PLOS ONE

RESEARCH ARTICLE

Channel semantic mutual learning for visible-thermal person re-identification

Yingjie Zhu[1,2], **Wenzhong Yang**[2]*

1 College of Software, Xinjiang University, Urumqi, China, **2** Xinjiang Multilingual Information Technology Laboratory, Xinjiang University, Urumqi, China

* ywz_xy@163.com

Citation: Zhu Y, Yang W (2024) Channel semantic mutual learning for visible-thermal person re-identification. PLoS ONE 19(1): e0293498. https://doi.org/10.1371/journal.pone.0293498

Editor: Feng Ding, Nanchang University, CHINA

Received: June 6, 2023

Accepted: October 13, 2023

Published: January 19, 2024

Data Availability Statement: RegDB dataset is available from http://dm.dongguk.edu/link.html SYSU-MM01 dataset is available from http://isee.sysu.edu.cn/project/RGBIRReID.htm.

Funding: This is supported by the [Research and Application of Multilingual and Multimodal Information Content Security] grant number [No.202304120002],[National Natural Science Foundation of China] grant number [No.202204120017], [Autonomous Region Special Research and Development Task] grant number [No. 2022B01008-2], [Autonomous Region Major Science and Technology Special Project] grant number [No. 2020A02001-1] and [Optimization of

Abstract

Visible-infrared person re-identification (VI-ReID) is a cross-modality retrieval issue aiming to match the same pedestrian between visible and infrared cameras. Thus, the modality discrepancy presents a significant challenge for this task. Most methods employ different networks to extract features that are invariant between modalities. While we propose a novel channel semantic mutual learning network (CSMN), which attributes the difference in semantics between modalities to the difference at the channel level, it optimises the semantic consistency between channels from two perspectives: the local inter-channel semantics and the global inter-modal semantics. Meanwhile, we design a channel-level auto-guided double metric loss (CADM) to learn modality-invariant features and the sample distribution in a fine-grained manner. We conducted experiments on RegDB and SYSU-MM01, and the experimental results validate the superiority of CSMN. Especially on RegDB datasets, CSMN improves the current best performance by 3.43% and 0.5% on the Rank-1 score and mINP value, respectively. The code is available at https://github.com/013zyj/CSMN.

1 Introduction

Person re-identification (ReID) [1] is a technology that employs computer vision algorithms to locate and retrieve a pre-defined individual from non-overlapping camera views. Previous studies [2–7] have mainly focused on ReID in visible light, capturing all images of a person with visible light cameras. However, visible light cameras may not capture a person's appearance at night. As a result, VI-ReID [8] is proposed.

Compared to single-modality ReID, VI-ReID faces the problem of intra-class variations, such as illumination and occlusion, and the challenge of significant modality discrepancy. Therefore, VI-ReID is more challenging. Currently, common methods for VI-ReID mainly include the following aspects: On the one hand, modality-invariant features are extracted to address the cross-modality problem [9, 10]. However, modality-invariant features are frequently difficult to ensure quality, leading to the loss of information in pedestrian image representations. On the other hand, using GAN methods for cross-modality transformation [11–14] can convert cross-modality matching problems into within-modality matching tasks to improve retrieval accuracy. However, such methods inevitably increase the computational complexity of the model and introduce noise, resulting in poor model performance. In addition, some work has been devoted to improving the

low-resolution device defect recognition algorithm based on image enhancement] grant number [No. SGXJXTOOJFJS2200076]. The funder's role in this research includes study design and decision to publish.

Competing interests: The authors have declared that no competing interests exist.

performance of metric learning methods [15–18]. But the above methods only learn the sample distribution at the instance level and lack handling of outlier samples.

To reduce the discrepancy between the channel semantics within a modality and between modalities, we designed a novel Channel Semantic Mutual Learning Network (CSMN), which simultaneously learns channel semantic consistency from two aspects: Intra-Modality Channel Semantic Mutual Learning(ICSM), which focuses on learning fine-grained information by increasing the similarity of feature distributions between channels, and Cross-Modality Channel Semantic Mutual Learning (CCSM), aiming at learning global information by reducing the distance between feature distributions across modalities. In addition, we proposed a Channel-level Auto-directed Metric Learning loss (CADM) to optimise intra-class and inter-modality feature distributions in a more fine-grained manner. Specifically, our approach reduces intra-class instance discrepancies and aggregates semantic information for the same identity while also narrowing the gap between modalities by strengthening the correlation between semantic information for the same identity across different modalities. Additionally, we designed an auto-guided function to mitigate the generation of noisy samples. Since infrared images cannot be viewed as normal RGB images, we use the gray-to-color method to convert infrared images to colored images. Fig 1 shows the overall structure of the model.

In summary, the main contributions of this paper are:

We propose a channel semantic mutual learning network (CSMN) for VI-ReID that treats modality discrepancy as inter-channel discrepancy and reduces intra-modality channel discrepancy while learning inter-modal channel information to bridge modality discrepancy.

We suggest a channel-level auto-guided double metric loss (CADM) to optimise the sample distribution intra- and inter-modality through multiple aspects, including reducing the intra-class instance differences, strengthening the correlation between the same identity across different modalities, and handling outlier samples.

We have conducted numerous experiments on two benchmarks. Specifically, on SYSU-MM01, CSMN achieves state-of-the-art performance and improves the Rank-1 score of the current best performance on the RegDB dataset by 3.43%.

2 Related work

2.1 Single-modality ReID

Single-modality ReID attempts to retrieve a specific person from a library of images obtained from different cameras during the day, where the images obtained have the same modalities.

