Machine Vision and Applications (2023) 34:105
https://doi.org/10.1007/s00138-023-01458-3

ORIGINAL PAPER

Beyond a strong baseline: cross-modality contrastive learning for visible-infrared person re-identification

Pengfei Fang[1,2] · Yukang Zhang[3] · Zhenzhong Lan[4]

Received: 15 March 2023 / Revised: 28 July 2023 / Accepted: 22 August 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
Cross-modality pedestrian image matching, which entails the matching of visible and infrared images, is a vital area in person re-identification (reID) due to its potential to facilitate person retrieval across a spectrum of lighting conditions. Despite its importance, this task presents considerable challenges stemming from two significant areas: cross-modality discrepancies due to the different imaging principles of spectrum cameras and within-class variations caused by the diverse viewpoints of large-scale distributed surveillance cameras. Unfortunately, the existing literature provides limited insights into effectively mitigating these issues, signifying a crucial research gap. In response to this, the present paper makes two primary contributions. First, we conduct a comprehensive study of training methodologies and subsequently present a strong baseline network designed specifically to address the complexities of the visible-infrared person reID task. This strong baseline network is paramount to the advancement of the field and to ensure the fair evaluation of algorithmic effectiveness. Second, we propose the Cross-Modality Contrastive Learning (CMCL) scheme, a novel approach to address the cross-modality discrepancies and enhance the quality of image embeddings across both modalities. CMCL incorporates intra-modality and inter-modality contrastive loss components, designed to improve the matching quality across the modalities. Thorough experiments show the superior performance of the baseline network, and the proposed CMCL can further bring performance over the baselines, outperforming the state-of-the-art methods considerably.

Keywords Cross-modality · Person re-identification · Strong baseline · Cross-modality contrastive learning

1 Introduction

This article studies the person re-identification (reID) problem, particularly focusing on a challenging setting: matching pedestrian images cross daytime and nighttime.

A person reID is a classical application in metric learning, which aims to learn a latent embedding space via training data. In such a latent space, the unseen query can be correctly matched by simply comparing the distance/similarity to images in the support set. In recent years, the person reID task has made a significant improvement, and the latest person reID machines can achieve human-level retrieval performance [1, 2]. However, such a reID machine cannot work successfully in complicated real-life situations. For example, it is difficult for a system trained by a single modality of visible images, to extract discriminative features for both visible and infrared images, leading to a mismatch of cross-modality data.

This newly emerged problem is named visible-infrared person reID (VI-reID) [3]. The main challenging comes from

✉ Pengfei Fang
fangpengfei@seu.edu.cn

Yukang Zhang
zhangyk@stu.xmu.edu.cn

Zhenzhong Lan
lanzhenzhong@westlake.edu.cn

[1] School of Computer Science and Engineering, Southeast University, Nanjing 211189, Jiangsu, China

[2] MOE Key Laboratory of Computer Network and Information Integration (Southeast University), Nanjing 211189, Jiangsu, China

[3] School of Informatics, Xiamen University, Xiamen 361005, Fujian, China

[4] School of Engineering, Westlake University, Hangzhou 310030, Zhejiang, China

Published online: 15 September 2023

the large modality discrepancy between visible images and infrared images. A diverse set of solutions, mainly including generative-based methods and representation-based methods, have been investigated to mitigate the modality gap of two types of images, resulting in learning a common embedding space for cross-modality pedestrian images [4–7]. Yet, it remains an open problem and requires more effort to improve the retrieval performance in practice.

Meanwhile, along with novel algorithms pushing the boundaries of state-of-the-art values, an effective baseline is also a necessary integral part of a reID system for reasons that a common strong baseline can evaluate the superiority of algorithms and establish the apple to apple comparison of algorithms. As illustrated in Fig. 1, we can observe that the performance of baselines in the state-of-the-art methods varies; thus comparing with algorithms developed on top of different baselines is unfair. That said, in some cases, the improvement is mainly attributed to training tricks, instead of the algorithm itself. This requires us to develop a strong baseline for academia. In this paper, we study a set of training methods in the literature and present a strong baseline for the VI-reID task. We believe our baseline will promote the development of the VI-reID community.

In the proposed baseline network, we follow the common practice of the VI-reID task to use cross-entropy loss and triplet loss as object functions [4, 8]. The cross-entropy loss only optimizes the representation of images and ignores the modality gap [9]. In contrast, the triplet loss usually leverages the hard mining strategy to mine a hard positive sample and a hard negative sample for a given anchor sample. It means that in the triplet loss, a positive pair only contrasts with one negative pair. However, recent studies in contrastive learning reveal contrasting to more negative pairs is helpful for representation learning [10, 11]. This inspires us to develop a cross-modality contrastive learning (CMCL) scheme for the task at hand. In the setup of contrastive learning for self-supervised learning (SSL), a core component is to create a positive pair by applying different data augmentations to the same image. Different from the setup in SSL, we leverage the label information to mine several positive pairs[1] without applying various data augmentations to the same image, which halves the computation cost. Also, in the proposed CMCL, a positive sample w.r.t. an anchor can either be a visible image or an infrared image, which simultaneously decreases the inter-modality variance and intra-variance in the embedding space. Empirically, we find that employing the proposed CMCL is not without difficulties in the proposed VI-reID baseline and we conduct experiments to explore the correct setup to make the contrastive loss work in our task. The **contributions** of this paper can be summarized as follows:

- We first study a set of training methods and verify the effectiveness of each training method on top of a vanilla baseline network. Having the training methods at hand, a strong baseline network can be presented. Our baseline significantly outperforms existing baselines. For example, our baseline performance on the RegDB [12] dataset is 90.57% / 83.05%, which outperforms the state-of-the-art method, i.e., CM-NAS [13], by 6.03% / 4.60% w.r.t. R-1 / mAP.
- We further propose a cross-modality contrastive learning (CMCL) scheme, in which intra-modality contrastive loss and inter-modality contrastive loss are developed to explicitly align the embeddings of two modalities. The correct setup to make the contrastive learning scheme work for our task is also studied.
- A thorough battery of experiments performed on two public datasets, i.e., SYSU-MM01 [3] and RegDB [12], verify the superiority of our baseline and the proposed cross-modality contrastive learning scheme.

Code is available for academic use.[2]

2 Related work

2.1 Person re-identification

An increasing number of solutions have been studied to report steady benchmark improvements over time in person reID and this task aims to create discriminative feature embeddings for pedestrian images [1]. In the era of deep learning, Convolutional Neural Networks (CNNs) have become the popular tool to establish such embedding spaces via extracting the feature of images [14–16]. Along with global features, local features [8, 17], or low-level features [18, 19] are employed to increase the discriminative power of the embeddings. Other auxiliary information, i.e., human poses [20, 21], human attributes [22, 23], and visual attention [24, 25], also provide cues to distinguish person appearance features.

Apart from the single modality retrieval task, a more difficult setting, where pedestrians are matched in different modalities of images, is first proposed in [3]. In this setting, a solution is supposed to align the feature embeddings for inter-modality images and intra-modality images of the same person jointly. To achieve so, two popular groups of solutions are studied.

One group is to learn an embedding space for visible and infrared images. The work in [3] augments the visible or infrared images via zero-padding, which enables the network to easily learn domain-specific features. Its extension

[1] A mini-batch contains several samples per person identity.

[2] https://github.com/PengfeiFANGanu/Cross-Model-Re-ID-PCB-baseline.

