability of the model by adaptively generating intermediate representations with multi-modality characteristics.

References

[1] X. Li, Y. Lu, B. Liu, Y. Liu, G. Yin, Q. Chu, J. Huang, F. Zhu, R. Zhao, N. Yu, Counterfactual intervention feature transfer for visible-infrared person re-identification, in: European Conference on Computer Vision, 2022. doi:10.1007/978-3-031-19809-0_22.

[2] B. Gaikwad, A. Karmakar, End-to-end person re-identification: Real-time video surveillance over edge-cloud environment, Comput. Electr. Eng. 99 (2022) 107824. doi:10.1016/j.compeleceng.2022.107824.

[3] W. Li, K. Qi, W. Chen, Y. Zhou, Bridging the distribution gap of visible-infrared person re-identification with modality batch normalization, in: 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2021, pp. 23–28. doi:10.1109/ICAICA52286.2021.9498199.

[4] Y. Zhang, Y. Yan, Y. Lu, H. Wang, Towards a Unified Middle Modality Learning for Visible-Infrared Person Re-Identification, Proceedings of the 29th ACM International Conference on Multimedia (2021) 788–796. doi:10.1145/3474085.3475250.

[5] Y. Yang, T. Zhang, J. Cheng, Z. Hou, P. Tiwari, H. M. Pandey, et al., Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification, Neural Networks 128 (2020) 294–304. doi:10.1016/j.neunet.2020.05.008.

[6] L. Tan, P. Dai, R. Ji, Y. Wu, Dynamic prototype mask for occluded person re-identification, Proceedings of the 30th ACM International Conference on Multimedia (2022) 531–540. doi:10.1145/3503161.3547764.

[7] H. Liu, J. Cheng, Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification, Neurocomputing 398 (2020) 11–19. doi:10.1016/j.neucom.2020.01.089.

[8] M. Ye, X. Lan, Z. Wang, P. C. Yuen, Bi-directional center-constrained top-ranking for visible thermal person re-identification, IEEE Transactions on Information Forensics and Security 15 (2019) 407–419. doi:10.1109/TIFS.2019.2921454.

[9] G. Gao, H.-C. Shao, Y. Yu, F. Wu, M. Yang, Leaning compact and representative features for cross-modality person re-identification, World Wide Web 25 (2022) 1649–1666. doi:10.1007/s11280-022-01014-5.

[10] X. Xu, K. Lin, L. Gao, H. Lu, H. T. Shen, X. Li, Learning cross-modal common representations by private–shared subspaces separation, IEEE Transactions on Cybernetics 52 (2020) 3261–3275. doi:10.1109/TCYB.2020.3009004.

[11] Y. Gao, T. Liang, Y. Jin, X. Gu, W. Liu, Y. Li, C. Lang, Mso: Multi-feature space joint optimization network for rgb-infrared person re-identification, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5257–5265. doi:10.1145/3474085.3475643.

[12] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, D. Tao, Hetero-center loss for cross-modality person re-identification, Neurocomputing 386 (2020) 97–109. doi:10.1016/j.neucom.2019.12.100.

[13] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, H. T. Shen, Cross-modal attention with semantic consistence for image–text matching, IEEE Transactions on Neural Networks and Learning Systems 31 (2020) 5412–5425. doi:10.1109/TNNLS.2020.2967597.

[14] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, S. Tian, Feature refinement and filter network for person re-identification, IEEE Transactions on Circuits and Systems for Video Technology 31 (9) (2020) 3391–3402. doi:10.1109/TCSVT.2020.3043026.

[15] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 480–496. doi:10.1007/978-3-030-01225-0_30.

[16] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 274–282. doi:10.1145/3240508.3240552.

[17] Z. He, H. Zhao, J. Wang, W. Feng, Pose matters: Pose guided graph attention network for person re-identification, Chinese Journal of Aeronautics (2022). doi:10.1117/12.2646845.

[18] Z. Feng, J. Lai, X. Xie, Learning view-specific deep networks for person re-identification, IEEE Transactions on Image Processing 27 (7) (2018) 3472–3483. doi:10.1109/TIP.2018.2818438.

[19] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, S. Zhang, Towards rich feature discovery with class activation maps augmentation for person re-identification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1389–1398. doi:10.1109/CVPR.2019.00148.

[20] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, Rgb-infrared cross-modality person re-identification, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5380–5389. doi:10.1109/ICCV.2017.575.

[21] C. Ma, X. Li, Y. Li, X. Tian, Y. Wang, H. Kim, S. Serikawa, Visual information processing for deep-sea visual monitoring system, Vol. 1, 2021, pp. 3–11. doi:10.1016/j.cogr.2020.12.002.

[22] H. Liu, X. Tan, X. Zhou, Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification, IEEE Transactions on Multimedia 23 (2020) 4414–4425. doi:10.1109/TMM.2020.3042080.

[23] N. Huang, J. Liu, Q. Zhang, J. Han, Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification, Pattern Recognit. 135 (2023) 109145. doi:10.1016/j.patcog.2022.109145.

[24] S. Zhang, Y. Yang, P. Wang, G. Liang, X. Zhang, Y. Zhang, Attend to the difference: Cross-modality person re-identification via contrastive correlation, IEEE Transactions on Image Processing 30 (2021) 8861–8872. doi:10.1109/TIP.2021.3120881.

[25] D. T. Nguyen, H. G. Hong, K. W. Kim, K. R. Park, Person recognition system based on a combination of body images from visible light and thermal cameras, Sensors 17 (3) (2017) 605. doi:10.3390/s17030605.

[26] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, S. 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, 2019, pp. 618–626. doi:10.1109/CVPR.2019.00071.

[27] S. Choi, S. Lee, Y. Kim, T. Kim, C. 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, 2020, pp. 10257–10266. doi:10.1109/CVPR42600.2020.01027.

[28] Y. Nakayama, H. Lu, Y. Li, T. Kamiya, Widesegnext: Semantic image segmentation using wide residual network and next dilated unit, IEEE Sensors Journal 21 (2021) 11427–11434. doi:10.1109/JSEN.2020.3008908.

latex file of manuscript