Intermediary-Generated Bridge Network for RGB-D Cross-modal Re-identification

JINGJING WU, School of Computer Science and Information Engineering, Hefei University of Technology, China
RICHANG HONG, Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology) Ministry of Education, School of Computer Science and Information Engineering, Hefei University of Technology, China
SHENGENG TANG, Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology) Ministry of Education, School of Computer Science and Information Engineering, Hefei University of Technology, China

RGB-D cross-modal person re-identification (re-id) targets at retrieving the person of interest across RGB and depth image modalities. To cope with the modal discrepancy, some existing methods generate an auxiliary mode with either inherent properties of input modes or extra deep networks. However, such useful intermediary role included in generated mode is often overlooked in these approaches, leading to insufficient exploitation of crucial bridge knowledge. By contrast, in this paper, we propose a novel approach that constructs an intermediary mode through the constraints of self-supervised intermediary learning, which is freedom from modal prior knowledge and additional module parameters. We then design a bridge network to fully mine the intermediary role of generated modality through carrying out multi-modal integration and decomposition. For one thing, this network leverages a multi-modal transformer to integrate the information of three modes via fully exploiting their heterogeneous relations with the intermediary mode as the bridge. It conducts the identification consistency constraint to promote cross-modal associations. For another, it employs circle contrastive learning to decompose the cross-modal constraint process into several subprocedures, which provides the intermediate relay during pulling two original modalities closer. Experiments on two public datasets demonstrate that the proposed method exceeds the state-of-the-arts. The effectiveness of each component in this method is verified through numerous ablation studies. Additionally, we have demonstrated the generalization ability of the proposed method through experiments.

CCS Concepts: • **Computing methodologies**; • **Image representations**; • **Object identification**;

Additional Key Words and Phrases: RGB-D cross-modal person re-identification, Auxiliary modal generation, Self-supervised intermediary learning, Heterogeneous relation integration, Cross-modal contrastive learning decomposition.

1 INTRODUCTION

Person re-identification (re-id) has been a fundamental task in computer vision for decades, aiming to retrieve the persons across non-overlapping cameras. Recently, with the flourishing of radar technology, the high-quality

Authors' addresses: Jingjing Wu, School of Computer Science and Information Engineering, Hefei University of Technology, Tunxi Road 193, Hefei, China, hfutwujingjing@mail.hfut.edu.cn; Richang Hong, Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology) Ministry of Education, School of Computer Science and Information Engineering, Hefei University of Technology, Tunxi Road 193, Hefei, China, hongrc@hfut.edu.cn; Shengeng Tang, Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology) Ministry of Education, School of Computer Science and Information Engineering, Hefei University of Technology, Tunxi Road 193, Hefei, China, tangsg@hfut.edu.cn.

ACM 2157-6912/2024/7-ART
https://doi.org/10.1145/3682066

depth image can be obtained more conveniently, which has been introduced into pedestrian recognition tasks to realize the person matching across RGB and depth image modalities, i.e., RGB-Depth (RGB-D) cross-modal person re-identification [8, 9, 36, 43, 49]. As depth image avoids the impacts of light variances, RGB-D cross-modal person re-identification can achieve all-weather recognition. Hence, this task can be deployed in a wider range of scenarios, which draws increasing attention in the person re-id community.

However, there exists an enormous discrepancy between RGB and depth image modalities. RGB image represents the color and texture of the person while depth mode depicts the distance information of the scene. This brings tremendous challenges to RGB-D person re-id.

In accordance with the patterns of narrowing modal gap, prevailing cross-modal recognition methods can be broadly categorized into two types: non-generation-based methods and generation-based methods. **Non-generation-based methods** [8, 9, 36, 43, 49] aim to lessen the modal discrepancy by capturing the common features between two modalities. Their framework is summarized in Fig. 1(a), which has been as a dominant cross-modal recognition paradigm. Typically, these methods concentrate on promoting feature extractors and loss functions to bolster feature representation capability. For instance, some approaches [39, 46] design part-based feature extraction networks and introduce attention mechanisms to incorporate spatial contextual clues for enriching feature expression. Additionally, some works [1, 36] strengthen the constraints of feature learning between the two modalities to align them. Nevertheless, these works straightly handle such significant modal discrepancy by mapping the representations of two modalities into the same space, whose tolerance for the modal gaps is limited.

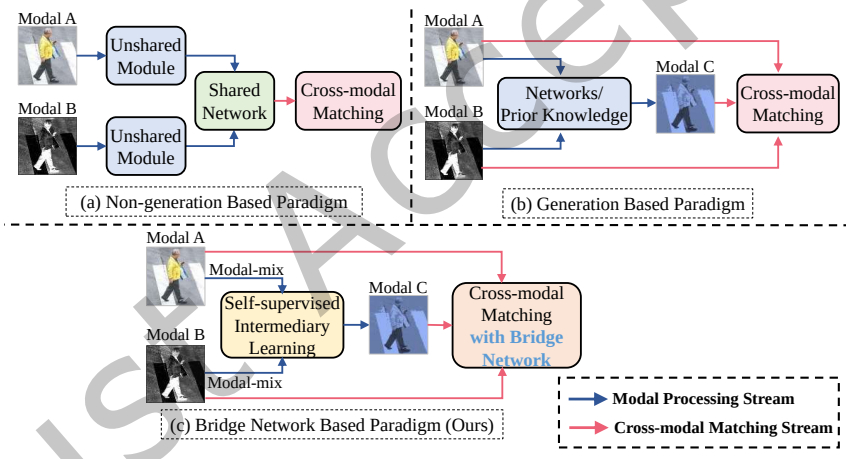

Fig. 1. Difference demonstrations between the proposed method and existing cross-modal person identification approaches. (a) Most non-generation-based methods aim to capture the common features from distinct modalities. (b) Prior generation-based methods either exploit additional deep networks or prior information to produce transition modes, and directly perform cross-modal matching on original and generated modalities. (c) The proposed method produces an intermediary mode of input modalities with self-supervised intermediary learning instead of prior properties and additional network parameters, which fully exploits the generated modal to bridge modal discrepancy by the bridge network during cross-modal matching.

To further narrow the modal gap, some **generation-based methods** [3, 14, 33, 34] have been investigated to provide transition modality of RGB and depth modalities, which build a bridge for lessening their discrpancy. Prior generation-based methods either leverage auxiliary networks or prior information to produce transition modes. Their overviews are summarized in Fig. 1(b). For example, some attempts [3, 4, 33, 34] adopt generative

adversarial networks (GANs) to produce cross-modal images for each modality, which are then matched with real images of the same modality. Partmix [16] applys part detection network to blend the part-based features from two modalities, which regularizes the models from overfitting to the training data via Mixup technology [42]. MID [14] designs a deep reinforcement learning framework to determine the combination ratio of two modalities for providing auxiliary modality. Work [40] fully utilizes the modal prior information, which generates a grayscale modality by preserving the structural information of both modalities and discarding the color information that only exists in one modality.

Despite significant progress achieved by existing generation-based cross-modal works, applying them on RGB-D cross-modal identification may account for the following two drawbacks: On one hand, it is difficult to create high-quality cross-modal images for RGB and depth modalities through deep networks, because generating three-dimensional (3D) depth images from two-dimensional (2D) RGB images has always been an ill-posed problem in the field of computer vision. Besides, the utilization of deep network attributes to expensive computational cost. On the other hand, it is tough to select a realistic mode that contains the contents of both RGB and depth modalities. Moreover, the exploitation of modal prior information hinders the modules from applying in other scenarios.

To build a generation-based approach that is free from the fetter of generative model and prior knowledge, this paper proposes a modal-mix operation inspired by Mixup [42] to generate an auxiliary mode with two input modalities. The key idea of modal-mix is to introduce a virtual modal for each instance, whose intermediary effect is deduced theoretically. Since it produces this intermediary modal with randomness instead of the inherent properties of original modality for advancing network generalization, it may yield some inferior samples. To alleviate ambiguous quality in modal generation, inspired by self-supervised learning [10, 37], the cross-modal self-supervised intermediary learning is leveraged on the generated and two input modalities to restrict the modal generation process with avoiding adding extra module. In contrast to previous Mixup-based method [16] that relies on additional modules to fuse two modalities for preventing model overfitting, this paper adopts a self-supervised intermediary learning approach to bridge the disparities between RGB and depth modalities. The objectives and implementation processes of these approaches differ significantly.