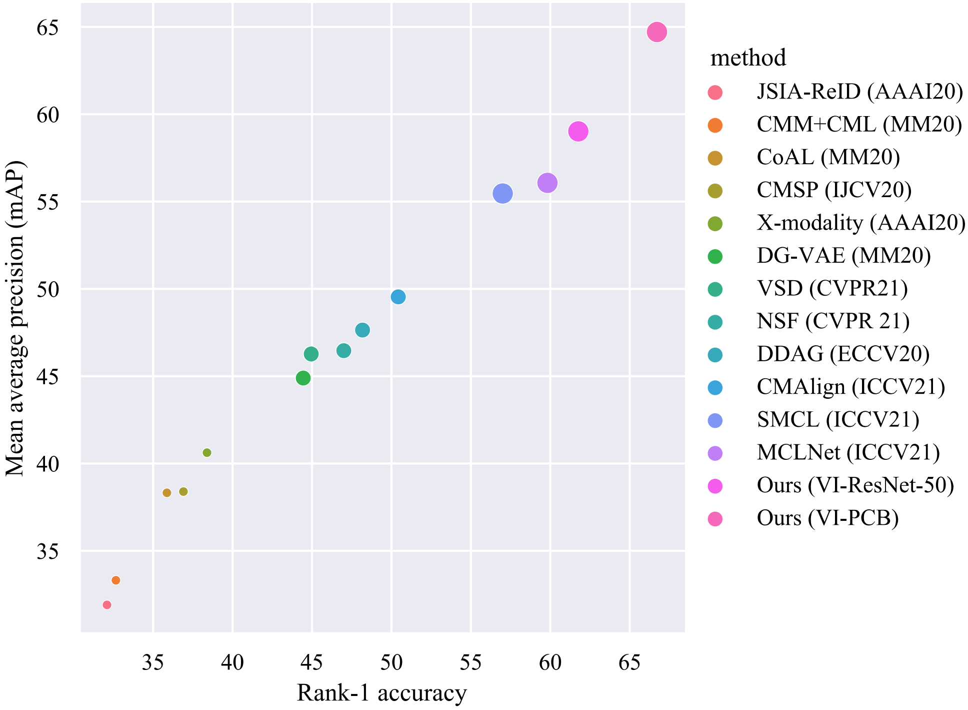


Fig. 1 The comparison of different baselines on the SYSU-MM01 dataset. Our baseline (i.e., VI-PCB) can considerably boost the performance over the baselines in existing state-of-the-art methods

work [26] further proposes to learn domain-shared features by constraining a cross-modality similarity preservation loss to the network. The following works improve the discrimination of features by extracting modality-shared features [27] and modality-invariant features [28]. Considering the domain gap, the works in [29, 30] enable the network to focus on the informative regions in person images across two modalities. The X-modality strategy, which bridges the images of two modalities, can alleviate the domain gap by a learned middle modality [31]. This idea is also elaborated in [32], where a new syncretic modality is adaptatively generated to minimize the modality gap by aligning the distributions of two modalities. The domain gap can also be eliminated either by leveraging the attention mechanism to match the visual-similar features between different modalities [4, 33], or by searching an optimal neural architecture for both modalities of data [13].

Another group of solutions leverages the superiority of generative models, which reduce the modality discrepancy at the pixel level of images. In [34], the AlignGAN translates visible images to infrared images and aligns the images at pixel level and feature level jointly. A similar idea is proposed in [35], where a generative model generates a unified latent space for images and therefore a feature extractor is trained to encode images in a common embedding space. The person image is also paired in both light conditions using a generative model, hence the network can focus on learning appearance features of images [36]. In contrast to [36], the work in [37] disentangles person images and only keeps the person-discriminative factors for the retrieval task.

Alongside the diverse set of seminal algorithms for the VI-reID task, an effective baseline network is also a necessary integral part of the community. In the single modality person reID task, Luo et al investigated a number of training methods in the baseline network, and found a useful training method, i.e., batch normalization neck, to jointly optimize the triplet loss and classification loss [2, 38]. Our work follows the works [2, 38] to develop our strong baseline for the VI-reID task. The closest study to our work is the work in [39], where the baseline network is developed via modifying the triplet loss, whereas our work systematically studies the training methods in the field and presents a better baseline.

2.2 Contrastive learning

Minimizing the objective function via contrastive scheme has shown its power in self-supervised learning (SSL) [10, 11, 40–42]. The contrastive scheme in SSL can be derived from the infoNEC loss [43], which is to maximize the mutual information (MI) between global and local representations of images. Similarly, contrastive learning aims to learn consistent image representations under different data augmentations and the work in [44] proves that reducing the MI between views of an image leads to an increasing in the representation power. Along with the data augmentation techniques, Chen et al first proposed an effective framework for SSL via contrastive scheme [10]. To alleviate the large batch size in [10], MoCo introduces a momentum-updated encoder, which utilizes the data in previous batches. The work in [40] further extends the SSL framework to

Siamese networks. Under the SSL framework, more effective algorithms are further proposed to push the boundaries of state-of-the-art performance [41, 42, 44–46].

In the cross-modality analysis, contrastive loss also becomes an option to match cross-modal data [47–50]. For example, the works in [47–49] establish correspondences between the text and image [48]. In [50], abnormality analysis is achieved by contrasting image features and radiomic features. Different from existing works, our proposal exploits the full potential of cross-modality contrastive learning, which takes into account the intra-modality similarities and inter-modality similarities in the loss, and is the very first time to address the VI-reID task.

3 A strong baseline

3.1 Problem formulation

Let a third-order tensor, $\boldsymbol{V}_i \in \mathbb{R}^{C\times H\times W}$ or $\boldsymbol{I}_i \in \mathbb{R}^{C\times H\times W}$, denote the i-th visible image or infrared image of a person, in which C, H and W are the number of channels, height and width, respectively. Each image is associated to a person identity y_i. The goal of VI-reID learning is to regress a non-linear function $F_\theta : \mathcal{X} \rightarrow \mathbb{R}^n$ to embed visible images or infrared images to a possible common feature space. To be specific, the training set consists of both visible image set and infrared image set, represented as $\mathcal{X}^{\text{trn}} = [\mathcal{V}^{\text{trn}} \mid \mathcal{I}^{\text{trn}}]$, with each set given by $\mathcal{V}^{\text{trn}} = \{\boldsymbol{V}_i, y_i\}_{i=1}^{N_V}$ and $\mathcal{I}^{\text{trn}} = \{\boldsymbol{I}_i, y_i\}_{i=1}^{N_I}$. The VI-reID task formulates the training as:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{\boldsymbol{V}\in\mathcal{V}^{\text{trn}},\boldsymbol{I}\in\mathcal{I}^{\text{trn}}} \mathcal{L}\big(F_\theta(\boldsymbol{V}), F_\theta(\boldsymbol{I})\big), \tag{1}$$

where $\mathcal{L}$ indicates the training objective. In the remainder of this section, we first describe the vanilla baseline network and then systematically study the training methods on the baseline.

3.2 A Vanilla baseline network

Figure 2a illustrates the architecture of the vanilla baseline network. We employ a ResNet-50 [51], pretrained on ImageNet [52], as the backbone network. Training a VI-reID machine includes the following steps.

Batch sampling A mini-batch contains M visible images and M infrared images, denoted by $\mathcal{B} = [\ \{\boldsymbol{V}_k\}_{k=1}^{M} \ \mid\ \{\boldsymbol{I}_k\}_{k=1}^{M}\ ]$. In each modality, we randomly sample N_P person identities and N_K samples per identity, satisfying that $M = N_P \times N_K$. Hence the size of a mini-batch is $2 \times M$. In this paper, we set $N_P = 8$ and $N_K = 4$.

Data augmentations In this vanilla baseline network, various data augmentation techniques are used, such as `zero-padding`, `random crop`, `random horizontal flip` and `normalize`.

Loss functions The network is optimized via multi-task learning (MTL) scheme. As the name suggests, MTL formulates the overall learning procedure as a combination of several sub-tasks; each having its own importance in the overall learning mechanism. In this work, we train our network for the tasks of ranking and classification. Specifically, a network first encodes images to a batch of features, i.e., $\boldsymbol{B} = [\ \{\boldsymbol{v}_k\}_{k=1}^{M} \ \mid\ \{\boldsymbol{i}_k\}_{k=1}^{M}\ ]$. For simplicity, we denote the batch of features as $\boldsymbol{B} = [\ \{\boldsymbol{b}_i\}_{i=1}^{2M}\ ]$. For all possible triplet $[\boldsymbol{b}_i, \boldsymbol{b}^+, \boldsymbol{b}^-]$, with $\boldsymbol{b}_i, \boldsymbol{b}^+$ and $\boldsymbol{b}^-$ presenting an anchor sample, a positive sample and a negative sample, the cross-modality triplet loss function is given by:

$$\mathcal{L}_{\text{tri}} = \frac{1}{2M}\sum_{i=1}^{2M}\big[d(\boldsymbol{b}_i, \boldsymbol{b}^+) - d(\boldsymbol{b}_i, \boldsymbol{b}^-) + \xi\big]_+, \tag{2}$$