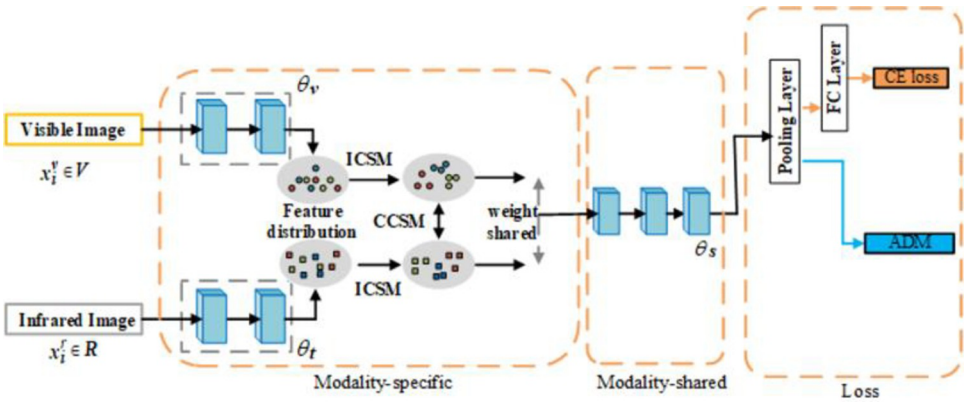

Fig 1. The figure shows the overall structure of the model.

https://doi.org/10.1371/journal.pone.0293498.g001

Person re-identification (ReID) methods have greatly improved as deep learning technology has advanced. Many methods have focused on building local-based models to fully explore fine-grained features in a person's images [19–21]. Fu et al. [10] learned local features at different scales using a pyramid structure and eventually obtained multi-scale fused features. Lian et al. [20] designed an attention-aligned network for feature learning that uses channel and spatial attention. Wang et al. [21] proposed a multi-branch network where one branch captures global representations and the other branch focuses on local information. In addition, attention models [22–26] are essential for designing novel network architectures that highlight salient regions and alleviate misalignment to learn robust features.

2.2 Visible-infrared person re-identification

VI-ReID is to match and identify the same pedestrian between different cameras, not different modalities. Wu et al. [8] published SYSU-MM01dataset and proposed a model to extract modality-shared person features. Dai et al. [27] suggested that cmGAN reduces the modality differences between visible and infrared images. Thus, dual-stream networks have been widely used to address modality discrepancy problems [16, 28]. Ye et al. [16] proposed a model to address intra-class variation caused by viewpoint non-variation. Ye et al. [28] proposed a novel DDAG learning method for VI-ReID by mining modal contextual cues. However, the methods above focus on reducing modality differences at the instance level. At the same time, this paper aims to learn more discriminative clues at the channel level, enabling semantic consistency between channels.

2.3 Metric learning

Metric learning plays a crucial role in inter-sample similarity measures for Re-ID. Ye et al. [16] provided a loss to learning discriminative feature representations using a two-stream network [29]. They also introduced a major constraint to enhance performance [30]. To reduce intra- and cross-modal variation, Hao et al. [31] proposed a network with classification and recognition constraints. Zhao et al. [32] introduced the hard pentaplet loss to improve VI-ReID performance. Wu et al. [33] designed a novel loss for focal modality-aware that guides inter-modal similarity learning with intra-modal similarity. However, the above methods only use the Euclidean metric, which cannot learn modality-shared discriminative features from multiple perspectives. And ignore the impact of noisy samples on model performance.

3 Method

In this section, as shown in Fig 1, we introduce CSMN, which consists of three parts: Modality-specific, Modality-shared and Loss, with the Modality-specific containing two essential elements: 1) Intra-modality Channel Semantic Mutual Learning (ICSM), which reduces differences between instances by learning semantic information among channels of instances within the same modality. 2) Cross-modality Channel Semantic Mutual Learning (CCSM), which learns the relationships between channels of different modalities and aggregates semantic information of the same identity at the channel level. Then the features learned from different branches go through Modality-shared for further feature learning. In terms of metric learning, we design Channel-level Auto-guided Double Metric loss (CADM), which optimizes the distribution of samples within and between modalities.

3.1 Intra-Channel Semantic Mutual Learning

RGB image channels contain different semantic information and have certain correlations. As depicted in Fig 2, modality-specific features are extracted by specific feature layers. As visible light and infrared images are captured based on different imaging principles, modality-specific features correspond to different semantic information for the same identity. Since infrared images are obtained based on the temperature distribution on the surface of objects, they cannot be treated as ordinary three-channel images. In this paper, we attribute the differences between modalities to differences between channels. Hence, the key to this problem is to ensure the identity correlation of channel features and reduce semantic changes between channels. Since the extended three-channel infrared image exhibits heterogeneity in the R/G/B channels, we aim to train the network to learn R/G/B channel distributions similar to those of visible images. To reach this goal, we made an Intra-Modality Channel Semantic Mutual Learning (ICSM) module, as shown in Fig 2, which uses the colours red, blue, and green to show how similar the channel feature distributions are to each other. Our method focuses on maximizing the intra-modality channel-level semantic consistency within each modality. We represent channel-level consistency as the logical distribution similarity between channel features. It can be formulated as follows:

$$L_{\mathrm{ICMC}}(\theta_v,\theta_t)=\frac{1}{2}*\left[\left(\sum_{i=1}^{N} f_i^{R_v}.log\frac{f_i^{R_v}}{f_i^{G_v}+f_i^{R_v}}+f_i^{R_t}.log\frac{f_i^{R_t}}{f_i^{G_t}+f_i^{R_t}}\right)\right.$$

$$+\left(\sum_{i=1}^{N} f_i^{B_v}.log\frac{f_i^{B_v}}{f_i^{G_v}+f_i^{B_v}}+f_i^{B_t}.log\frac{f_i^{B_t}}{f_i^{G_t}+f_i^{B_t}}\right)$$

$$\left.+\left(\sum_{i=1}^{N} f_i^{B_v}.log\frac{f_i^{B_v}}{f_i^{R_v}+f_i^{B_v}}+f_i^{B_t}.log\frac{f_i^{B_r}}{f_i^{R_t}+f_i^{B_t}}\right)\right] \quad (1)$$

where L_{ICMC} represents the semantic consistency between the three channels, and ICSM aims to minimize the semantic difference between channels. To achieve the above goals, the following formula is used to optimize the parameters θ_v and θ_t:

$$(\hat{\theta}_v\hat{\theta}_t)=\arg\min(L_{\mathrm{ICMC}}(\theta_v,\theta_t)) \quad (2)$$