Afterward, this paper adequately utilizes the bridge role of generated modes by presenting a bridge network, which differs from previous generation-based methods [4, 14, 16, 33] that feed their features into cross-modal matching directly. This network first exploits a multi-modal transformer to aggregate the information of triple modalities after fully mining multi-modality clues. Specifically, it attends to the generated modal to propagate the transition information while exploring the correlations between original modalities. The utilization of intermediate modes accounts for a smoother modeling process of heterogeneous relationship. Then, we apply the identification consistency learning on the integrated triple-mode features to strengthen multi-mode constraints, consequentially enhancing the cross-modal associations.

To further produce the bridge effect of generated modality, this bridge network decomposes the process of cross-modal constraints via circle contrast learning. It enforces the characteristics of two original modalities to be similar to the features of the intermediary mode at the category-level during pulling two original modes closer. In this way, the generated modal acts as an intermediate relay for two original modalities, which is capable to provide intermediate effects, resulting in the link ability for modal gaps.

To summarize, we propose a generation-based RGB-D cross-modal person re-identification method, presented in Fig. 1(c). It sufficiently exploits the generated intermediary modal to bridge the modal gaps between two input modals, which is termed as intermediary-generated bridge network (IBN). The contributions of this paper are as follows:

(1) This paper proposes a modal generation approach to construct the intermediary modality of two input modalities. It designs cross-modal self-supervised intermediary learning to guarantee the quality of generated samples, which is freedom from modal prior knowledge and additional module parameters.

(2) This paper puts forward a bridge network to adequately mine the bridge role of generated intermediary modal, which employs the multi-modal transformer and circle contrast learning to provide an intermediate relay through generated modality during multi-modal integration and decomposition.

(3) Experiments on several public datasets demonstrate that the performance of the proposed method is superior in RGB-D cross-modal person re-id tasks. In addition, extensive ablation studies illustrate the effectiveness of each part in IBN.

2 RELATED WORKS

In this section, we mainly introduce the differences and relations between the proposed method and existing methods in related fields.

2.1 RGB-IR Person Re-identification

The performance of most RGB-based methods [6, 22, 31] would deteriorate under scenes with significant illumination variations. Especially in inferior light environment, such as the night scene, it is infeasible to capture RGB images with sufficiently high-quality for person re-id task. To address this issue, RGB-IR cross-modal re-id is presented [7, 35] to match the person across RGB and infrared (IR) modalities.

Existing RGB-IR methods are mainly divided into two categories: non-generation-based methods and generation-based ones. Non-generation-based methods [39, 45, 46] focus on capturing the common features of two modals, which attempt to narrow modal gap by enhancing feature representation. Some efforts [39, 46] have been made to design part-based attention modules to obtain part-based relations for enriching feature expression. Besides, work [21] proposes a memory-augmented unidirectional learning method to learn explicit cross-modality metrics in two uni-directions and further enhances them with memory-based augmentation. Nevertheless, these works straightly handle such significant modal discrepancy by mapping the representations of two modalities into the same space, whose tolerance for the modal gaps is limited.

To further narrow the modal gap, some attempts have investigated auxiliary mode generation for this task. Prior RGB-IR generation-based methods can be categorized into two finer types: generation-network-based approaches and prior-information-based methods. The former approachs [3, 14, 16, 29, 33, 34, 47] employ deep networks to provide generated modal. For example, some methods [3, 33, 34] adopt generative adversarial networks (GANs) to produce cross-modal images for each modality. Subsequently, a feature alignment module is adopted to match the real images with the generated images of the same modality. Work [16] utilizes part detection network to fuze the part-based features of two modals with Mixup-based idea, which regularizes the models from overfitting to the training data. It then adopts contrastive learning to pull positive sample pairs close and push negative sample pairs far. MID [14] designs a deep reinforcement learning framework to decide the combination ratio of two modalities for providing auxiliary mode. By contrast, the latter ones [17, 40, 48] leverage prior properties of original modals to construct the transition modal. For instance, in the work [40], cross-modal identification is achieved by generating a grayscale modality that preserves the structural information of RGB images and discards the color information not presented in IR images.

Despite significant progress achieved by existing RGB-IR generation-based works, applying them on RGB-D cross-modal identification may account for the following two drawbacks: On one hand, it is difficult to create high-quality cross-modal images for RGB and depth modalities through deep networks, because generating three-dimensional (3D) depth images from two-dimensional (2D) RGB images has always been an ill-posed problem in the field of computer vision. Besides, the utilization of deep network attributes to their expensive computational cost. On the other hand, it is tough to select a realistic mode that contains the contents of both RGB and depth modalities. Moreover, the exploitation of modal prior information hinders the modules from applying in other scenarios.

Thus, this paper proposes a generation-based identification method tailored for RGB-D cross-modal person re-id, which is freedom from modal prior knowledge and additional module parameters. It exploits a modal-mix operation inspired by Mixup to construct the intermediary mode of two input modes, whose quality is guaranteed by cross-modal self-supervised intermediary learning. Although minor cross-modal methods [14, 16] also utilize Mixup-based approaches, the proposed generation-based method differs from them in the following aspects: 1) Unlike these methods [14, 16], which rely on auxiliary networks to produce transition modes, the proposed generation approach operates independently of deep models and prior knowledge. 2) This paper adequately utilizes the bridge role of generated modes by presenting a bridge network, which differs from previous methods [14, 16] that feed their features into cross-modal matching directly. 3) Work [16] promotes the Mixup method to realize a data augmentation method suitable for cross-modal pedestrian recognition, which aims to enhance model generalization performance. By contrast, this paper leverages the Mixup technology to generate transition modalities, which devotes to bridging the gap between two original modalities. 4) In contrast with [16], which employs contrastive learning to enhance feature discrimination by pulling positive sample pairs close and pushing negative sample pairs far, this paper utilizes contrastive learning to ensure the quality of generated modalities and narrow the modal gap by enforcing a bridge constraint between them. Their purposes and applications are fundamentally different. In summary, our method diverges from existing approaches in terms of mode generation and usage, reflecting diverse goals and methodologies.

2.2 RGB-D Person Re-identification

Infrared images are easily affected by environmental temperature. For example, if the human body temperature is similar to the ambient temperature, especially in summer, it is difficult to obtain high-quality infrared images. By contrast, the depth image that reflects the depth information of the scene avoids the impacts of light and heat variance. These images remain unchanged even if the pedestrians change their clothes. Nowadays, with the rapid development of radar technology, it is more convenient to capture depth and skeleton information with depth camera, such as Microsoft Kinect. Therefore, depth image has been widely exploited in many computer vision fields. Specifically, in the field of pedestrian recognition, with the acquisition of RGB and depth images, RGB-D dual-modal and RGB-D cross-modal pedestrian re-id tasks are proposed.

Some works [15, 25, 38] devote to the research of RGB-D dual-modal person re-id. They combine RGB and depth information to jointly realize pedestrian recognition, which enhance the accuracy of recognition by forming the complementation of two images. For example, literature [25] combines the color histogram extracted from the RGB image and the pedestrian height feature captured from the depth image. John et al. [15] integrate RGB height histogram and depth gait feature information. Xu et al. [38] exploit the depth data to assist RGB-based pedestrian recognition. Work [28] boosts the accuracy of recognition by fusing the appearance features extracted from RGB images and the anthropometric features captured from depth images.

A few papers [8, 9, 36, 43, 49] begin to concentrate on RGB-D cross-modal pedestrian re-id. Literatures [43, 49] utilize hand-crafted features to identify pedestrians, which lack semantically abstract expression. To solve this problem, several efforts [8, 9] have been devoted to constructing a RGB-D cross-modal deep network. They first train the single-mode pedestrian recognition network with depth images. And then, these methods train the single-mode recognition network with RGB images, which adopt distillation learning to constrain the similarity between RGB and depth images taken at the same time, so as to reduce the gaps between these two modes. This method has achieved significant performances. However, it can not be trained in an end-to-end pattern. To solve this problem, work [36] puts forward an end-to-end heterogeneous restraint network, which suppresses the differences between the two modalities by fully exploring cross-modal relationships. Cross-modal RGB-D pedestrian recognition can be applied to scenarios where two modalities cannot be simultaneously obtained, whose application scope is wider. Therefore, this paper focuses on the implementation of this task.

The aforementioned RGB-D cross-modal methods intend to directly handle such huge modal discrepancy by constraining and aligning two modalities, which may be difficult to converge. Hence, this paper proposes a novel solution to narrow the differences between two modalities. Specifically, inspired by the success achieved by RGB-IR cross-modal identification, we generate an auxiliary mode with two original modes and fully exploit it as the bridge to decrease the modal gaps.