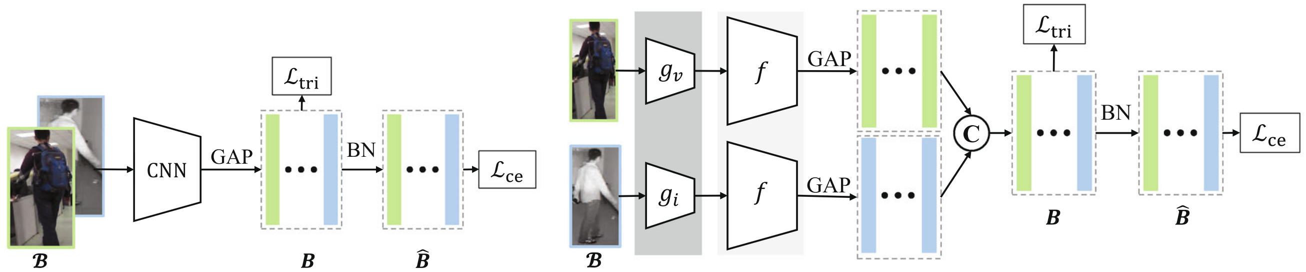
where ξ is the margin and we set $\xi = 0.3$ throughout the paper. The triplet loss aims to learn an embedding space that increases the between-class variance by optimizing the relative similarity information in a triplet. Following the good practice in [38], we use a `batch normalization neck` (BNN) on $\boldsymbol{B}$ to obtain new features $\hat{\boldsymbol{B}} = [\{\hat{\boldsymbol{b}}_i\}_{i=1}^{2M}] = \text{BN}(\boldsymbol{B})$, where BN is a batch normalization layer, defined as $\text{BN}(\cdot) : \mathbb{R}^m \rightarrow \mathbb{R}^m, \text{BN}(\boldsymbol{x}) := \gamma\frac{\boldsymbol{x}-\text{E}[\boldsymbol{x}]}{\sqrt{\text{Var}[\boldsymbol{x}]}} + \beta$. Then a fully connected layer $\boldsymbol{W}$ is used to produce logits, as $\boldsymbol{p} = \text{softmax}(\boldsymbol{W}^\top\hat{\boldsymbol{b}}_i)$. The cross-entropy loss is to maximize the log likelihood of $\hat{\boldsymbol{b}}_i$ w.r.t. its identity c, as follows:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{2M}\sum_{i=1}^{2M}\log\big(p(y_i = c \mid \hat{\boldsymbol{b}}_i)\big). \tag{3}$$

This classification task encodes the class-specific information, which minimizes the within-class variance.

Optimization SGD optimizer is adopted to optimize the network. The learning rate is initialized to 0.01. The learning rate is decayed by a factor of 0.1 at the 20-th, 50-th epoch respectively for all the datasets. The network is trained for 80 epochs in total. The values of weight decay and momentum are set to 0.0005 and 0.9, respectively. Given a batch of data, one first need to calculate the gradient of the loss function $\mathcal{L}$ w.r.t θ^i in the i-th iteration. Then the update role of SGD is given by $\theta^{i+1} = \theta^i - \eta\Delta_{\theta^i}\mathcal{L}$.

Inference In the inference stage, the features $\hat{\boldsymbol{B}}$ is employed to evaluate the model's performance.


$\mathcal{L}_{\text{tri}}$
CNN
GAP
BN
$\mathcal{L}_{\text{ce}}$
$\mathcal{B}$
$\boldsymbol{B}$
$\widehat{\boldsymbol{B}}$
g_v
g_i
f
f
GAP
GAP
C
$\mathcal{L}_{\text{tri}}$
BN
$\mathcal{L}_{\text{ce}}$
$\mathcal{B}$
$\boldsymbol{B}$
$\widehat{\boldsymbol{B}}$

(a) The architecture of the vanilla baseline network.

(b) The architecture of the strong baseline network.

Fig. 2 The network architectures of the vanilla baseline and strong baseline

3.3 Training methods

On top of the vanilla baseline, we further introduce some training methods, which are the essential components of our strong baseline network (see Fig. 2b). Those training methods have appeared in published papers and open-source codes. We aim to systematically evaluate the effectiveness of such methods in VI-reID task. Noted that most of the training methods can be seamlessly employed in the vanilla baseline.

Modality Learning To learn modality-specific features of visible and infrared images, we follow the work in [4] to incorporate the `modality learning` module in the network. This is achieved by using two convolutional blocks with different parameters to capture the modalities features (see g_v and g_i in Fig. 2b). Then another convolutional block with share parameters is used to encode a common feature space for two modalities (see f in Fig. 2b). In our implementation, we set g_i = conv1, g_v = conv1, and f = conv2_x − conv5_x in the ResNet-50.

Warmup learning rate The learning rate is an important hyper-parameter, which decides the step size of updating parameters. We employ the `warmup` strategy to slowly increase the value of the learning rate at the beginning of the training stage. This can help the network to adapt to the training data and avoids over-fitting of the network in the early stages of training. In practice, we use the linear `warmup` scheme [38] as follows:

$$\mathrm{lr}(epoch) = \begin{cases} 0.1 \times \frac{epoch}{10}, & epoch \leq 10 \\ 0.1, & 10 < epoch \leq 20 \\ 0.01, & 20 < epoch \leq 50 \\ 0.001, & 50 < epoch \leq 80. \end{cases} \tag{4}$$

Random erasing To improve the generalization of a reID machine, we adopt the `random erasing` [53] data augmentation to train the neural network as shown in Fig. 3a and b. In the training stage, the `random erasing` first produces a rectangle region with various sizes. Then this rectangle region randomly attends to images and replaces the pixel values with random values. Such a process of data augmentation can reduce the risk of over-fitting, thereby improving the generalization of the network. In the implementation, we use the default values of hyper-parameters for `random erasing` as in [53].

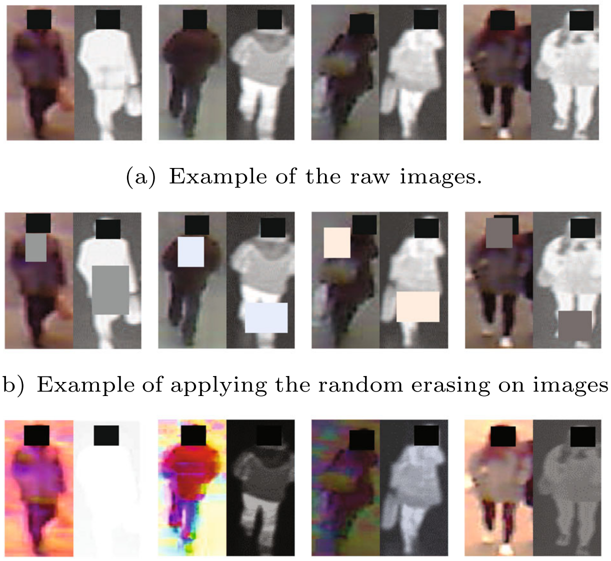
Color Jitter In the VI-reID task, different imaging processes of spectrum cameras [54] lead to the changes of color between visible images and infrared images, whereas the texture information of images is preserved. That said, the color of images significantly affects appearance features encoded by the network. To mitigate this issue, we use the `color jitter` to augment the data, as illustrated in Fig. 3a and c. The `color jitter` can randomly change the brightness, contrast, and saturation of images. In our baseline, all values of hyper-parameters are set to 0.5.

Part feature learning Many studies in the person reID field have shown that part-features of person images can enhance the matching performance of the system [8, 17, 55]. We also equip our network with the ability of part-based learning. We follow the seminal work, e.g., PCB [17], to develop the `part feature learning` model. To this end, the backbone network encodes an image to a 3-D feature map, $\boldsymbol{T} \in \mathbb{R}^{c \times h \times w}$. We horizontal partition the feature map into t parts. Then each part $\boldsymbol{T}_i$ is summarized using global average pooling into a part feature $\boldsymbol{g}_i$. The part feature $\boldsymbol{g}_i$ is optimized by triplet loss and cross-entropy loss. In our baseline network, we set $t = 4$.

Noted the ablation study on the vanilla baseline (i.e., image size or batch size) and strong baseline (i.e., number of part features t) is reported in the supplementary material.

4 Cross-modality contrastive loss

In this section, we propose to enable the use of contrastive loss in the cross-modality matching problem as shown in Fig. 4. To align features of both intra-modality



(a) Example of the raw images.

(b) Example of applying the random erasing on images.

(c) Example of applying the color jitter on images.