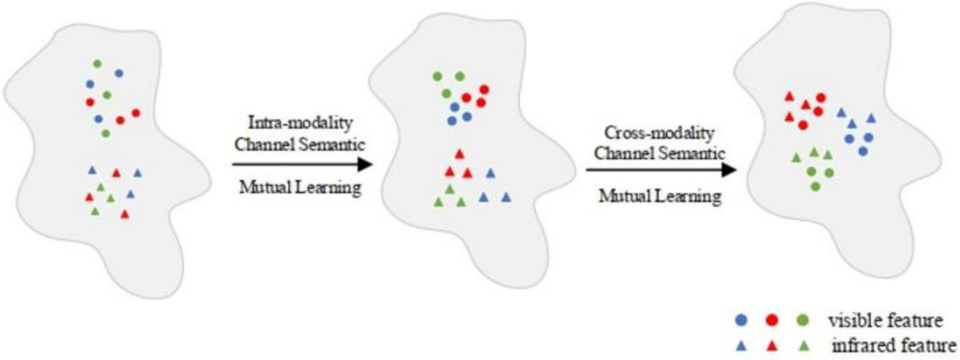

Fig 2. Channel semantic mutual learning.

https://doi.org/10.1371/journal.pone.0293498.g002

3.2 Cross-Modality Channel Semantic Mutual Learning

In addition, to reduce the differences between modalities, we propose the Cross-Modality Channel Semantic Mutual Learning (CCSM) method, which aims to maximize the inter-modality channel semantic consistency. This method uses inter-modality semantic consistency to aggregate features from different modalities under the same identity. We further reduce the inter-modality channel semantic differences based on the modality-specific semantic consistency features. Since the modality-specific extractors θ_v and θ_t extract features within each modality, the extracted modality-specific features are independent. The following formula can represent the features of each modality:

$$C_v = (1/\sum_{i=1}^{S_v} w_i^v) * \sum_{i=1}^{S_v} w_i^v * f_i^v \tag{3}$$

$$C_t = (1/\sum_{i=1}^{S_t} w_i^t) * \sum_{i=1}^{S_t} w_i^t * f_i^t \tag{4}$$

where S_t and S_v represent the number of samples in the infrared and visible images and represent the weights of the i-th feature vector in different modalities, which are adjusted to reduce the influence of outliers. C_v and C_t are batch-computed. CCSM aims to learn semantic information between modalities rather than identity information. Using metric learning enables the alignment of the distance between C_v and C_t, and the features θ_v and θ_t will have more semantic consistency information. In CCSM, the goal is to maximize the inter-modality semantic consistency between visible and infrared image features:

$$L_{ccsm} = L(\theta_v, \theta_t) = \|C_v - C_t\|^2 \tag{5}$$

Fig 3 shows the collaborative processing process of ICSM and CCSM. The combination of the two can not only reduce the differences between instances of the same identity within each modality and improve the matching accuracy of the same identity between modalities.

3.3 ADM

Most existing metric learning methods are at the instance level, a coarse-grained learning method that is also vulnerable to the influence of noisy samples. To learn fine-grained features and reduce the impact of noisy samples on the feature space, we propose channel-level auto-guided double metric loss (CADM). We obtain semantically consistent features f^v and f^t from

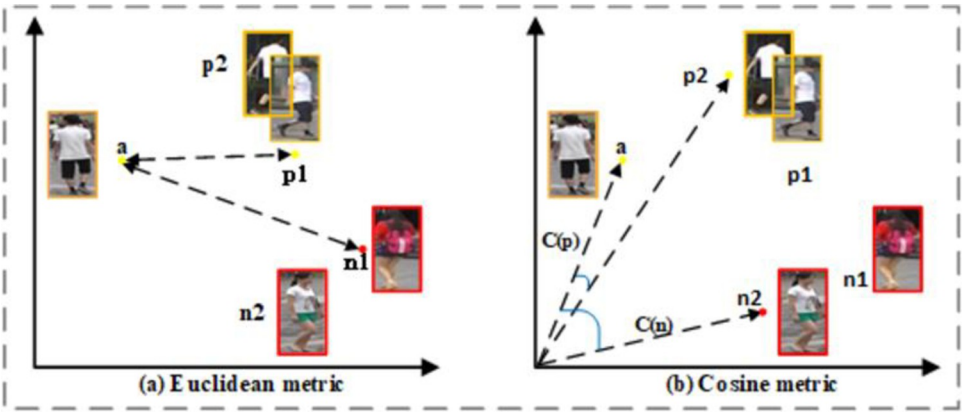

Fig 3. The diagram depicts the single and double metric learning methods. Where *C(p)* and *C(n)* represent the cosine values between the positive and negative samples.