2.3 Self-supervised Learning

Self-supervised learning has drawn lots of attention due to its recent success, aiming at learning representations from unlabeled data through accomplishing a pretext task that is derived from self-supervision. There are many manually designed pretext tasks for pre-training, such as image colorization [44], jigsaw puzzle solving [27]. Contrastive learning has shown to be effective in self-supervised learning. Most existing methods [2, 11] perform contrastive learning to enhance the representation ability for a single modality. The concept of contrastive learning is applicable to any modal.

However, in the context of this downstream task, the dissimilarities between the two modalities are substantial, posing challenges to direct contrastive learning. To tackle this issue, this paper introduces a modal-mix generation method tailored to the task requirements. As a result, the paper is capable of leveraging a virtual modality as an intermediary to bridge the gaps between the original modalities. This is achieved through a cross-modal contrastive learning process that involves the splitting of the modality alignment. Distinguishing itself from existing pre-training methods that combine data augmentation and contrastive learning to enhance representation learning, this paper utilizes these techniques with a more specific and intuitive objective: the generation, decomposition, and integration of multiple modalities in downstream tasks. This approach facilitates the mitigation of actual modality differences.

3 METHODS

3.1 Problem Definition

RGB sample and depth sample are denoted as $S^R_{i,j}$ and $S^D_{i,j}$, where R and D stand for RGB modal and depth modal, respectively. Subscript i is the category of the sample, $i \in \{1, 2, \cdots, C\}$, and C refers to the category number of person. Subscript j represents the j^{th} image in the class i, $j \in \{1, 2, \cdots, N_i\}$, and N_i refers to the image number in the class i.

3.2 Overview

The overall framework of this paper is depicted in Fig. 2. We first produce intermediary mode $S^T_{i,j}$ with depth image $S^D_{i,j}$ and RGB image $S^R_{i,j}$ through modal-mix, which is introduced in Sec. 3.3. Let T denote the generation modal type.

Afterward, we input three modalities into the proposed triple-stream network. The network first adopts three un-shared shallow networks to extract modal-specific features $R^D_{i,j}$, $R^T_{i,j}$ and $R^R_{i,j}$ from the images of three modals, $i \in \{1, 2, \cdots, C\}$, $j \in \{1, 2, \cdots, N_i\}$. Then it exploits the shared deep network to extract the deep features of three modes, $F^D_{i,j}$, $F^T_{i,j}$ and $F^R_{i,j}$, $i \in \{1, 2, \cdots, C\}$, $j \in \{1, 2, \cdots, N_i\}$. D, R and T represent the type of modal. The meanings of subscripts i and j are the same as those of the given samples. These features are first input into self-supervised intermediary learning to promote the feature quality of generated modality, which is introduced in Sec. 3.5.1. We then feed the deep features of three modes into the base loss, which is stated in Sec. 3.5.3.

We further input the obtained three modal features into the multi-modal transformer to fully explore the relationships among three modals. This component aggregates the characteristics of three modalities by attending to the learned heterogeneous relationships, whose output is constrained by identification consistency loss to

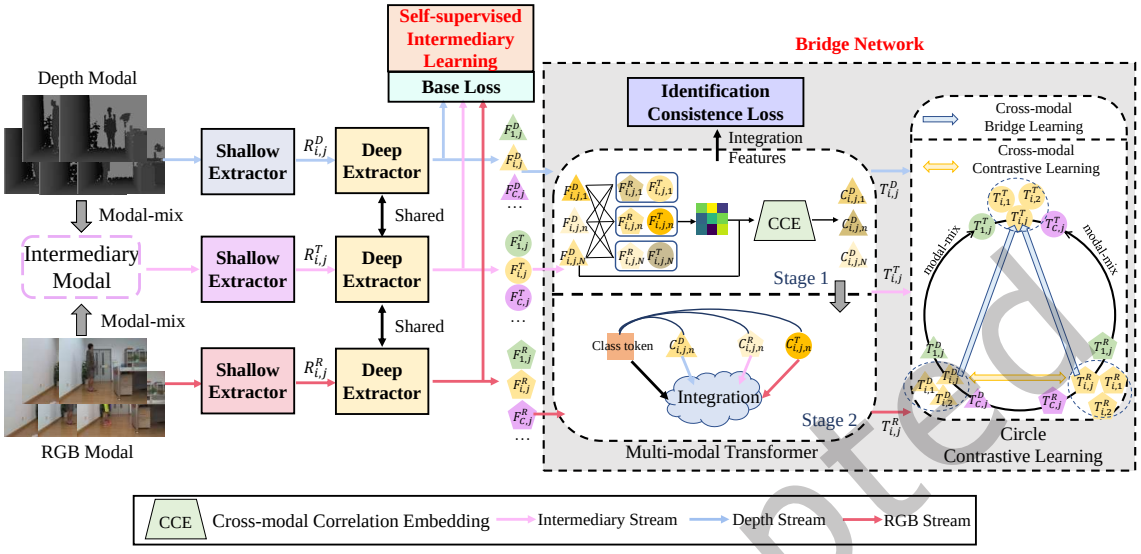

Fig. 2. The overall framework of the intermediary-augmented bridge network (IBN). It consists of triple streams. We adopt round, square and pentagon symbols to stand for the features. The symbol shape represents the feature modality and color denotes the feature category. Best viewed in color.

enhance the cross-modalities interactions. The multi-modal transformer and identification consistency loss are explained in Sec. 3.4 and Sec. 3.5.4, respectively.

Subsequently, we feed three modal features, which are integrated with multi-modal relationships, into circle contrast learning. It decomposes the cross-modal contrast learning process of two original modalities into several subprocedures, which carries out multiple contrastive learning to provide an intermediate relay for modal gap suppression. The circle contrast learning is explained in Sec. 3.5.2.

3.3 Modal-mix

RGB images mainly contain the color and texture of the pedestrian, which are not included in depth images. Besides, the depth image can display the three-dimensional geometric surface information of the person, which is not available in the RGB image. Therefore, there is a tremendous gap between these two images.

To alleviate these gaps, we propose a modal-mix operation to generate the intermediate mode of the two modes. Inspired by data augmentation method Mixup [42], we randomly sample RGB image $S^R_{i,j1}$ and depth image $S^D_{i,j2}$ from class i, $i \in \{1, 2, \cdots, C\}$, $j1, j2 \in \{1, 2, \cdots, N_i\}$. In order to simplify the notation, this paper uniformly records $j1$, $j2$ as j in this subsection. We directly exploit the random linear combination of two modal images to generate their virtual mode $S^T_{i,j}$, which is calculated as follows:

$$S^T_{i,j} = \eta S^R_{i,j} + (1 - \eta) S^D_{i,j} \tag{1}$$

Among that, $\eta \sim \beta(1, 1)$, that is, variable η obeys beta distribution.

We construct N_i samples in each category i. In order to deduce and prove that the generated mode is the intermediate state of two input modes, we first calculate the difference between RGB image $S^R_{i,j}$ and depth image $S^D_{i,j}$:

$$D^{RD} = \| S^R_{i,j} - S^D_{i,j} \| \tag{2}$$

The difference between generation modal and RGB modal can be computed:

$$\begin{aligned} D^{RM} &= \| S^R_{i,j} - S^T_{i,j} \| \\ &= \| S^R_{i,j} - (\eta S^R_{i,j} + (1-\eta) S^D_{i,j}) \| \\ &= \| S^R_{i,j} - \eta S^R_{i,j} - (1-\eta) S^D_{i,j} \| \\ &= (1-\eta) \| S^R_{i,j} - S^D_{i,j} \| \end{aligned} \tag{3}$$

Since $\eta \in [0,1]$, $D^{RM} \leq D^{RD}$. This illustrates the difference between RGB modality and generation modality is smaller than that between RGB modality and depth modality.

Analogously, the difference between the generation modal and the depth modal is:

$$\begin{aligned} D^{DM} &= \| S^D_{i,j} - S^T_{i,j} \| \\ &= \| S^D_{i,j} - (\eta S^R_{i,j} + (1-\eta) S^D_{i,j}) \| \\ &= \eta \| S^D_{i,j} - S^R_{i,j} \| \end{aligned} \tag{4}$$

Since $\eta \in [0,1]$, we can infer that $D^{DM} \leq D^{RD}$. Consequently, we can reach a similar conclusion: The discrepancy between the generated modality and each original modality is smaller compared to the disparity between the two original modalities. Hence, the generated virtual modality can be considered as an intermediary mode between RGB and depth modalities. It serves as a bridge between the two modalities, effectively reducing their dissimilarities.