Fig. 3 Samples for raw images, and augmented images by `random erasing` and `color jitter` in RegDB dataset. Faces are covered using black boxes for privacy purposes. Best viewed in color

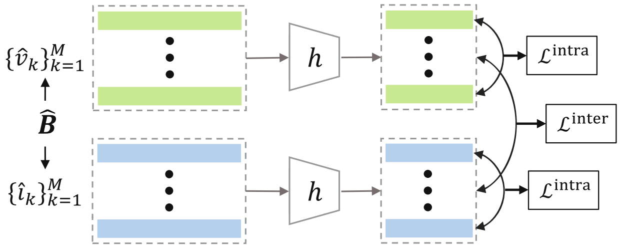

$\{\hat{\boldsymbol{v}}_k\}_{k=1}^{M}$
$\hat{\boldsymbol{B}}$
$\{\hat{\boldsymbol{i}}_k\}_{k=1}^{M}$
h
h
$\mathcal{L}^{\text{intra}}$
$\mathcal{L}^{\text{inter}}$
$\mathcal{L}^{\text{intra}}$

Fig. 4 Intra-modality and inter-modality contrastive loss in the cross-modality contrastive learning scheme

and inter-modality in the embedding space, we propose two modifications of the contrastive loss,[3] i.e., intra-modality contrastive loss and inter-modality contrastive loss.

Consider a batch of vector representations for visible images and infrared images, $\hat{\boldsymbol{B}} = [\ \{\hat{\boldsymbol{v}}_k\}_{k=1}^{M}\ \mid\ \{\hat{\boldsymbol{i}}_k\}_{k=1}^{M}\ ]$, we use a projection head h[4] to project the vector representations to the space where contrastive loss is applied as $\tilde{\boldsymbol{B}} = h(\hat{\boldsymbol{B}})$, $\tilde{\boldsymbol{B}} = [\ \{\tilde{\boldsymbol{v}}_k\}_{k=1}^{M}\ \mid\ \{\tilde{\boldsymbol{i}}_k\}_{k=1}^{M}\ ]$. Let $I = \{1 \ldots M\}$ be the index of samples in a single modality. In the presence of label information and the sampling strategy (e.g., N_k samples per person identity) in our task, each anchor sample contains at least one positive sample in a mini-batch. In other words, positive pairs for any samples can be guaranteed without using data augmentations in the popular contrastive learning pipeline. We describe the intra-modality contrastive loss and inter-modality contrastive loss below.

4.1 Intra-modality contrastive loss

The intra-modality contrastive loss aims to maximize the mutual information between the anchor sample and its positive sample in a single modality, thereby aligning the feature embeddings for the same class. Given a batch of feature vectors for visible images, i.e., $\tilde{\boldsymbol{V}} = \{\tilde{\boldsymbol{v}}_k\}_{k=1}^{M}$, $P(k)$ indicates the index set for positive samples w.r.t. the anchor (i.e., $\tilde{\boldsymbol{v}}_k$) in $\tilde{\boldsymbol{V}}$, and $S(k) = I \backslash \{k\}$. The intra-modality contrastive loss for visible images is formulated by:

$$\mathcal{L}_V^{\text{intra}} = \sum_{k=1}^{M} \frac{-1}{|P(k)|} \sum_{p\in P(k)} \log \frac{\exp\left(\text{sim}(\tilde{\boldsymbol{v}}_k, \tilde{\boldsymbol{v}}_p)/\tau\right)}{\sum_{s\in S(k)} \exp\left(\text{sim}(\tilde{\boldsymbol{v}}_k, \tilde{\boldsymbol{v}}_s)/\tau\right)}, \tag{5}$$

where τ is the temperature and $\text{sim}(\cdot,\cdot) : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$, $\text{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j) := \frac{\boldsymbol{x}_i^\top \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|}$ is cosine similarity. Throughout the paper, we empirically set $\tau = 0.1$. The intra-modality contrastive loss for infrared images (i.e., $\mathcal{L}_I^{\text{intra}}$) can be obtained in a similar way.

4.2 Inter-modality contrastive loss

The inter-modality contrastive loss receives the feature vectors of two modalities as input and aims to project visible images and infrared images into a common embedding space. Having both the visible image embeddings (i.e., $\tilde{\boldsymbol{V}} = \{\tilde{\boldsymbol{v}}_k\}_{k=1}^{M}$) and infrared image embeddings (i.e., $\tilde{\boldsymbol{I}} = \{\tilde{\boldsymbol{i}}_k\}_{k=1}^{M}$) as input, the inter-modality contrastive loss from visible images to infrared images is defined as:

$$\mathcal{L}_{V\rightarrow I}^{\text{inter}} = \sum_{k=1}^{M} \frac{-1}{|P'(k)|} \sum_{p\in P'(k)} \log \frac{\exp\left(\text{sim}(\tilde{\boldsymbol{v}}_k, \tilde{\boldsymbol{i}}_p)/\tau\right)}{\sum_{s\in I} \exp\left(\text{sim}(\tilde{\boldsymbol{v}}_k, \tilde{\boldsymbol{i}}_s)/\tau\right)}, \tag{6}$$

where $P'(k)$ indicates the index set for positive samples w.r.t. the anchor (e.g., $\tilde{\boldsymbol{v}}_k$) in another modality of data (e.g., $\tilde{\boldsymbol{I}}$). Noted that the inter-modality contrastive loss from infrared images to visible images (e.g., $\mathcal{L}_{I\rightarrow V}^{\text{inter}}$) can also be defined. In both intra-modality contrastive loss and inter-modality contrastive loss, one can find that the optimization of the two contrastive loss indeed can create an embedding space that accommodates two modality of data jointly.

Hence, the total cross-modality contrastive loss can be formally formulated as: $\mathcal{L}_{\text{cts}} = \lambda_1 \mathcal{L}_V^{\text{intra}} + \lambda_2 \mathcal{L}_I^{\text{intra}} + \lambda_3 \mathcal{L}_{V\rightarrow I}^{\text{inter}} + \lambda_4 \mathcal{L}_{I\rightarrow V}^{\text{inter}}$. Having the cross-modality contrastive loss at hand, we can give the final training loss:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}} + \beta \mathcal{L}_{\text{cts}}. \tag{7}$$

In this paper, we set λ_1, λ_2, λ_3, λ_4 and β to 1.0.

[3] For simplicity, we formulate the contrastive loss without considering the part model. While, the usage of contrastive loss in the part model is identical.

[4] In our work, the projection head is a 2-layer MLP with ReLU activation (i.e., 2048 → 2048 → ReLU → 128) as the common practice in [10].

Table 1 Evaluation of different training components on SYSU-MM01 and RegDB datasets

Training methods	SYSU-MM01		RegDB	
	R-1	mAP	R-1	mAP
Vanilla baseline	38.44	38.21	41.67	39.77
+ `modality learning`	42.93	42.06	48.41	46.47
+ `warmup`	51.11	50.01	73.44	67.31
+ `random erasing`	61.76	59.02	79.30	72.53
+ `color jitter`	60.35	57.68	84.82	75.60
+ `part feature learning` ($t = 4$)	**66.71**	**64.71**	**90.57**	**83.05**

Bold values indicate the best result

5 Experiment

5.1 Datasets and evaluation protocol

SYSU-MM01 [3] is a large VI-reID dataset, collected by four RGB cameras and two near-infrared cameras. The training set of SYSU-MM01 contains 22, 258 visible images and 11, 909 infrared images of 395 identities in total. In the testing set, the query set has 3, 803 infrared images of 96 identities, and the gallery set has 301 visible images of the same identities as the query set. Specifically, RGB camera 1 and camera 2 are located in two bright indoor rooms. Infrared camera 3 and camera 6 are placed in a dark environment. Cameras 4 and 5 are two RGB surveillance cameras to capture pedestrians in outdoor scenes. In evaluation, we report results on both *All Search* and *Indoor Search* modes. In All Search mode, the gallery set contains all visible images captured by four RGB cameras, while in Indoor Search mode, the gallery set only contains the images captured by two indoor cameras.

RegDB [12] is another popular dataset for VI-reID task. This dataset uses a paired camera system (one RGB camera and one far-infrared camera) to captures 412 person, with each having 10 visible images and 10 infrared images. Two cameras are rigidly attached closely together on a panel such that two modalities of images can be collected without any differences in capturing time. Then, the dual camera system is placed on top of the building, simulating the normal working condition of the surveillance system. That said, the RGB images and the far-infrared images are captured simultaneously. Following the standard protocol, 206 identities are randomly selected as training set and the remaining 206 identities are used as testing set in a trial. All results are reported over 10 trials of random split to the dataset. In evaluation, both *Visible to Infrared* and *Infrared to Visible* modes are adopted to evaluate our model.

Following the common practice of person reID, we evaluate our model using rank-k values in cumulative matching characteristic (CMC) curve and mean average precision (mAP) metric.