https://doi.org/10.1371/journal.pone.0293498.g003

the modality-specific extractors θ_v and θ_t respectively. Then we use a weight-shared feature extractor θ_s to obtain rich semantic features $[f_i^{R_v}, f_i^{G_v}, f_i^{B_v}]$, $[f_i^{R_t}, f_i^{G_t}, f_i^{B_t}] \in R^{B\times C\times H\times W}$. Since different metric learning methods will learn different the hardest samples. For the Euclidean metric, $p1$ and a in Fig 3(A) are the pair of positive samples, and n1 and a are the pair of negative samples, but for the cosine metric in Fig 3(B), $p2$ and a are only the pair of positive samples, and n2 and a are the pair of negative samples. So, to learn the sample distribution from multiple perspectives, we propose a double metric loss(DM), which introduces a cosine metric that takes the direction of the feature vector into account based on the Euclidean metric:

$$L_e = \|f_i^{R_v} - f_i^{R_t}\|^2 + \|f_i^{G_v} - f_i^{G_t}\|^2 + \|f_i^{B_v} - f_i^{B_t}\|^2 \tag{6}$$

$$L_c = \left(1 - \frac{f_i^{R_v} \cdot f_i^{R_t}}{\|f_i^{R_v}\|\|f_i^{R_t}\|}\right) + \left(1 - \frac{f_i^{G_v} \cdot f_i^{G_t}}{\|f_i^{G_v}\|\|f_i^{G_t}\|}\right) + \left(1 - \frac{f_i^{B_v} \cdot f_i^{B_t}}{\|f_i^{B_v}\|\|f_i^{B_t}\|}\right) \tag{7}$$

Additionally, to deal with noisy samples, we introduce an auto-guided function. Specially, we utilize the Euclidean metric to calculate the similarity between samples and construct the corresponding similarity matrix. We initially extract the positions of all positive and negative samples from the distance matrix using the Euclidean metric to generate a position mask. The position mask calculates the distances between each sample in the distance matrix. All sample distances are then combined using the proposed auto-guided function. The following is the auto-guided function.

$$\mathrm{P(x)} = \begin{cases} 0, & \mathrm{x} < 0 \\ \dfrac{x}{2\mathrm{d}} + \delta, & 0 < \mathrm{x} < 2d \\ \mathrm{x}, & \mathrm{x} > 2\mathrm{d} \end{cases} \tag{8}$$

where d is the constant slope that controls the auto-guided function. δ is a very small constant that ensures the function value is greater than zero.

The CADM can finally be expressed as:

$$L_{cadm} = L_e + b \cdot L_c + c \cdot L_p \tag{9}$$

Where c are the auto-guided function loss coefficients of L_p.Therefore, the final expression of the loss function is as follows:

$$L_{total} = L_{ICMC} + L_{ccsm} + L_{cadm} \tag{10}$$

4 Experiment

4.1 Experimental settings

4.1.1 Dataset and setting. We evaluate the performance of our proposed approach on the VI-ReID task through experiments on two widely used benchmark datasets, SYSU-MM01 [8] and RegDB [34]. The SYSU-MM01 dataset, the largest VI-ReID dataset, comprises four visible and two near-infrared cameras. The training set consists of 22,258 visible images and 11,909 thermal images of 395 individuals. The test set has 96 distinct identities, with 3,803 thermal images used as queries and 301 visible images used as galleries. We used single-shot outdoor and indoor search modes in our experiments. The dataset's configuration details can be found in [35]. The RegDB dataset consists of images

Table 1. Ablation experiments regarding L_{ICSM} and L_{CCSM} were studied on the RegDB dataset, where Base refers to CAJL [34].

Methods			RegDB(Visible to Infrared)			RegDB(Infrared to Visible)		
Base	L_{ICSM}	L_{CCSM}	Rank-1	mAP	mINP	Rank-1	mAP	mINP
√			85.03	79.14	65.33	84.75	77.82	61.56
√	√		86.50	77.25	64.89	85.67	76.02	59.34
√		√	88.01	79.99	65.59	85.98	77.08	62.42
√	√	√	88.11	80.06	65.75	86.21	77.12	62.54

https://doi.org/10.1371/journal.pone.0293498.t001

captured by one visible camera and one far-infrared camera. It contains 412 identities, each represented by 20 images (10 visible and 10 infrared) per person. According to the current VI-ReID settings [36], 206 identities are chosen randomly for training, and the remaining 206 identities are allocated to the test set.

4.1.2 Evaluation metrics. To assess the performance of our method, we use cumulative matching characteristics (CMC), mean average precision (mAP), and mean inverse negative penalty (mINP) [36]. mAP evaluates the retrieval system's performance when a gallery set contains multiple matched images. CMC measures the probability that the top-ranked retrieval results have the correct image of the person. mAP evaluates the retrieval system's performance when a gallery set contains multiple matched images. Furthermore, mINP considers the most difficult match to calculate the amount of work for inspectors.

4.1.3 Implementation details. We use CAJL [37] as the baseline network. The pre-trained weights of ImageNet are used to initialize the network parameters. We employ a PK sampling design with P = 8 and K = 4 parameters. We use zero-padded, randomly cropped images (288 × 144) as training data to supplement the original dataset. The SGD optimizer is used during the optimization process's learning phase. The learning rate is reduced from its initial value of 0.1 after 20 and 50 iterations. There are 100 training epochs in total. All tests were performed on an Nvidia 3090 GPU with PyTorch 1.6 and cuda11.0.

4.2 Ablation study

To verify the effectiveness of ICSM, CCSM, DM, and CADM, we conducted detailed experiments on the RegDB and SYSU-MM01dataset.

4.2.1 Effectiveness of Intra-Channel Semantic Mutual Learning (ICSM). As shown in Table 1, taking the visible to infrared mode as an example, based on the Base model, using only L_{ICSM} achieved a Rank-1 score of 86.5% and an mAP score of 77.25%. This improved the Rank-1 score of the baseline model by 1.47%. The experimental results show that ICSM can learn fine-grained features among channels within a modality, reducing the differences between channels within the same modality.