3.4 Multi-modal Transformer

In order to investigate the relationship among three modes for a given instance, we introduce the concept of a multi-modal transformer. This module is composed of two distinct stages, as illustrated in Fig. 3. The primary objective of the first stage is to acquire cross-modal correlations that enhance the representation of the input modalities. This process facilitates a deeper understanding of the interdependencies among the distinct modes. The second stage is specifically designed to integrate the appearance representations of the three modalities. In this way, it allows for a comprehensive analysis of the combined information provided by each mode, leading to a more holistic understanding of the instance.

Specifically, the features of three modalities, including $F^D_{i,j}$, $F^R_{i,j}$ and $F^T_{i,j}$, are evenly partitioned into N stripes along the horizontal direction to capture detailed part-based features. Take $F^D_{i,j} \in \mathbb{R}^{H \times W \times \bar{C}}$ as an example, H and W represent the feature size, and $\bar{C}$ represents the number of feature channels, it is divided into part-based feature $F^D_{i,j,n} \in \mathbb{R}^{\frac{H}{N} \times W \times \bar{C}}$, where $n \in \{1, 2, \cdots, N\}$ represents the stripe index. Then, average pooling is performed on each stripe $F^D_{i,j,n}$ to derive the pooled feature $\bar{F}^D_{i,j,n} \in \mathbb{R}^{1 \times 1 \times \bar{C}}$, which serves as a spatial element for the depth feature. These N elements $\bar{F}^D_{i,j,n}$ are concatenated to create the feature sequence $\bar{F}^D_{i,j}$. We adopt the same operations on $F^R_{i,j}$ and $F^T_{i,j}$ to capture feature sequences $\bar{F}^R_{i,j}$ and $\bar{F}^T_{i,j}$. We feed the sequences of three modalities into the multi-modal transformer network to comprehensively explore the cross-modal relations.

In stage 1, it captures the cross-attention between various modalities with the shared cross-attention module from feature sequences of three modalities. Take the left cross-attention module in Fig. 3 as an instance, the cross-attention $Catt$ is formulated as follows:

$$Q = Ln([\bar{F}^D_{i,j}, PE]) \tag{5}$$

$$K = V = Ln([Avg(\bar{F}^R_{i,j}, \bar{F}^T_{i,j}), PE]) \tag{6}$$

$$Catt(Q, K, V) = softmax(\frac{Q \cdot Tr(K)}{\sqrt{d_k}})V \tag{7}$$

$Ln()$, $Tr()$ and $Avg()$ denote the linear layers, transposition and average operation, respectively. PE refers to position embedding. $\sqrt{d_k}$ denotes the dimension of K. We assign the combination of $\bar{F}^D_{i,j}$ and position encoding (PE) as Q. As the difference between RGB and depth modalities is significant, directly modeling their relationships is rigid. Consequently, the cross-modal attention module attends to the cross-attention between depth mode and the combination of RGB and intermediary modes. To this end, it conducts an average operation on $\bar{F}^R_{i,j}$ and $\bar{F}^T_{i,j}$, incorporating the resulting outputs with PE to form K and V. This module is designed to capture the cross-attention between each depth element $\bar{F}^D_{i,j,n}$ and the average outputs of elements $\bar{F}^R_{i,j,n}$ and $\bar{F}^T_{i,j,n}$ by computing $softmax(\frac{Q \cdot Tr(K)}{\sqrt{d_k}})$. Thus, this paper captures the relationships between any two elements in RGB and depth modals with the generated modal as the bridge. With the intermediate mode as the bridge, it is conducive to smoothing the relation modelling procedure.

Next, it performs the cross-modal correlation embedding (CCE) operation to introduce the learned cross-modal relationships into depth modal. It accumulates the obtained cross-attention relations for each depth element $\bar{F}^D_{i,j,n}$ by multiplying cross-attention values $softmax(\frac{Q \cdot Tr(K)}{\sqrt{d_k}})$ with V. We add these relations to the depth element, outputting $C^D_{i,j,n}$. All elements $C^D_{i,j,n}$ are concatenated to form cross-modal sequence $C^D_{i,j}$. $C^R_{i,j}$ is gained with the same way. Meanwhile, it inputs the features of generated mode into the linear layer, which outputs $C^T_{i,j}$.

In stage 2, the enhanced features of three modals obtained in stage 1 are connected together and inserted with a class token at the beginning to form the triple-mode sequence. Then, multi-mode transformer exploits the transformer encoder to obtain the pair-wise relationship from the triple-mode sequence due to its powerful relationship modeling ability. Transformer encoder consists of L blocks in series. Each block includes M attention units in parallel. It accumulates the influence of all elements in the sequence for each element. In this way, each element contains not only the relationships within the same mode, but also the relationships across various modalities. Correspondingly, with the intermediate mode, the modeling process of heterogeneous relations becomes more smooth. The outputs of L blocks are denoted as T_i.

There are $(3N+1)$ elements in T_i. The first element in T_i, i.e., $T_i[0]$, integrates the information of three modalities, which is denoted as integration features. This element is initialized randomly and it does not contain any modal information. Therefore, integrating all modal clues into this element can avoid the bias towards a certain modality, which is capable to reflect the statistical characteristics of the three modalities. The last $3N$ elements contain three modal features that have considered the impacts of other modals, including $T^R_{i,j}$, $T^D_{i,j}$ and $T^T_{i,j}$. We input the last $3N$ elements of T_i into circle contrast learning and $T_i[0]$ into identification consistency loss respectively, which are described in Sec. 3.5.2 and Sec. 3.5.4.

Unlike existing methods that utilize the transformer structure solely for establishing self-attention within single-mode images, this paper takes a different approach. Here, the transformer component is employed to model the intricate relationships among various modes and facilitate modal fusion. As a result, the objective of this paper is more focused and specific. Instead of applying transformers within individual modes, the aim is to

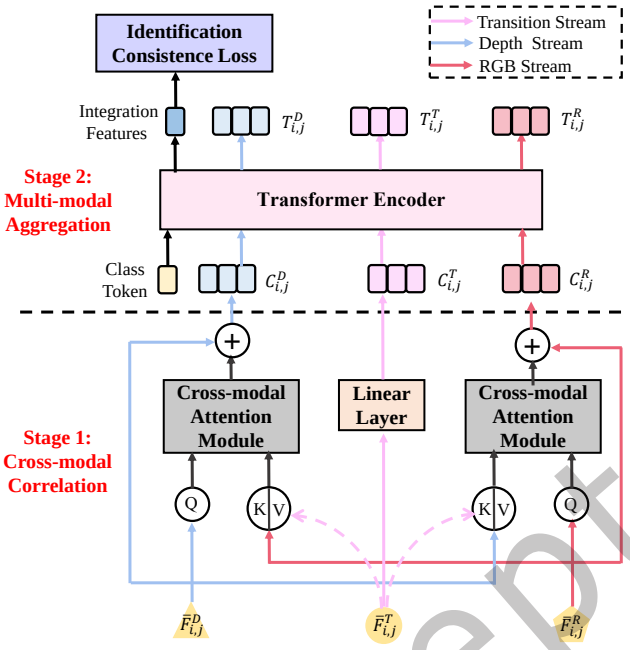

Fig. 3. The overview of the proposed multi-modal transformer.

leverage their capabilities to capture cross-modal dependencies and achieve fusion across different modalities. This novel approach allows for a more comprehensive analysis and understanding of the interplay between modes, leading to enhanced overall performance.

3.5 Loss Functions

We begin by introducing the contrastive learning constraints applied to the generated modality, referred to as self-supervised intermediary learning. Next, we provide detailed explanations of the contrastive learning constraints applied to the RGB/depth modalities, which are denoted as circle comparative learning. Following that, we present the base loss and identification consistency loss. Lastly, we define the overall loss function.

3.5.1 Self-supervised Intermediary Learning. As aformentioned, this paper produces generated samples to bulid a bridge for RGB and depth modalities for narrowing their gaps. The quality of generated samples is determined by their intermediary effect. The more obvious the reduction in disparity between the RGB and depth modalities, the superior the quality of the generated samples.

However, the quality of certain generated samples cannot be guaranteed due to the inherent randomness in the modal generation process. Additionally, the paper does not leverage the prior properties of the original modalities as supplementary cues to guide the mode generation. It is time-consuming to measure the quality of generated samples by evaluating the differences between the RGB and depth modalities after introducing the generated samples. To enhance the quality of the intermediate mode features in such scenarios, we draw inspiration from self-supervised learning. Consequently, this paper adopts a contrast learning pattern to establish cross-modal constraints between the generated modality and the input modalities, achieving this without supervision.