We implement our method in the PyTorch [56] deep learning package and all experiments are performed on NVIDIA Tesla V100 GPUs. In Sects. 5.2, and 5.3, we use “All Search” mode for SYSU-MM01 dataset and “Visible to Infrared” mode for RegDB dataset for the study.

5.2 Training methods on baseline networks

We first conduct a thorough battery of experiments to systematically evaluate the effectiveness of each training component as shown in Table 1. The vanilla baseline network achieves 38.44%/38.21% and 41.67%/39.77% for Rank-1/mAP accuracies on t SYSU-MM01 dataset and RegDB dataset respectively, which is on par with most of baselines in recent works [31, 33, 57]. We further employ the training methods along with the vanilla baseline network. We can observe that such training methods are essential for training a generalizable visible-infrared image retrieval machine. Noted that the `color jitter` augmentation can improve the performance on the RegDB dataset, i.e., 5.84% for Rank-1 and 3.07% for mAP. However, it will lower the performance of the network on the SYSU-MM01 dataset. The main conjecture is that the SYSU-MM01 dataset is much larger than RegDB dataset, and the network can learn enough color-invariant information in the SYSU-MM01 dataset. As for the RegDB dataset, the network is easily over-fitting to the training data, and `color jitter` can effectively augment the training data, thereby improving the generalization of the network. We also report the ablation of the baseline network in the supplementary material.

5.3 Experiments on cross-modality contrastive learning

In this part, we continue to evaluate the superiority of the contrastive loss in VI-reID task. This study is conducted on VI-ResNet-50 and VI-PCB baselines.

5.3.1 Impact of contrastive loss

We first study the effect of contrastive loss in the VI-reID task. It is shown in Table 2 that each of the intra-modality contrastive loss (i.e., $\mathcal{L}^{\text{intra}}$) and inter-modality contrastive

Table 2 Evaluation of the proposed cross-modality contrastive learning on SYSU-MM01 and RegDB datasets

Contrastive Loss	SYSU-MM01		RegDB	
	R-1	mAP	R-1	mAP
VI-ResNet-50	61.76	59.02	84.82	75.60
$\mathcal{L}^{\text{intra}}$	62.94	60.17	66.38	78.21
$\mathcal{L}^{\text{inter}}$	63.52	60.34	86.08	78.79
$\mathcal{L}^{\text{intra}}$ & $\mathcal{L}^{\text{inter}}$	**64.88**	**61.76**	**88.44**	**81.62**
VI-PCB	66.71	64.71	90.57	83.05
$\mathcal{L}^{\text{intra}}$	67.94	65.11	91.63	84.20
$\mathcal{L}^{\text{inter}}$	68.27	65.27	91.82	84.07
$\mathcal{L}^{\text{intra}}$ & $\mathcal{L}^{\text{inter}}$	**69.97**	**67.42**	**93.40**	**86.77**

Bold values indicate the best result

Table 3 Evaluation of different losses on SYSU-MM01 dataset

Loss		R-1	mAP
(i)	$\mathcal{L}_{\text{ce}}$	48.52	45.27
(ii)	$\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}}$	61.76	59.02
(iii)	$\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{stri}}$	58.72	56.81
(iv)	$\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{cts}}$	55.81	54.70
(v)	$\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{stri}}$	62.68	59.68
(vi)	$\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{cts}}$	**64.88**	**61.76**

Bold values indicate the best result

$\mathcal{L}_{\text{ce}}$, $\mathcal{L}_{\text{tri}}$, $\mathcal{L}_{\text{stri}}$ and $\mathcal{L}_{\text{cts}}$ indicate cross-entropy loss, triplet loss with hard mining, triplet loss with semi-hard mining and the proposed cross-modality contrastive loss, respectively

Table 4 Comparison of the effectiveness of the proposed cross-modality contrastive loss $\mathcal{L}_{\text{cts}}$ and the hetero-center loss $\mathcal{L}_{\text{hc}}$

Loss	SYSU-MM01		RegDB	
	R-1	mAP	R-1	mAP
VI-ResNet-50	61.76	59.02	84.82	75.60
$\mathcal{L}_{\text{hc}}$	63.42	60.70	86.47	78.66
$\mathcal{L}_{\text{cts}}$	**64.88**	**61.76**	**88.44**	**81.62**
VI-PCB	66.71	64.71	90.57	83.05
$\mathcal{L}_{\text{hc}}$	67.96	65.21	91.82	84.30
$\mathcal{L}_{\text{cts}}$	**69.97**	**67.42**	**93.40**	**86.77**

Bold values indicate the best result

Table 5 Evaluation of different projection heads on SYSU-MM01 and RegDB datasets

Projection Head (h)	SYSU-MM01		RegDB	
	R-1	mAP	R-1	mAP
VI-ResNet-50	61.76	59.02	84.82	75.60
`identity head`	56.62	55.78	84.62	73.20
`linear head`	61.98	59.90	86.12	77.69
`non-linear head`	**64.88**	**61.76**	**88.44**	**81.62**
VI-PCB	66.71	64.71	90.57	83.05
`identity head`	62.07	60.29	88.61	80.20
`linear head`	67.11	64.89	90.82	83.82
`non-linear head`	**69.97**	**67.42**	**93.40**	**86.77**

Bold values indicate the best result

loss (i.e., $\mathcal{L}^{\text{inter}}$) can improve the retrieval performance over baseline networks. We also find that combining both contrastive losses can further bring performance gain, indicating that two modifications of contrastive loss learn complementary information in the dataset.

Both triplet loss and contrastive loss are popular options for representation learning. We use both loss functions to optimize our network. To verify and compare the effectiveness of the two schools of losses, we include another study. We also include a loss, triplet loss with a semi-hard mining strategy (denoted by $\mathcal{L}_{\text{stri}}$) [8], to justify our assumption that optimizing non-hard samples benefits the network training. Table 3 suggests that optimizing hard samples by triplet loss is much more important than non-hard samples by contrastive loss for learning a good representation (see (ii), (iii) and (iv)). Beyond optimizing hard samples by triplet loss, such non-hard samples can also provide useful information for the optimization process of a network. That said, the usage of contrastive loss in our work is non-trivial and the network indeed benefits from the combination of triplet loss and contrastive loss.

The hetero-center loss [9, 32] is also a popular loss in the VI-reID. We further study the hetero-center loss on two baselines, and the empirical results are reported in Table 4. It shows the hetero-center loss (i.e., $\mathcal{L}_{\text{hc}}$) can improve the performance of both baselines. While the proposed cross-modality contrastive loss (i.e., $\mathcal{L}_{\text{cts}}$) attains superior performance than the hetero-center loss. Along with the empirical evaluation, we further conduct a t-test to demonstrate the statistical significance of the improvement, and obtain the p-value of 0.0026, meaning that our results are statistically significant ($p < 0.05$ is significant). Thus, we believe that our the proposed contrastive loss is superior to the hetero-center loss.

5.3.2 Impact of projection head

The work in [10] suggests that the projection head is an essential component for contrastive learning, in the sense that the projection head can maintain more information about the data information (i.e., color, pose of objects etc). Thus we evaluate three architectures of projection head in Table 5. The `identity head` indicates an identity mapping. Table 5 reveals that in both baseline networks, the `non-linear head` is better than the `linear head` and the `identity head` counterparts, and a similar observation is also made in [10]. We can also observe that using the contrastive loss without projection head (i.e., `identity head`) even degrades the performance over baseline networks. A possible

Table 6 Evaluation of different temperatures on SYSU-MM01 and RegDB datasets

Temperature (τ)	SYSU-MM01		RegDB	
	R-1	mAP	R-1	mAP
VI-ResNet-50	61.76	59.02	84.82	75.60
$\tau = 0.05$	62.50	60.23	85.91	77.73
$\tau = 0.1$	**64.88**	**61.76**	**88.44**	**81.62**
$\tau = 0.5$	53.85	52.92	83.69	73.69
$\tau = 1.0$	56.71	53.70	82.76	72.62
VI-PCB	66.71	64.71	90.57	83.05
$\tau = 0.05$	67.18	65.08	91.54	83.97
$\tau = 0.1$	**69.97**	**67.42**	**93.40**	**86.77**
$\tau = 0.5$	59.43	60.20	87.62	79.93
$\tau = 1.0$	63.37	60.92	85.44	76.59

Bold values indicate the best result

conjecture of this situation is that the loss value of contrastive loss is large, which affects the optimization direction of the network; thereby affecting the discrimination of features.