4.2.2 Effectiveness of Cross-Modality Channel Semantic Mutual Learning(CCSM). Unlike ICSM, CCSM learns global information between modalities to aggregate samples of the same identity across modalities. As shown in Table 1, L_{CCSM} represents the loss between modalities. Taking the visible-to-infrared mode as an example, using only L_{CCSM} on the Base model increased the Rank-1 score and mAP score by 2.98% and 0.85%, respectively. Additionally, we observed that using both L_{CCSM} and L_{ICSM} on the Base model further improved the model's performance. Specifically, compared to Base + L_{ICSM}, the Rank-1 and mAP scores improved by 1.61% and 2.81%, respectively. The Rank-1 score of Base + L_{CCSM} was improved by 0.1%, and the mAP score was improved by 0.07%. From the experimental results, CCSM can effectively reduce the differences between modalities, and combining ICSM and CCSM further enhances the model's performance.

Table 2. DM ablation experiment.

Method	Metric		RegDB		SYSU-MM01	
	Euclidean	cos	Rank-1	mAP	Rank-1	mAP
Baseline1	√		88.11	80.06	69.88	66.89
Baseline2		√	87.86	79.77	69.95	66.34
DM(ours)	√	√	**88.20**	**80.83**	**70.29**	**66.94**

https://doi.org/10.1371/journal.pone.0293498.t002

4.2.3 Effectiveness of Double Metric Loss (DM). As shown in Table 2, we conducted a series of experiments on the methods based on Euclidean and cosine metrics to demonstrate that the double metric consisting of Euclidean and cosine metrics can effectively improve the baseline performance. It should be noted that our experiments were conducted based on Base $+L_{ICSM}+L_{CCSM}$. Baseline1 only uses the Euclidean metric, whereas Baseline2 only uses the cosine metric. The DM-based baseline model outperformed Baseline1 on the RegDB dataset by 0.53% and 0.39% on Rank-1 and mAP, respectively, and by 0.41% and 0.05% on the SYSU-MM01 dataset. The experimental results show that DM can learn the feature distribution from multiple perspectives in a fine-grained manner, thereby improving the performance of the model.

4.2.4 Effectiveness of Channel-level Auto-guided Double Metric Loss (CADM). We compared our proposed method with commonly used loss functions, as shown in Table 3, and found that CADM outperformed other loss functions, specifically CELoss by 5.09% at Rank-1 and TripletLoss by 4.19% on the RegDB dataset. In addition, Rank-1 score over 1.02% of DM. On the SYSU-MM01 dataset, the Rank-1 score of the proposed method is 0.92% higher than that of DM. Experimental results show that CADM can effectively handle abnormal samples.

4.3 Parameter analysis

Furthermore, we examined different b's effects on DM on the RegDB dataset. As shown in Fig 4, b = 0 is equivalent to using only the Euclidean metric, and the performance of DM gradually improves as b increases. When b = 1, DM performs optimally; however, as b increases, DM's performance decreases. This indicates that the two metrics have an equal impact on model performance, demonstrating their complementarity.

As shown in Table 4, on the RegDB and SYSU-MM01 datasets, we tested the effect of different weights c on the loss of the auto-guided function. When c is small, experimental results show that CADM performance is poor, even worse than DM performance. The weight coefficients are so small that the model parameters do not converge sufficiently. The CADM's performance is optimal when the weight coefficient c is set to 1. In this case, CADM outperforms DM by 1.02%/2.74% on dataset RegDB on Rank-1/mAP and 0.92%/0.74% on dataset SYSU-MM01. When c exceeds 1, the CADM's performance decreases rather than increases.

Table 3. Comparison of our proposed loss with other common loss functions.

Loss	RegDB		SYSU-MM01	
	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)
CenterLoss	83.56	77.96	68.01	65.98
CELoss	84.13	78.34	68.94	66.13
TripletLoss	85.03	79.14	69.88	66.89
DM(Ours)	88.20	80.83	70.29	66.94
CADM(Ours)	**89.22**	**83.57**	**71.21**	**67.68**

https://doi.org/10.1371/journal.pone.0293498.t003

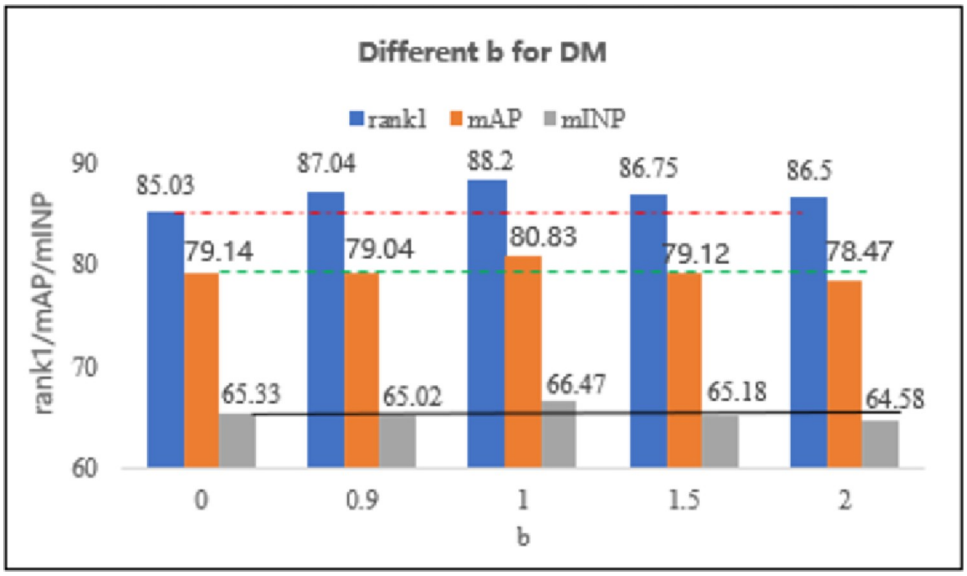

Fig 4. The figure shows the effect of variation in b on the DM performance on RegDB. Rank-1 is represented by the long blue bar, mAP by the long orange bar, and mINP by the long gray bar. The red "—" line represents the rank1 baseline, the green "—" line represents the mAP baseline, and the black "-" line represents the mINP baseline.

https://doi.org/10.1371/journal.pone.0293498.g004

One possibility is that it only amplifies the gradient when it is very large while the model parameters have already been optimized to their maximum.