Specifically, we carry out self-supervised intermediary learning to enhance the representations ability of intermediate modal by constraining the features of the generated sample to be more similar to the original ones. Namely, it enforces the features of intermediary sample $F^T_{i,j}$ to be more similar to those of RGB sample $F^R_{i,j}$ and depth sample $F^D_{i,j}$, $j \in \{1, 2, \cdots, N_i\}$, where each element $F^T_{i,j}$ is generated with $F^R_{i,j}$ and $F^D_{i,j}$. Here, $F^R_{i,j}$ and $F^D_{i,j}$ are uniformly denoted as $F^M_{i,j}$, $M \in \{R, D\}$. To this end, this method calculates the cosine similarity between $F^T_{i,j}$ and $F^M_{i,j}$, i.e., $F^T_{i,j}F^M_{i,j}$. Here we enforce all vectors to be L2-normalized feature embeddings, i.e., $\| F^R_{i,j} \| = 1$, $\| F^M_{i,j} \| = 1$. Afterward, the self-supervised intermediary learning loss is formed as follows:

$$L_{t-rd} = \sum_{M, M \in \{R,D\}} L^M_t \tag{8}$$

Among that, L^M_t is computed as follows:

$$\begin{aligned} L^M_t &= -\log \frac{exp(F^T_{i,j}F^M_{i,j}/\tau)}{exp(F^T_{i,j}F^M_{i,j}/\tau) + \sum_n \sum_k exp(F^T_{i,j}F^M_{n,k}/\tau)}, M \in \{R, D\} \\ & n \in \{1, 2, \cdots, C\} \& n \neq i; k \in \{1, 2, \cdots, N_n\} \end{aligned} \tag{9}$$

τ refers to a hyper parameter that controls data distribution level. Higher τ leads to a softer probability distribution. We set it to 0.2 in this paper.

This loss function is designed to strengthen the similarity between $F^T_{i,j}$ and its corresponding $F^R_{i,j}$ and $F^D_{i,j}$, effectively bringing the generated modality closer to the two original modalities. By doing so, it effectively enhances the transition effect of the generated samples, thereby improving their overall quality.

3.5.2 Circle Contrastive Learning. Pulling two original modalities closer is the core of cross-modal constraints in this paper. To this end, this paper conducts contrastive learning on RGB and depth modalities with the bridge of generated modality. As depicted in Fig. 2, the bidirectional cross-modal contrastive learning forms a closed loop, so we term it as circle contrastive learning L_{cm}.

As there exists tremendous modal gap between RGB and depth modalities, circle contrastive learning initially prompts the features of the two original modalities to approach those of the generated modality before bringing them closer together. By doing so, the generated modality acts as an intermediary relay, facilitating the establishment of connections and bridging the gaps between the input modalities. This procedure notably enhances the effectiveness of the subsequent pulling constraints on the two original modalities.

Specifically, for the RGB modality, circle contrastive learning performs two essential restricts. Firstly, it executes a cross-modal bridge loss L^T_r which encourages the RGB samples $T^R_{i,j}$ to resemble more closely the intermediary modality $T^T_{i,j}$. In this way, the generated modality can function as an intermediary relay. Secondly, it employs a cross-modal contrast loss L^D_r to bring $T^R_{i,j}$ closer to $T^D_{i,j}$, effectively reducing the existing modal gaps. Thus, the circle contrastive learning for the RGB modality can be collectively denoted as $L^A_r, A \in \{T, D\}$, and is computed as follows:

$$\begin{aligned} L^A_r &= -\log \frac{exp(T^R_{i,j}T^A_{i,j}/\tau)}{exp(T^R_{i,j}T^A_{i,j}/\tau) + \sum_n \sum_k exp(T^R_{i,j}T^A_{n,k}/\tau)}, A \in \{T, D\} \\ & n \in \{1, 2, \cdots, C\} \& n \neq i; k \in \{1, 2, \cdots, N_n\} \end{aligned} \tag{10}$$

Similarly, for the depth modality, circle contrastive learning not only carries out cross-modal bridge loss L^T_d to encourage the depth samples $T^D_{i,j}$ to resemble more closely the intermediary modality $T^T_{i,j}$, but also employs

cross-modal contrast loss L_d^R to bring the representations of $T_{i,j}^D$ closer to $T_{i,j}^R$. The circle contrastive learning for the depth modality can be uniformed denoted as $L_d^B, B \in \{T, R\}$, which is computed as follows:

$$L_d^B = -\log \frac{exp(T_{i,j}^D T_{i,j}^B/\tau)}{exp(T_{i,j}^D T_{i,j}^B/\tau) + \sum_n \sum_k exp(T_{i,j}^D T_{n,k}^B/\tau)}, B \in \{T, R\}$$
$$n \in \{1, 2, \cdots, C\} \& n \neq i; k \in \{1, 2, \cdots, N_n\} \tag{11}$$

To narrow the gap between the RGB and depth modalities after establishing their bridge with the intermediate modality, the circle contrastive learning L_{cm} is formulated as follows:

$$L_{cm}(e) = L_r^T(e) + L_d^T(e) + \frac{1}{1 + \mathbb{E}(L_r^T(e-1) + L_d^T(e-1))} \times (L_r^D(e) + L_d^R(e)) \tag{12}$$

Where e represents the current training epoch. $\mathbb{E}(.)$ refers to the average loss value from previous epochs. The coefficient associated with $(L_r^D(e) + L_d^R(e))$ is inversely proportional to the sum of $\mathbb{E}(L_r^T(e-1) + L_d^T(e-1))$. This deliberate relationship between the coefficients and the average loss values allows for a gradual increase in the constraint weights between the two original modals. As the features of the input modes become more similar to those of the intermediary mode, the intermediate relay is effectively established. Subsequently, the method incrementally enhances the emphasis on constraints pertaining to the original modes. By gradually increasing the proportion of these constraints, the process of reducing the modal gaps is smoothed out, resulting in a more efficient and effective alignment of the modalities. This progressive adjustment of the constraint coefficients aids in achieving a more balanced and coherent fusion of the original modes within the intermediate representation.

In addition to reducing the dissimilarity between positive cross-modal sample pairs, this loss function also contributes to diminishing the similarities between negative cross-modal sample pairs. By doing so, it reinforces the discriminative nature of the learned features, ultimately enhancing the overall performance of the model.

3.5.3 Base Loss. Base loss is comprised of triplet loss [13] and softmax loss [23]. Specifically, we input features $F_{i,j}^D$, $F_{i,j}^T$ and $F_{i,j}^R$ into the triplet loss, which first selects sample of one modality as an anchor, and chooses the positive and negative samples from features of the remaining two modalities. The distances between the positive sample pairs are denoted as d_p, and those of negative sample pairs are termed as d_n. The triple-modal triplet loss is recorded as L_t in this paper, which is calculated as follows:

$$L_t = [d_p - d_n + \alpha]_+ \tag{13}$$

α is the margin of the triplet loss, which is set to 0.3 and $[z]_+$ represents the function $max(z, 0)$.

Subsequently, we input $F_{i,j}^D$, $F_{i,j}^T$ and $F_{i,j}^R$ into batch normalization (BN) layer and fully connected (FC) layer to obtain features $B_{i,j}^D$, $B_{i,j}^T$ and $B_{i,j}^R$. These features are input into softmax loss function which is recorded as L_s in our method. The probabilities assigned to the class c, $c \in \{1, 2, \ldots, C\}$, for feature $B_{i,j}^a$, $a \in \{R, D, T\}$, are obtained as follows:

$$p(c|B_{i,j}^a) = \frac{e^{B_{i,j,c}^a}}{\sum_{k=1}^C e^{B_{i,j,k}^a}} \tag{14}$$

$B_{i,j,c}^a$ refers to the c^{th} channel of feature $B_{i,j}^a$. L_s is formulated as below:

$$L_s = -\sum_{c=1}^{C} \log(p(c))q(c) \tag{15}$$

Let g be the ground-truth, $q(g) = 1$. Otherwise, $q(n) = 0$ for $n \neq g$. In this case, minimizing the softmax loss is equivalent to maximizing the possibility of being assigned to the ground-truth class.

We record the sum of the L_t and L_s as L_b.