5.3.3 Impact of temperature

Temperature (i.e., τ) is another important hyper-parameter that affects the loss value of contrastive loss. The results and comparisons shown in Table 6 reveal that: (1) A small value of temperature (i.e., 0.05 and 0.1) in contrastive loss can bring performance gain over the baseline network. (2) In contrast, the performance of the network will be degraded when the temperature value in contrastive loss is large (i.e., 0.5 and 1.0).

5.3.4 Qualitative results

We illustrate some qualitative results in Fig. 5, to verify the superiority of the proposed CMCL on the SYSU-MM01 dataset. It is observed that our method can improve the hit positions in the ranking list and retrieve more images correctly as compared to its baseline, which indicates the superior property of the proposed contrastive learning scheme.

To further justify the effectiveness, we produce some visualizations of the feature maps on the baseline, baseline $+\mathcal{L}_{\text{intra}}$ and baseline $+\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}$, as shown in Fig. 6. Images are sampled from the SYSU-MM01 dataset. Figure 6a and b indicate the samples from ID7 and ID53, respectively. In each person, from left to right, (1) the input person image, (2) the feature map of baseline, (3) the feature map of baseline $+\mathcal{L}_{\text{intra}}$ and (4) the feature map of baseline $+\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}$. It shows that the proposed cross-modality contrastive loss can help the network to focus on the discriminative and common areas of two modalities of images, resulting in performance gain over the baseline.

We provide the similarity distribution of the positive and negative pairs, as shown in Fig. 7. Figure 7a–d show the distance distributions generating from (a): initial features, (b): baseline, (c): baseline $+\mathcal{L}_{\text{intra}}$ and (d) baseline $+\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}$, respectively. This study justifies the effectiveness of each proposed loss component. Specifically, Fig. 7a and b

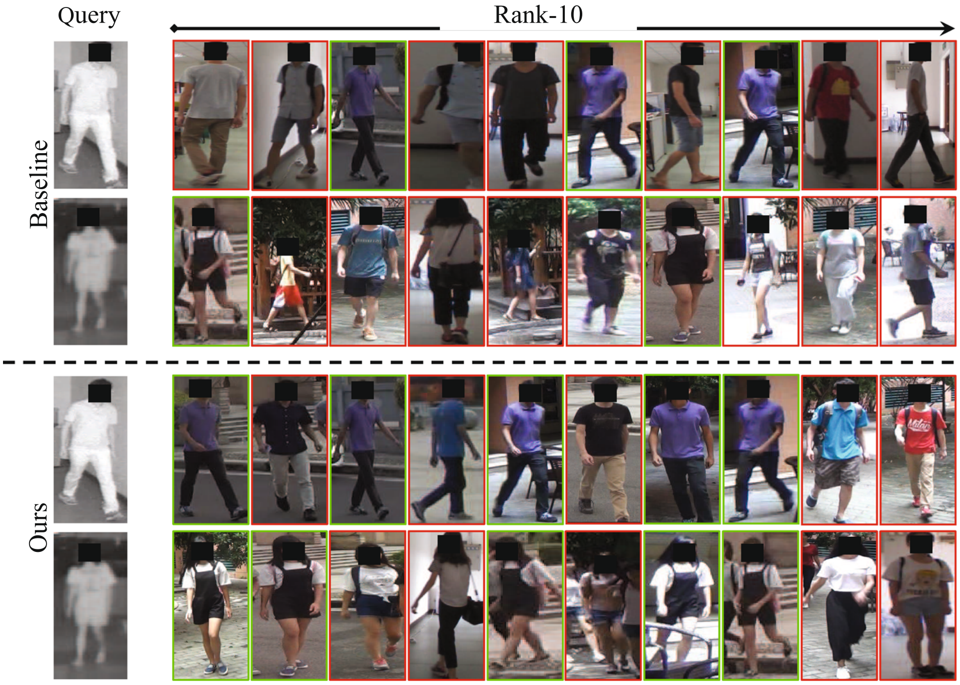


Fig. 5 The ranking lists obtained from baseline network and our method. The correct and false hits are enclosed in green and red boxes. Faces are covered using black boxes for privacy purposes. Best viewed in color