4.4 Visualization analysis

To demonstrate the effectiveness of our proposed method more intuitively, we use heat maps to display the features learned from pedestrian images. The heat maps of pedestrian images in different modalities are shown in Fig 5(A) and 5(B), respectively. The heat map obtained from the CSMN below focuses more on identity-related information than the heat map obtained from the baseline (CAJL [37]) network above, as shown in the figure. This suggests that the CSMN is not particularly sensitive to some distressing information (light, occlusion, etc.). As a result, the model has a high degree of generalizability.

4.5 Comparison to the state-of-the-art methods

We compared CSMN with the existing state-of-the-art VT-ReID methods on two benchmark datasets. Tables 5 and 6 show the detailed results for different evaluation metrics.

Table 4. The effect of different values of c on the performance of CADM on the RegDB and SYSU-MM01 datasets.

Methods	c	RegDB			SYSU-MM01		
		Rank-1	mAP	mINP	Rank-1	mAP	mINP
Base	-	85.03	79.14	65.33	69.88	66.89	53.61
DM	0	88.20	80.83	66.02	70.29	66.94	54.07
CADM	0.5	86.95	78.89	64.56	68.32	64.96	51.83
CADM	0.9	88.38	80.87	67.40	70.34	66.92	54.02
CADM	**1**	**89.22**	**83.57**	**65.93**	**71.21**	**67.68**	**54.12**
CADM	1.25	88.59	79.77	66.16	70.65	67.01	53.97
CADM	1.5	86.26	79.34	65.84	68.47	65.70	51.93

https://doi.org/10.1371/journal.pone.0293498.t004

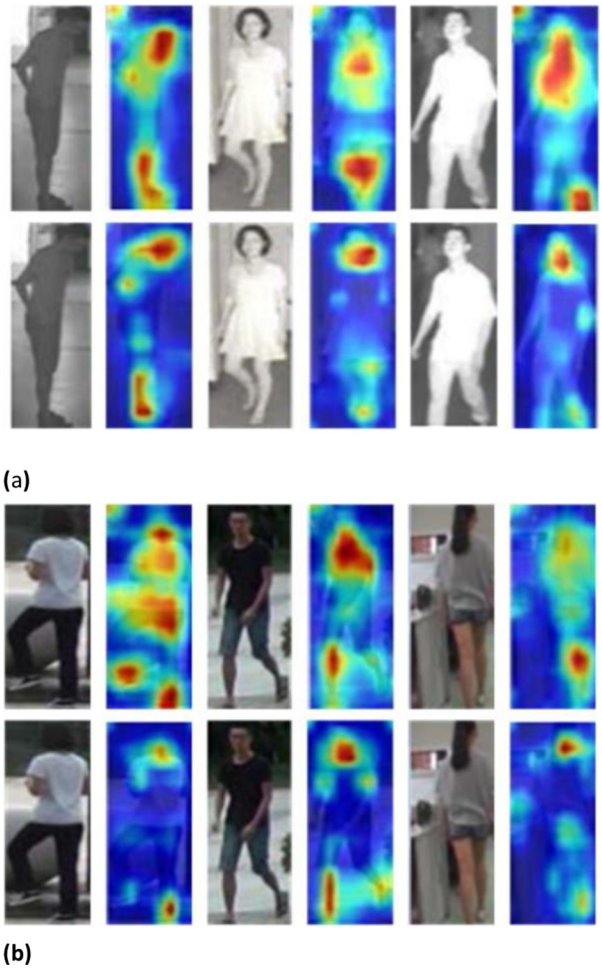

Fig 5. Heat maps extracted by the baseline network (CAJL) and CSMN are displayed on top and bottom, respectively. Note that the pedestrian images is similar but not identical to the original image and is therefore for illustrative purposes only. (a) Comparison of heat maps extracted by the DCMN and the baseline network (CAJL) in infrared modality. (b) Comparison of heat maps extracted by the CSMN and the baseline network (CAJL) in visible modality.

https://doi.org/10.1371/journal.pone.0293498.g005

As shown in Table 5, on the RegDB dataset. SCFNet [45] also designs loss functions to reduce the impact of outlier samples on the spatial features and achieves excellent retrieval accuracy. However, our method achieves better results. Specifically, the proposed method outperforms SCFNet by 3.43% in Rank-1 and 1.66% in mAP scores. For the SYSU-MM01 dataset, as shown in Table 6, CSMN achieves state-of-the-art performance. CSMN outperforms the CAJL [37] by 1.33% in Rank-1 and 0.51% in mINP in the more complicated full search mode. To alleviate the strict constraints of traditional triplet loss, the HCTri [39] method, which also improves the loss function, proposes hetero-center triplet loss. Our proposed CSMN, on the other hand, outperforms HCTri on Rank-1 and mAP by 9.53% and 10.17%, respectively. These findings imply that CSMN can effectively reduce the differences between modalities and channels within a modality. On the other hand, CADM can learn the sample distribution in a more fine-grained manner and deal with outlier samples.

Table 5. Comparison to the state-of-the-art methods on the RegDB dataset.