3.5.4 Identification Consistence Loss. In order to strengthen the multi-modal constraint, we input the integration features into the identification consistency loss. This loss adopts the widely used softmax pattern shown in Eq. 15, which is termed as L_{ic} in this paper. It drives the features that aggregate three modalities and their relations to meet with identification constraints. Therefore, it can effectively exploit the correlations among various modalities to optimize the features of each modality towards the direction of identification consistency. It favors promoting multi-modal correlations by reinforcing heterogeneous constraints, consequentially narrowing the modal discrepancy.

3.5.5 Overall Loss. The overall loss function of network is calculated as follows:

$$L = L_b + L_{t-rd} + L_{cm} + L_{ic} \tag{16}$$

4 EXPERIMENTS

4.1 Datasets and Evaluation Protocol

4.1.1 Dataset. **RobotPKU dataset:** This work [20] captures the RobotPKU dataset through Microsoft Kinect camera, which consists of 90 people with 16512 images in depth and RGB modalities. There exists a slight time delay between these two modalities. Some depth images in this dataset miss a part of body, which brings a greater challenge to cross-modal identification. Following the works [8, 36], we randomly sample 40 persons for training, 10 persons for validation and the remaining individuals for testing.

BIWI dataset: Work [26] utilizes the Microsoft Kinect camera to capture the long-term depth and RGB sequence pairs in BIWI dataset. Specifically, it is comprised of 78 individuals with 22038 images in two modals, which regards the same human with distinct clothes as a separate instance. We conduct the same partitions as methods [8, 36] for the fair evaluation on BIWI dataset. Namely, we randomly select 32 instances for training, 8 individuals for validation and 38 persons for testing.

4.1.2 Evaluation Metrics. In RGB-D cross-modal re-id task, there are two testing mechanisms: RGB-D and D-RGB. For RGB-D testing mode, query is given in RGB modal, gallery is comprised of depth images. For D-RGB testing pattern, query is given in depth modal while gallery consists of RGB images.

For these two mechanisms, the cumulative matching characteristic (CMC) curve and the mean average precision (mAP) are adopted for the performance evaluation. For each query, its average precision (AP) is calculated from the precision-recall curve, and mAP refers to the mean value of AP within all queries. This paper lists the cumulated matching result at selected Rank-n, $n \in \{1, 5\}$. The experiments are repeated 10 times to gain the average results for stable comparison.

4.2 Implementation Details

In RGB-D cross-modal person recognition task, it reduces the influence of complex background in the original image following method [36]. Specifically, we perform the detection algorithm [30] to detect pedestrians, and cut out the pedestrian regions accordingly. We discard the images without detected pedestrians. The proposed method uniformly resizes the input images to 288×144 size in two tasks.

The shallow networks all consist of the initial convolution layer of ResNet50 [12]. The deep network is composed of ResNet50 block1-4. The offline network is trained for 60 epochs. Each training batch samples 4 instances with 32 image pairs from two modalities. This paper applies stochastic gradient descent optimizer, which sets the weight_decay to 5e-4 and the momentum to 0.9. The setting of learning rate uses a warm-up learning strategy

with the initial learning rate of 0.1, which follows method [24]. The ResNet50 network is pre-trained on ImageNet dataset [5]. This paper trains the whole network on a single Nvidia GTX 1080 Ti GPU with Pytorch framework.

During the testing phase, we adopt the features outputted by shared deep extractor for cross-modal matching. The optimization process of the features extractors is guided by the decomposition and integration learning within multiple modalities, which are beneficial to strengthen modal correlation and narrow modal differences.

4.3 Comparisons with State-of-the-arts

Table 1. Comparisons with the state-of-the-art on RobotPKU dataset.

Method	RGB-D		D-RGB	
	Rank-1	mAP	Rank-1	mAP
LOMO+XQDA [18]	12.9	10.1	12.3	12.3
WHOS+XQDA [19]	10.0	8.2	9.8	9.8
Zero-padding network [35]	7.8±0.9	7.7±0.6	6.6±0.6	8.3±0.6
One-stream network [35]	11.9±0.6	11.4±0.5	12.5±1.0	14.2±1.4
Cross-modal distillation network [8]	17.5±2.2	17.1±1.9	19.5±2.0	19.8±2.1
Work [9]	25.3±2.0	23.5± 2.0	22.9±1.8	22.4±1.9
HRN [36]	23.1	17.3	25.7	23.5
IBN (Ours)	**31.8**	**25.1**	**29.1**	**24.9**

Results on RobotPKU dataset: Table 1 displays the comparison results between our method and state-of-the-arts on RGB-D RobotPKU dataset, which lists two testing cases. The comparison methods include: LOMO+XQDA [18], WHOS+XQDA [19], zero-padding network [35], one-stream network [35], cross-modal distillation network [8], work [9] and heterogeneous restraint network (HRN) [36]. Since only some methods begin to concentrate on this task, the number of listed compared methods is small.

For RGB-D testing mode, our method IBN can achieve 31.8% Rank-1 matching rate and 25.1% mAP, whilst 12.9% and 10.1% for LOMO+XQDA, 10.0% and 8.2% for WHOS+XQDA, 7.8% and 7.7% for zero-padding network, 11.9% and 11.4% for one-stream network, 17.5% and 17.1% for cross-modal distillation network, 25.3% and 23.5% for work [9], 23.1% and 17.3% for HRN. It can be observed that the proposed method realizes the advantages in both Rank-1 matching rate and mAP compared with these existing methods.

For D-RGB testing pattern, our method IBN surpasses all current approaches, which reaches the Rank-1 matching rate of 29.1% and mAP of 24.9%. It yields relative improvements of 3.4% at Rank-1 over the second-best method HRN. This indicates the effectiveness of our proposed approaches. IBN is capable of obtaining outstanding identification results in both testing cases, which illustrates that its application range is wide.

Results on BIWI dataset: We compare our approach with the state-of-the-arts in terms with RGB-D cross-modal recognition task on BIWI dataset, such as LOMO+XQDA [18], method [49], ICMDL [43], zero-padding network [35], one-stream network [35], cross-modal distillation network [8], work [9] and heterogeneous restraint network (HRN) [36]. The results of the two testing cases are shown in Table 2.

For RGB-D testing mode, our approach outperforms all listed methods. More precisely, IBN achieves 49.7% Rank-1 matching rate and 42.2% mAP. In contrast to the most relevant method HRN, IBN is superior to it by 5.8% and 11.3% at Rank-1 and mAP. All these experiments validate that the improvement components proposed by our method are beneficial to enhance the performance of cross-modal identification.

In terms of D-RGB testing case, our method can still achieve competitive performances. IBN obtains Rank-1 matching rate of 48.6% and mAP of 44.9%. Specifically, it raises the Rank-1 matching rate over the best compared

Table 2. Comparisons with the state-of-the-art on BIWI dataset.

Method	RGB-D		D-RGB	
	Rank-1	mAP	Rank-1	mAP
LOMO+XQDA [18]	13.7	12.9	16.3	15.9
Work [49]	12.1	-	11.3	-
ICMDL [43]	-	-	7.1	17.5
Zero-padding network [35]	5.9±2.2	7.3±4.0	10.3±2.7	9.8±3.8
One-stream network [35]	15.7±0.8	16.9±0.9	19.8±0.3	23.8±0.3
Cross-modal distillation network [8]	26.9±1.8	27.3±1.7	29.2±2.3	30.5±2.0
Work [9]	40.4±2.1	41.3± 1.8	42.8±3.9	43.9±3.9
HRN [36]	43.9	30.9	47.1	44.6
IBN (Ours)	**49.7**	**42.2**	**48.6**	**44.9**

method HRN by 1.5% on this dataset. The proposed method is capable to obtain a promising performance in various RGB-D cross-modal recognition benchmarks.

4.4 Ablation Studies

4.4.1 Parameter Validation. Multi-modal transformer embraces L blocks in series, and each block contains M parallel attention units. Therefore, L and M are two key hyperparameters in the proposed network. We carry out the person identification on the BIWI validation dataset to determine the parameters L and M. We initialize variables L and M with 3 and 4. First, we fix the value of M and linearly increase L from the initial value 3 to 7 with interval of 1. The results are shown in Fig. 4. It can be observed that the identification performance curve (mAP) first displays a growing trend with the increase of L. This is because deeper networks can embrace more powerful modeling capabilities. When L equals 6, the network performance reaches the best in both mAP and Rank-1. Therefore, L is set to 6 in subsequent experiments.