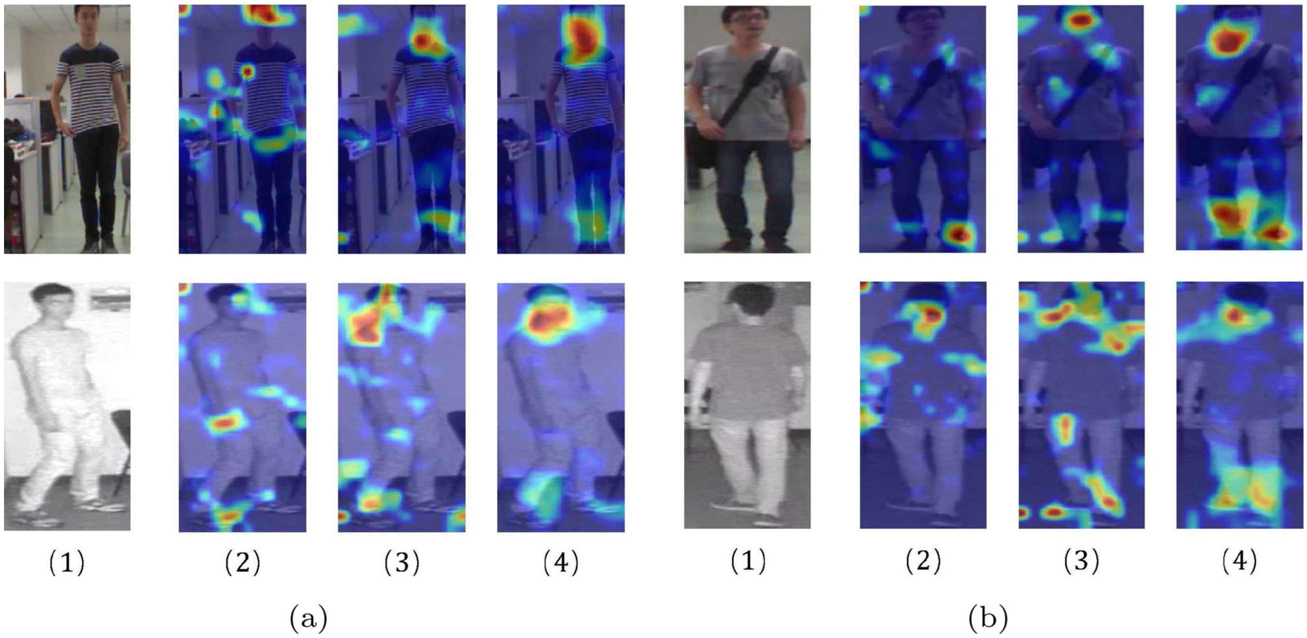


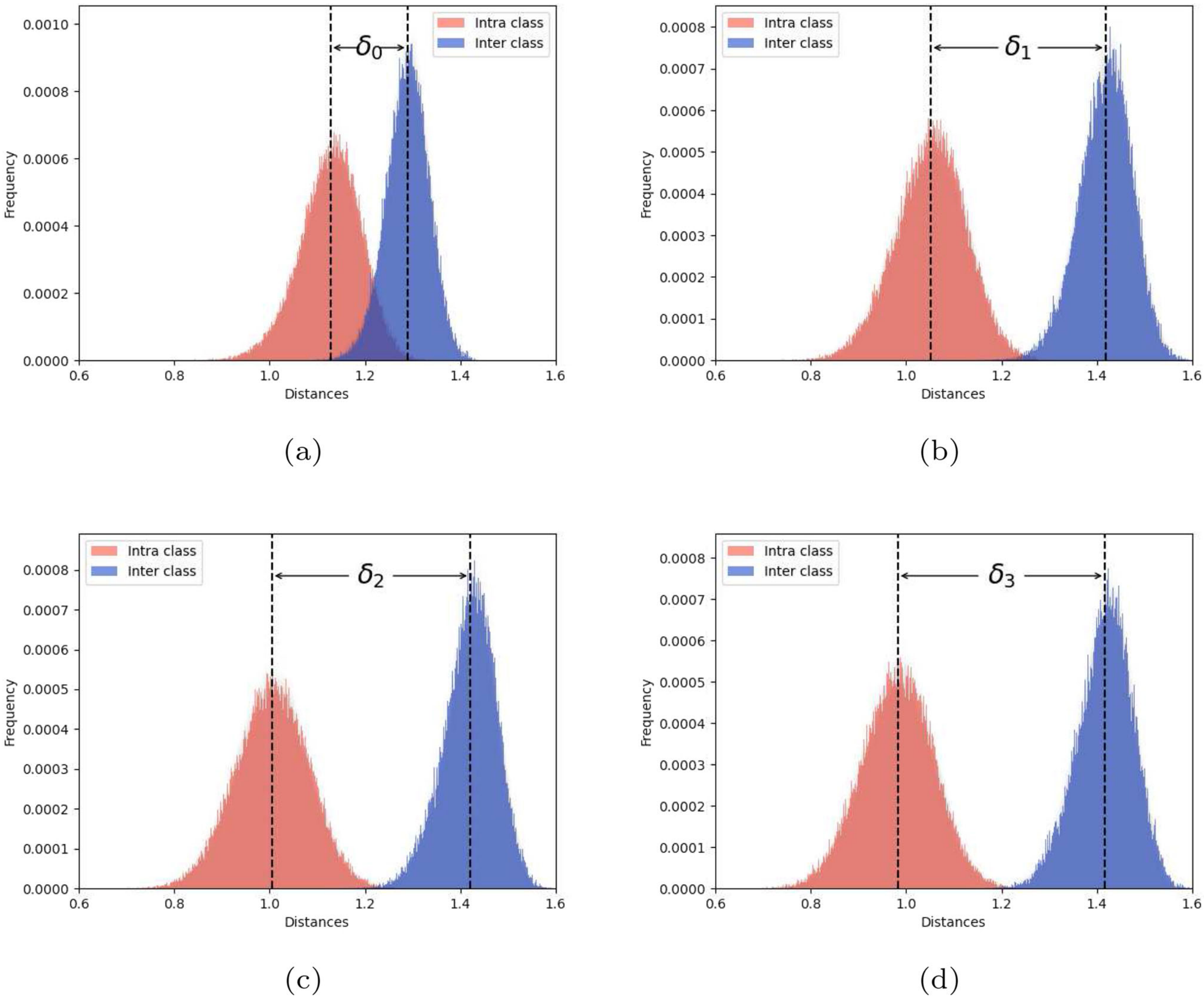
Fig. 6 Visualization of the visible images and infrared images using Grad-CAM. The images are sampled from SYSU-MM01 dataset. In each person, from left to right, (1) the input person image, (2) the feature map of baseline, (3) the feature map of baseline $+\mathcal{L}_{\text{intra}}$ and (4) the feature map of baseline $+\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}$. In the heat map, the response increases from blue to red. Best viewed in color

Table 7 Comparison with the state-of-the-art algorithms on SYSU-MM01 dataset

Models	All search				Indoor search			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
♣ D^{2}RL [35]	28.9	70.6	82.4	29.2	–	–	–	–
♣ HI-CMD [37]	34.94	77.58	–	35.94	–	–	–	–
♣ AlignGAN [34]	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
♣ JSIA-ReID [36]	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9
♣ DG-VAE [58]	59.5	93.8	–	58.5	–	–	–	–
♠ SSFT [27]	47.7	–	–	54.1	–	–	–	–
♠ X-modality [31]	49.92	89.79	95.96	50.73	–	–	–	–
♠ DDAG [4]	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
♠ CMAlign [59]	55.41	–	–	54.14	58.46	–	–	66.33
♠ NFS [6]	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
♠ CoAL [33]	57.2	92.3	97.6	57.2	63.9	95.4	98.9	70.8
♠ CICL+IAMA [54]	57.2	94.3	98.4	59.3	66.6	98.8	99.7	74.7
♠ VSD [5]	60.01	94.18	98.14	58.80	66.05	96.59	99.38	72.98
♠ CM-NAS [13]	61.99	92.87	97.25	60.02	67.01	97.02	99.32	72.95
♠ MCLNet [7]	65.40	93.33	97.14	61.98	72.56	96.98	99.20	76.58
♠ SMCL [32]	67.39	92.87	96.76	61.78	68.84	96.55	98.77	75.56
VI-ResNet-50	61.76	91.28	96.07	59.02	66.01	95.46	98.77	71.68
+ CMCL (Ours)	64.88	93.67	97.12	61.76	71.63	97.82	99.12	76.49
VI-PCB	66.71	94.37	98.01	64.71	72.73	**97.93**	99.59	77.63
+ CMCL (Ours)	**69.97**	**95.26**	**98.27**	**67.42**	**76.48**	97.92	**99.68**	**79.94**

Bold values indicate the best result

The symbols ♣ and ♠ indicate the generative-based algorithms and representation-based algorithms, respectively



(a)

(b)

(c)

(d)

Fig. 7 The distributions of the four types of distances between the cross-modality features. The distance distribution of positive and negative pairs are indicated by red and blue color, respectively. From (**a**) to (**b**), the distance distributions are generate from (**a**): initial features, (**b**): baseline, (**c**): baseline $+\mathcal{L}_{\text{intra}}$ and (**d**) baseline $+\mathcal{L}_{\text{itra}} + \mathcal{L}_{\text{inter}}$. δ is distance of the median value of two distributions

show that the baseline network can significantly improve the gap between positive and negative pairs. Moreover, the $\mathcal{L}_{\text{intra}}$ brings performance gain over the baseline (Fig. 7b vs. c) and the $\mathcal{L}_{\text{inter}}$ can further improve the performance (Fig. 7c vs. d).

5.4 Comparison with state-of-the-art algorithms

To evaluate the superiority of the proposed baseline network and cross-modality contrastive learning scheme, we continue to compare our results with current state-of-the-art algorithms, shown in Tables 7 and 8.

5.4.1 Evaluation on SYSU-MM01 dataset

We first evaluate our model on both “All Search” and “Indoor Search” modes for the SYSU-MM01 dataset. Table 7 shows our method significant outperforms the current state-of-the-art. As observed, “VI-PCB + CMCL” improves the Rank-1/mAP over SMCL [32] by 2.58% / 5.64% on the “All Search” mode, and 7.64% / 4.38% on the “Indoor Search” mode. As compared to another state-of-the-art method, i.e., MCLNet [7], the improvement reads as 4.57% / 5.44% on the a“All Search” mode and 3.92% / 3.36% on the “Indoor Search” mode, w.r.t. Rank-1 / mAP values, vividly showing the superiority of our method.

Table 8 Comparison with the state-of-the-art algorithms on RegDB dataset

Models	Visible to Infrared				Infrared to Visible			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
♣ D^{2}RL [35]	43.4	66.1	76.3	44.1	–	–	–	–
♣ AlignGAN [34]	57.9	–	–	53.6	56.3	–	–	53.4
♣ JSIA-ReID [36]	48.5	–	–	49.3	48.1	–	–	48.9
♣ HI-CMD [37]	70.93	86.39	–	66.04	–	–	–	–
♣ DG-VAE [58]	73.0	86.9	–	71.8	–	–	–	–
♠ SSFT [27]	65.4	–	–	65.6	63.8	–	–	64.2
♠ X-modality [31]	62.21	83.13	91.72	60.18	–	–	–	–
♠ DDAG [4]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
♠ VSD [5]	73.2	–	–	71.6	71.8	–	–	70.1
♠ CMAlign [59]	74.17	–	–	67.64	72.43	–	–	65.46
♠ CICL+IAMA [54]	78.8	–	–	69.4	77.9	–	–	69.4
♠ CoAL [33]	74.1	90.2	94.5	70.0	–	–	–	–
♠ MCLNet [7]	80.31	92.70	96.03	73.07	75.93	90.93	94.59	69.49
♠ NFS [6]	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
♠ SMCL [32]	83.93	–	–	79.83	83.05	–	–	78.57
♠ CM-NAS [13]	84.54	95.18	97.85	80.32	82.57	94.51	97.37	78.31
VI-ResNet-50	84.82	95.53	97.63	75.60	84.69	95.60	97.79	75.73
+ CMCL (Ours)	88.44	97.62	98.56	81.62	88.62	97.51	98.60	81.74
VI-PCB	90.57	97.47	98.61	83.05	90.32	97.27	98.60	82.95
+ CMCL (Ours)	**93.40**	**97.63**	**98.90**	**86.77**	**94.16**	**97.70**	**98.69**	**86.69**

Bold values indicate the best result

The symbols ♣ and ♠ indicate the generative-based algorithms and representation-based algorithms, respectively

5.4.2 Evaluation on RegDB dataset

We further evaluate our proposed method against the state-of-the-art algorithms on both "Visible to Infrared" and "Infrared to Visible" modes for the RegDB dataset. Like before, our method outperforms the state-of-the-art method, i.e., CM-NAS [13] by a tangible margin. Particularly, "VI-ResNet-50 + CMCL" outperforms CM-NAS Rank-1/mAP by 3.90%/1.30% on the "Visible to Infrared" mode, and 6.05%/3.43% on the "Infrared to Visible" mode. Equipped with `part feature learning`, the performance gain of Rank-1 / mAP value is 8.86%/6.45% on the "Visible to Infrared" mode, and 11.59%/8.38% on the "Infrared to Visible" mode, respectively. This huge performance gain again shows the effectiveness of the proposed method.