Method	Venue	Visible to Infrared			Infrared to Visible		
		Rank-1	mAP	mINP	Rank-1	mAP	mINP
Zero-Pad [8]	ICCV-17	14.80	15.95	-	16.63	17.82	-
HCML [16]	AAAI-18	24.44	20.08	-	21.70	22.24	-
HSME [17]	AAAI-19	50.85	47.00	-	50.15	46.16	-
D^2RL [38]	CVPR-19	43.40	44.10	-	-	-	-
AlignGAN [13]	ICCV-19	57.90	53.60	-	56.30	53.40	-
Hi-CMD [39]	CVPR-20	70.93	66.04	-	-	-	-
AGW [36]	arXiv-20	70.05	66.37	50.19	70.49	65.90	51.24
DDAG [28]	ECCV-20	69.34	63.46	49.24	68.06	61.80	48.62
HAT [40]	TIFS-20	71.83	67.56	-	70.02	66.30	-
GMRN [41]	ICIP-21	78.25	71.00	-	-	-	-
MCLNet [42]	ICCV-21	80.31	73.07	57.39	75.93	69.49	52.63
CAJL [37]	ICCV-21	85.03	79.14	65.33	84.75	77.82	61.56
SCFNet [43]	CVPR-22	85.79	81.91	-	86.33	82.10	-
DML [44]	TCSVT-22	77.60	84.30	-	77.00	83.60	-
DSCNet [45]	TIFS-23	85.39	77.30	-	83.50	75.19	-
CSMN	**Ours**	**89.22**	**83.57**	**65.93**	**86.89**	**82.34**	**63.12**

https://doi.org/10.1371/journal.pone.0293498.t005

5 Conclusion

This paper proposes a CSMN framework for visible-thermal person re-identification, which considers cross-modality differences as differences between channels. We reduce the differences between channels in two aspects: On the one hand, we propose ICSM, which learns fine-grained features among channels within a modality to maximize the consistency between channels and minimize the differences between them. On the other hand, we propose CCSM,

Table 6. Comparison to the state-of-the-art methods on the SYSU-MM01 dataset.

Method	Venue	All Search			Indoor Search		
		Rank-1	mAP	mINP	Rank-1	mAP	mINP
Zero-Pad [8]	ICCV-17	14.80	15.95	-	20.58	26.92	-
HCML [16]	AAAI-18	14.32	16.16	-	24.52	30.08	-
cmGAN [27]	IJCAI18	26.97	27.80	-	31.63	42.19	-
HSME [17]	AAAI-19	20.68	23.12	-	-	-	-
D^2RL [38]	CVPR-19	28.90	29.20	-	-	-	-
AlignGAN [13]	ICCV-19	42.40	40.70	-	45.90	54.30	-
AGW [36]	arXiv-20	47.50	47.65	35.30	54.17	62.97	59.23
DDAG [28]	ECCV-20	54.75	53.02	39.62	61.02	67.98	62.61
HAT [40]	TIFS-20	55.29	53.89	-	62.10	69.37	-
HCTri [46]	TMM20	61.68	57.51	39.54	63.41	68.17	64.26
GMRN [41]	ICIP-21	57.67	54.88	-	-	-	-
MCLNet [42]	ICCV-21	65.40	61.98	47.39	72.56	76.58	72.10
CAJL [37]	ICCV-21	69.88	66.89	53.61	76.26	80.37	76.79
DML [44]	TCSVT-22	62.20	49.60	-	66.40	60.00	-
DLRL [47]	TIP-22	63.04	60.58	-	67.95	52.12	-
FMCNet [48]	TIP-22	66.34	62.51	-	68.15	74.09	-
CSMN	**Ours**	**71.21**	**67.68**	**54.12**	**77.32**	**81.56**	**77.12**

https://doi.org/10.1371/journal.pone.0293498.t006

which learns global channel features between modalities to aggregate samples of the same identity across modalities. In addition, to better optimize the sample distribution between and within modalities, we propose CADM. Unlike methods that learn sample distribution at the instance level, our method fully exploits the advantages of channel consistency to learn the sample distribution in a more fine-grained manner. Moreover, we use an auto-guided function to reduce the generation of outlier samples. Our experiments on two benchmark datasets indicate that CSMN outperforms the existing state-of-the-art methods for VI-ReID.

Acknowledgments

The authors are very thankful to the editor and the referees for their valuable comments for improving the paper.

Author Contributions

Conceptualization: Wenzhong Yang.

Funding acquisition: Wenzhong Yang.

Methodology: Yingjie Zhu, Wenzhong Yang.

Software: Yingjie Zhu, Wenzhong Yang.

Supervision: Wenzhong Yang.

Validation: Yingjie Zhu.

Visualization: Yingjie Zhu.

Writing – original draft: Yingjie Zhu.

Writing – review & editing: Yingjie Zhu.

References

1. Zheng L., Yang Y., Hauptmann A. G., "Person re-identification: Past, present and future," 2016, arXiv:1610.02984.
2. Koestinger M, Hirzer M, Wohlhart P, Roth P M, Bischof H. Large scale metric learning from equivalence constraints[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 2288–2295.
3. Liao S, Hu Y, Zhu X, Li S Z. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2197–2206.
4. Das A., Chakraborty A., Roy-Chowdhury A. K., Consistent re-identification in a camera network, in Proceedings of the European Conference on Computer Vision, 2014, pp. 330–345.
5. Hirzer M., Roth P. M., Köstinger M., Bischof H. Relaxed pairwise learned metric for person re-identification, in Proceedings of the European Conference on Computer Vision, 2012, pp. 780–793.
6. Zheng L, Shen L, Tian L, Wang J, Tian Q. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1116–1124.
7. Lv J, Chen W, Li Q, Yang C. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7948–7956.
8. Wu A, Zheng W S, Yu H X, Gong S, Lai J. RGB-infrared cross-modality person re-identification[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5380–5389.
9. Feng Z, Lai J, Xie X. Learning modality-specific representations for visible-infrared person re-identification[J]. IEEE Transactions on Image Processing, 2019, 29: 579–590.
10. Fu C, Hu Y, Wu X, Shi H, Mei T, He R. CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11823–11832.

11. Kniaz V V, Knyaz V A, Hladuvka J, Kropatsch W G, Mizginov V. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018: 0–0.