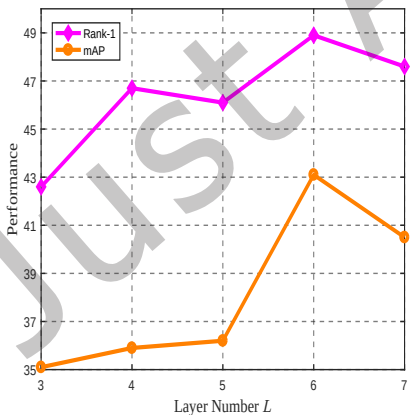

Fig. 4. Parameter validation of L on BIWI dataset.

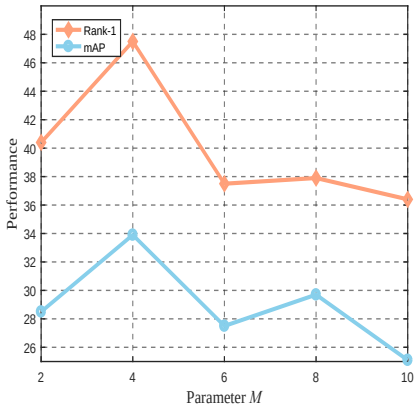

Fig. 5. Parameter validation of M on BIWI dataset.

In addition, we fix the value of L to 6, and linearly increase M from 2 to 6. The experimental results are shown in Fig. 5. It can be observed that the network performance has been improved with the increase of M. This is because

richer multi-modal relationships are obtained via larger network, which enhances recognition performance. When M reaches 4, a relatively superior result can be obtained. Then, with the increase of M, the recognition performance begins to drop. This is because the learned content is redundant with the expansion of the network. Therefore, M is set to 4 in this paper.

4.4.2 The Effectiveness of Intermediary Modal. In this paper, as the gap between two input modes is obvious, we introduce their intermediary mode for suppressing these differences. Although we have theoretically deduced its effectiveness, we still intend to testify its actual experimental effect in this part. To this end, we remove this mode from the proposed method to observe performance variation. Specifically, without the intermediary mode, the comparison approach becomes a dual-flow network. It replaces the triple-mode transformer with a dual-mode transformer, which is used to model the relationship between two original modes. Besides, there are L_r^D and L_d^R left in the cross-modal contrast learning. Namely, except the intermediary mode, the remaining parts of this comparison method keep unchanged for fair comparison.

We denote the comparison method as "w/o intermediary mode", and its experimental results are shown in Table 3. The method without intermediary mode deteriorates the performances of Rank-1 with absolute declines of 2.5% and 1.1% compared to our method with this modal on RobutPKU dataset for two testing mechanisms respectively. These results demonstrate the ability of intermediary modality in guiding the improvement of cross-modal person recognition. The intermediary mode builds a bridge for two original modes, which is more conducive to narrowing the differences between them, thus promoting the results of cross-modal identification.

To analyze the effectiveness of self-supervised intermediary learning, we evaluate the variant where this supervision is removed. The results on RobutPKU dataset are presented in Table 3. The method with this part tops over the variant with the absolute gains of 0.9% and 1.1% in terms of Rank-1 and mAP for RGB-D testing mechanism. This indicates that self-supervised contrastive learning offers a powerful constraint for the generation of intermediary mode, which helps promote the quality of the representations of intermediary mode.

Table 3. The effectiveness validation of the intermediary mode on RobutPKU dataset. "w/o" means "without".

Method	Testing Mode	Rank-1	Rank-5	mAP
w/o Intermediary mode	RGB-D	29.3	69.9	18.6
	D-RGB	28.0	58.3	17.9
w/o self-supervised intermediary learning	RGB-D	30.9	69.1	24.0
	D-RGB	28.6	59.1	24.1
IBN (Ours)	RGB-D	31.8	70.4	25.1
	D-RGB	29.1	60.7	24.9

4.4.3 The Effectiveness of Multi-modal Transformer. We remove the multi-modal transformer on RobutPKU dataset to analyze its effect on the visual cross-modal person re-id task. The remaining parts stay unchanged for fair comparison. The experimental results are displayed in Table 4. We observe that abandoning this module descends the Rank-1 by 1.2% and 1.5% in terms of two testing modes on RobutPKU dataset respectively, which emphasizes its role by contrary. This is because the multi-modal transformer can not only obtain the relationships within the same mode, but also model the relationship among different modes. In addition, it integrates the features of various modes, strengthening cross-modal association to enhance the performance of cross-modal person re-id.

4.4.4 The Effectiveness of Circle Contrastive Learning. In this paper, circle contrastive learning is exploited to pull heterogeneous modes closer. To demonstrate the contributions of this component, we conduct a comparative experiment by detaching it. The results on RobutPKU dataset are listed in Table 5. We find that the method

Table 4. The effectiveness validation of the multi-modal Transformer (denoted as multi-modal T) on RobutPKU dataset. "w/o" means "without".

Method	Testing Mode	Rank-1	Rank-5	mAP
w/o multi-modal T	RGB-D	30.6	70.2	21.3
	D-RGB	27.6	58.3	22.7
IBN (Ours)	RGB-D	31.8	70.4	25.1
	D-RGB	29.1	60.7	24.9

without the supervision of the circle contrastive learning degrades the performances of Rank-1 and mAP with the drops of 1.7% and 2.5% in contrast to our method with it in terms of RGB-D testing mode. Circle contrastive learning provides the intermediate relay for suppressing the gaps between two original modes. Therefore, this cross-modal learning can effectively reinforce the recognition accuracy of cross-modal networks.

Besides, we carry out the variant to validate the effect of cross-modal bridge learning by removing this component. The results on RobutPKU dataset are listed in Table 5. It can be seen that our approach with cross-modal bridge learning is superior to that of variant by 0.6% with regards to Rank-1. Cross-modal bridge learning produces the link effects for the reduction of modal gaps, which is beneficial to promote the performance of cross-modal recognition.

Table 5. The effectiveness validation of the circle contrastive learning on RobutPKU dataset. "w/o" means "without".

Method	Testing Mode	Rank-1	Rank-5	mAP
w/o circle contrastive learning	RGB-D	30.1	68.5	22.6
	D-RGB	27.5	58.3	23.9
w/o cross-modal bridge learning	RGB-D	31.2	69.9	24.8
	D-RGB	28.6	60.2	24.5
IBN (Ours)	RGB-D	31.8	70.4	25.1
	D-RGB	29.1	60.7	24.9

4.4.5 The Investigation of Usage of Mixup Ratio η. The proposed modal-mix mechanism determines the generated modality using a random mixup ratio, denoted as η. We aim to explore the impact of incorporating this random ratio into subsequent contrastive learning, observing how distinct usages of the mixup ratio affect the proposed approach. Specifically, in self-supervised intermediary learning and circle contrastive learning, we assign the random mixup ratio η (1-η) as the learning weight for contrastive learning between generated modal and RGB (depth) modal. This variant is denoted as "contrastive learning w/ mixup ratio". The results on RobutPKU dataset are presented in Table 6. It can be seen that our approach is superior to that of variant in terms of Rank-1 and mAP. In the majority of methods, the learning weights are either acquired through network training or manually configured based on validation experimental outcomes. In this experiment, the learning weights are randomly generated. Since the difficulty of contrastive learning cannot be accurately estimated through a random mixup ratio, incorporating it as the weight might introduce interference, thereby potentially weakening the identification performance.

4.4.6 The Investigation of Other Generation Methods. This paper adopts Mixup [42] to produce the generated modal, which is supervised by self-supervised intermediary learning. Mixup has inspired various variants, such as part-based Cutmix [41] and feature-level Mainfoid Mixup [32]. We substitute the Mixup opeation with these

Table 6. The investigation of usage of the mixup ratio η on RobutPKU dataset. "w/" means "with".

Method	Testing Mode	Rank-1	Rank-5	mAP
contrastive learning w/ mixup ratio	RGB-D	30.6	70.3	22.6
	D-RGB	28.6	60.1	23.4
IBN (Ours)	RGB-D	31.8	70.4	25.1
	D-RGB	29.1	60.7	24.9

typical variants, to explore the effects of different generation methods on the proposed approach. The results on RobutPKU dataset are listed in Table 7. It can be observed that the method with feature-level Manifold Mixup exhibits slightly improved rank-1 accuracy but lower performance in mAP compared with the method with Mixup. This suggests that the generation methods at the feature-level yield similar effects as those at the image-level in our method. This could be attributed to the fact that our method imposes self-supervised intermediary learning on the generated features. Besides, it can be seen that the method with part-based Cutmix significantly outperforms all other generation methods. Part-based generation methods may help provide more abundant generation samples to construct bridges for original modals, indicating potential avenues for further research in future work.