6 Conclusion

In this paper, we first contribute a strong baseline network for the visible-infrared person re-identification task, by means of extensively studying various training methods in pieces of literature. Our strong baseline (i.e., VI-PCB) is able to reach 90.57% Rank-1 accuracy and 83.05% mAP accuracy on RegDB dataset. Inspired by the successful practice of contrastive learning in self-supervised learning, we further propose a novel but simple cross-modality contrastive learning scheme, which explicitly aligns the embeddings for visible images and infrared images. We conduct thorough experiments to verify the superior performance of the proposed loss function, which improves the state-of-the-art results by a considerable margin. In the future, we will study more effective ways, including the memory mechanism [60], to improve the retrieval performance and build a large scale dataset, containing more complicated factors, for VI-reID task.

Acknowledgements This work was supported in part by the National Science Foundation of China under Grant 62306070 and in part by the Southeast University Start-Up Grant for New Faculty under Grant 4009002309. Furthermore, the work was also supported by the Big Data Computing Center of Southeast University.

Data Availability In this paper, we use two datasets to evaluate the proposed algorithms, namely, SYSU-MM01 and RegDB. The SYSU-MM01 dataset is available at: https://github.com/wuancong/SYSU-

MM01, and the REegDB dataset is available at: http://dm.dongguk.edu/link.html.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2021)
2. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Trans. Multimedia **22**, 2597–2609 (2020)
3. Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5380–5389 (2017)
4. Ye, M., Shen, J., Crandall, D.J., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: European Conference on Computer Vision, pp. 229–247 (2020)
5. Tian, X., Zhang, Z., Lin, S., Qu, Y., Ma, Y.X.L.: Farewell to mutual information: variational distillation for cross-modal person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1522–1531 (2021)
6. Chen, Y., Wan, L., Li, Z., an Zongyuan Sun, Q.J.: Neural feature search for rgb-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 587–597 (2021)
7. Hao, X., Zhao, S., Ye, M., Shen, J.: Cross-modality person re-identification via modality confusion and center aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16403–16412 (2021)
8. Fang, P., Zhou, J., Roy, S.K., Petersson, L., Harandi, M.: Bilinear attention networks for person retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8030–8039 (2019)
9. Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D.: Hetero-center loss for cross-modality person re-identification. Neurocomputing **386**, 97–109 (2020)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 36th International Conference on Machine Learning, pp. 1597–1607 (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
12. 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. Sensors **17**, 605 (2017)
13. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11823–11832 (2021)
14. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
15. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
16. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1249–1258 (2016)
17. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: European Conference on Computer Vision, pp. 501–518 (2018)
18. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2109–2118 (2018)
19. Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., Huang, G., Hariharan, B., Weinberger, K.Q.: Resource aware person re-identification across multiple resolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8042–8051 (2018)
20. Saquib Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 420–429 (2018)
21. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3960–3969 (2017)
22. Tay, C.-P., Roy, S., Yap, K.-H.: Aanet: attribute attention network for person re-identifications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7134–7143 (2019)
23. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: European Conference on Computer Vision, pp. 475–491 (2016)
24. Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3186–3195 (2020)
25. Fang, P., Zhou, J., Roy, S.K., Ji, P., Petersson, L., Harandi, M.: Attention in attention networks for person retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 4626 (2021)
26. Wu, A., Zheng, W.S., Gong, S., Lai, J.: RGB-IR person re-identification by cross-modality similarity preservation. Int. J. Comput. Vis. **128**, 1765 (2020)
27. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13379–13389 (2020)
28. Zhang, Q., Lai, J., Xie, X.: Learning modal-invariant angular metric by cyclic projection network for vis-nir person re-identification. IEEE Trans. Image Process. **30**, 8019 (2021)
29. Wei, Z., Yang, X., Wang, N., Gao, X.: Flexible body partition-based adversarial learning for visible infrared person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **33**, 4676–4687 (2022)
30. Wu, Q., Dai, P., Chen, J., Lin, C.-W., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4330–4339 (2021)
31. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4610–4617 (2020)

32. Wei, Z., Yang, X., Wang, N., Gao, X.: Syncretic modality collaborative learning for visible infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 225–234 (2021)
33. Wei, X., Li, D., Hong, X., Ke, W., Gong, Y.: Co-attentive lifting for infrared-visible person re-identification. In: ACM International Conference on Multimedia, pp. 1028–1037 (2020)
34. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3623–3632 (2019)
35. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.-Y., Satoh, S.: 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, pp. 618–626 (2019)
36. Wang, G.-A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12144–12151 (2020)
37. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: 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, pp. 10257–10266 (2020)
38. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 4321–4329 (2019)
39. Liu, H., Chai, Y., Tan, X., Li, D., Zhou, X.: Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification. IEEE Signal Process. Lett. **28**, 653–657 (2021)
40. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758 (2021)
41. Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent—a new approach to self-supervised learning. In: Thirty-fourth Conference on Neural Information Processing Systems, pp. 21271–21284 (2020)
42. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Thirty-fourth Conference on Neural Information Processing Systems, pp. 9912–9924 (2020)
43. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: The International Conference on Learning Representations, pp. 1–14 (2019)
44. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? In: Thirty-fourth Conference on Neural Information Processing Systems, pp. 6827–6839 (2020)
45. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3299–3309 (2021)
46. Han, J., Shoeiby, M., Petersson, L., Armin, M.A.: Dual contrastive learning for unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2021)
47. Udandarao, V., Maiti, A., Srivatsav, D., Vyalla, S.R., Yin, Y., Shah, R.R.: Cobra: contrastive bi-modal representation algorithm. arXiv preprint arXiv:2005.03687 (2020)
48. Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 833–842 (2021)
49. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: towards unified-modal understanding and generation via cross-modal contrastive learning. In: The 59th Annual Meeting of the Association for Computational Linguistics (2021)
50. Han, Y., Chen, C., Tewfik, A., Glicksberg, B., Ding, Y., Peng, Y., Wang, Z.: Cross-modal contrastive learning for abnormality classification and localization in chest x-rays with radiomics using a feedback loop. arXiv preprint arXiv:2104.04968 (2021)
51. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
52. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**, 211–252 (2015)
53. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 13001–13008 (2020)
54. Zhao, Z., Liu, B., Chu, Q., Lu, Y., Yu, N.: Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3520–3528 (2021)
55. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2018)
56. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: Thirty-first Conference on Neural Information Processing Systems, pp. 1–4 (2017)
57. 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. In: ACM International Conference on Multimedia, pp. 1028–1037 (2020)
58. Pu, N., Chen, W., Liu, Y., Bakker, E.M., Lew, M.S.: Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In: ACM International Conference on Multimedia, pp. 1028–1037 (2020)
59. Park, H., Lee, S., Lee, J., Ham, B.: Learning by aligning: visible-infrared person re-identification using cross-modal correspondences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12046–12055 (2021)
60. Tutsoy, O., Polat, A., Colak, S., Balikci, K.: Development of a multi-dimensional parametric model with non-pharmacological policies for predicting the covid-19 pandemic casualties. IEEE Access **8**, 225272 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Pengfei Fang is an Associate Professor at the School of Computer Science and Engineering, Southeast University (SEU), China, and he is also a member of the PALM lab. Before joining SEU, he was a postdoctoral fellow at Monash University in 2022. He received the Ph.D.

degree from the Australian National University and DATA61-CSIRO in 2022, and the M.E. degree from the Australian National University in 2017. His research interests include computer vision and machine learning.

Yukang Zhang received his B.E. degree and M.S. degree from China University of Geosciences and Central China Normal University, and is currently studying for a Ph.D. degree in computer science and technology in the School of Informatics, Xiamen University. His research interests include machine learning and computer vision.

Zhenzhong Lan received his Ph.D. degree from Carnegie Mellon University, Pittsburgh, in May 2017. He is currently an Assistant Professor with Westlake University, Hangzhou. From 2018 to 2020, he worked at Google AI, Los Angeles, as an NLP researcher. His current research interests lie in data-driven video and natural language understanding. He is also a member of the committees for several of the premier computer vision and multimedia conferences, including CVPR, ICCV, ECCV, and ACM MM.