12. Zhang Z, Jiang S, Huang C, Li Y, Da Xu R Y. RGB-IR cross-modality person ReID based on teacher-student GAN model[J]. Pattern Recognition Letters, 2021, 150: 155–161.

13. Wang G, Zhang T, Cheng J, Liu S, Yang Y, Hou Z. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3623–3632.

14. Hao Y, Li J, Wang N, Gao X. Modality adversarial neural network for visible-thermal person re-identification[J]. Pattern Recognition, 2020, 107: 107533.

15. Ling Y, Zhong Z, Luo Z, Rota P, Li S, Sebe N. Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 889–897.

16. Ye M., Lan X., J Li, Yuen P. Hierarchical discriminative learning for visible thermal person re-identification, in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

17. Hao Y., Wang N., Li J., and Gao X., "HSME: Hypersphere Manifold Embedding for Visible Thermal Person Re-Identification", AAAI, vol. 33, no. 01, pp. 8385–8392, Jul. 2019. https://doi.org/10.1609/aaai.v33i01.33018385

18. Ye M., Shen J. and Shao L., "Visible-Infrared Person Re-Identification via Homogeneous Augmented Tri-Modal Learning," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 728–739, 2021, https://doi.org/10.1109/TIFS.2020.3001665

19. Zhang Y, Liu S, Qi L, Coleman S, Kerr D, Shi W. Multi-level and multi-scale horizontal pooling network for person re-identification[J]. Multimedia Tools and Applications, 2020, 79: 28603–28619.

20. Lian Sicheng, Jiang Weitao, and Hu Haifeng. "Attention-aligned network for person re-identification." IEEE Transactions on Circuits and Systems for Video Technology 31. 8 (2020): 3140–3153.

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

22. Chen G, Lin C, Ren L, Lu J, Zhou J. Self-critical attention learning for person re-identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9637–9646.

23. Chen T, Ding S, Xie J, Yuan Y, Chen W, Yang Y, et al. Abd-net: Attentive but diverse person re-identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8351–8361.

24. Xia B N, Gong Y, Zhang Y, Poellabauer C. Second-order non-local attention networks for person re-identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 3760–3769.

25. Chen B., Deng W., and Hu J., "Mixed high-order attention network for person re-identification," in Proc. ICCV, Oct. 2019, pp. 371–381.

26. Martinel N, Foresti G L, Micheloni C. Deep pyramidal pooling with attention for person re-identification [J]. IEEE Transactions on Image Processing, 2020, 29: 7306–7316.

27. Dai P, Ji R, Wang H, Wu Q, Huang Y. Cross-modality person re-identification with generative adversarial training[C]//IJCAI. 2018, 1(3): 6.

28. Ye M, Shen J, J. Crandall D, Shao L, Luo J. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer International Publishing, 2020: 229–247.

29. Ye M, Wang Z, Lan X, Yuen P C. Visible thermal person re-identification via dual-constrained top-ranking[C]//IJCAI. 2018, 1: 2.

30. Ye M, Lan X, Wang Z, Yuen P C. Bi-directional center-constrained top-ranking for visible thermal person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 407–419.

31. Hao Y, Wang N, Li J, Gao X. HSME: Hypersphere manifold embedding for visible thermal person re-identification[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 8385–8392.

32. Zhao Y B, Lin J W, Xuan Q, Xi X. Hpiln: a feature learning framework for cross-modality person re-identification[J]. IET Image Processing, 2019, 13(14): 2897–2904.

33. Wu A, Zheng W S, Gong S, Lai J. RGB-IR person re-identification by cross-modality similarity preservation[J]. International journal of computer vision, 2020, 128: 1765–1785.

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

35. Luo H, Jiang W, Gu Y, Liu F, Liao X, Lai S, et al. A strong baseline and batch normalization neck for deep person re-identification[J]. IEEE Transactions on Multimedia, 2019, 22(10): 2597–2609.

36. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi S C. Deep learning for person re-identification: A survey and outlook[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(6): 2872–2893.

37. Ye M, Ruan W, Du B, Shou M Z. Channel augmented joint learning for visible-infrared recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13567–13576.

38. Wang Z, Wang Z, Zheng Y, Chuang Y Y, Satoh S I. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 618–626.

39. Choi S, Lee S, Kim Y, Kim T, Kim C. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10257–10266.

40. Ye M, Shen J, Shao L. Visible-infrared person re-identification via homogeneous augmented tri-modal learning[J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 728–739.

41. Sun J, Zhang T. RGB-Infrared Person Re-Identification Via Multi-Modality Relation Aggregation and Graph Convolution Network[C]//2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021: 1174–1178.

42. Hao X, Zhao S, Ye M, Shen J. Cross-modality person re-identification via modality confusion and center aggregation[C]//Proceedings of the IEEE/CVF International conference on computer vision. 2021: 16403–16412.

43. Su P, Liu R, Dong J, Yi P, Zhou D. SCFNet: A Spatial-Channel Features Network based on Hetero-centric Sample Loss for Visible-Infrared Person Re-Identification[C]//Proceedings of the Asian Conference on Computer Vision. 2022: 3552–3568.

44. Zhang D, Zhang Z, Ju Y, Wang C, Xie Y, Qu Y. Dual mutual learning for cross-modality person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5361–5373.

45. Zhang Y, Kang Y, Zhao S, Shen J. Dual-Semantic Consistency Learning for Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Information Forensics and Security, 2022.

46. Liu H, Tan X, Zhou X. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification[J]. IEEE Transactions on Multimedia, 2020, 23: 4414–4425.

47. Wu Y, He G D, Wen L H, Qin X, Yuan C A, Gribova V, et al. Discriminative local representation learning for cross-modality visible-thermal person re-identification[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2022, 5(1): 1–14.

48. Zhang Q, Lai C, Liu J, Huang N, Han J. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7349–7358.