Table 7. Comparisons with other generation methods on RobutPKU dataset.

Method	RGB-D		D-RGB	
	Rank-1	mAP	Rank-1	mAP
Ours + Mixup [42]	31.8	25.1	29.1	24.9
Ours + Mainfoid Mixup [32]	32.1	23.7	28.8	24.8
Ours + Cutmix [41]	**37.4**	**27.0**	**33.4**	**26.9**

4.5 Visualization

4.5.1 Modal Difference Visualization. In order to further prove that the proposed method can decrease the modal differences between two original modes, we count the feature difference between these two modes in our method and *baseline* (omitting all proposed components).

Because cross-modal positive sample pairs can reflect the modal difference situation, we calculate the Euclidean distances between testing features of all positive pairs in these two methods. Here we enforce all feature vectors to be L2-normalized embeddings to ensure a fair comparison. The statistical histogram results of BIWI dataset are listed in Fig. 6. It can be seen from the figure that the distributions of feature differences obey the normal distribution in the two methods. The mean of feature difference in our method (pink histogram) is smaller than that in the *baseline*, so IBN successfully reduces the feature differences between two original modes. This is because the proposed method introduces the intermediate mode of two input modes. Besides, it exploits cross-modal contrast learning to pull the representations of two modes closer by building a bridge with the intermediate modal. Furthermore, this approach makes full use of the relationships among three modes, thus promoting cross-modal interactions.

4.5.2 Training Process Analysis. In this paper, the intermediate mode is produced to decrease the gaps between two input modalities. Compared with the existing methods that directly pull these two modalities closer, our method extra utilizes the intermediate mode as the bridge, which smooths the learning process of cross-modal identification network. In order to validate this advantage, we visualize the loss situations of "w/o Intermediary mode" and our method in Fig. 7 for comparison. It can be observed that our method with intermediate mode

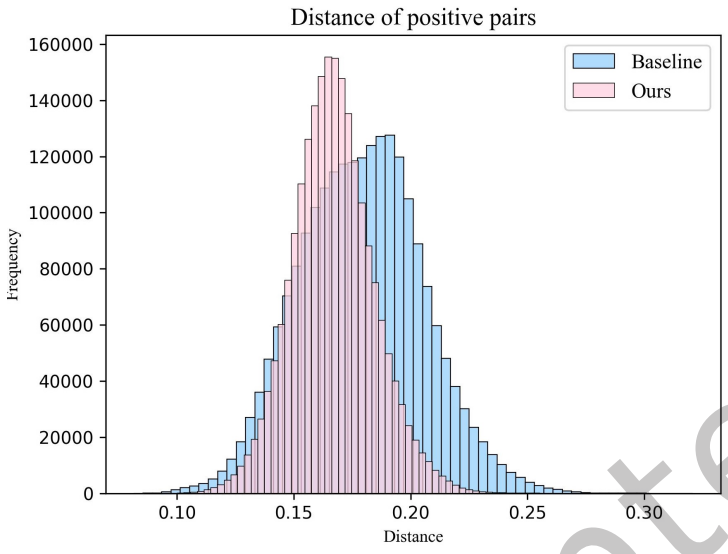

Fig. 6. Modal difference visualization. The distances of positive pairs in the proposed method are displayed in pink histogram, while those of *baseline* are shown in blue histogram. Best viewed in color.

can converge at the 40^{th} epoch, while the comparison network converges at the 50^{th} epoch. The decline slope of the loss function is larger in our method. So, our method is easier to converge, whose training process is more smooth. This is because the differences between the two modes are too large to directly decrease. In this paper, the intermediate mode of two original modes is constructed, which builds a bridge for narrowing the gaps between two modes. Therefore, the convergence process of the proposed network is more smooth.

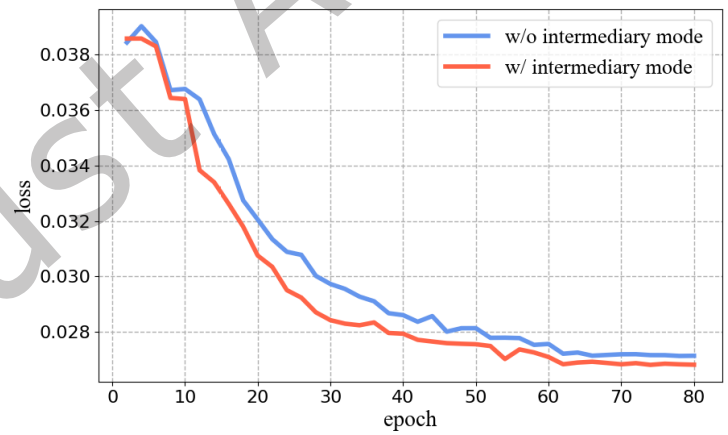

Fig. 7. Comparison of loss curves (after smooth) on BIWI dataset. The abscissa is the training epoch number and the ordinate is the loss value. "w/ Intermediary mode" and "w/o Intermediary mode" represent with and without the intermediary mode, resp.

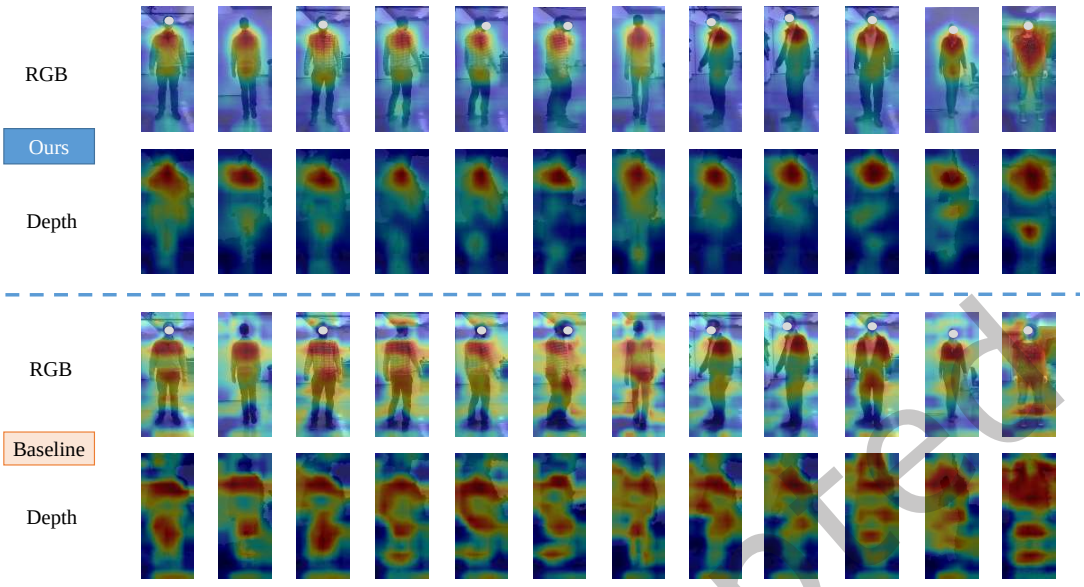

Fig. 8. Feature visualization of the proposed approach and the *baseline* method on the BIWI dataset.

4.5.3 Feature Visualization. In order to analyze the performance of the proposed method in depth, we visualize the characteristics of the proposed method and *baseline* method (omitting all proposed components). The images in the upper two lines display the feature visualization results of two modals in our method. The RGB and depth image pairs in the same column are taken at the same time. The images in the next two lines show the feature visualization results of the same image pairs in *baseline*. Through visualization, we can draw the following conclusions:

(1) Compared with *baseline*, the features of our method better depict the pedestrian contour. As we see, the pedestrian viewpoints are various in these images. The ratio information of head and neck owns superior robustness to the viewpoint variance, which is common in two input modes. Therefore, two modal features in our method and *baseline* mainly focus on the upper body of pedestrians, that is, the head-shoulder region. It can be seen that the focus area of our features is more consistent with the shape of the head-shoulder. The features of depth images in our method display the outline of pedestrians better compared with *baseline*. This is because the improvement components proposed in this paper are beneficial to learn these common features of the RGB and depth modalities, which can promote the discrimination of features.

(2) In contrast to *baseline*, the discrepancy between two modal features is smaller in the proposed method. Concretely, the features of two modalities pay attention to similar regions in our method, i.e., the upper body of person. The distributions of two features are more similar as well, that is, the feature weights mainly distribute in the upper body and a small part of weights locate in the lower body. However, there is a slight deviation for the focus regions of two features in *baseline*. Some RGB features focus on the upper body, while their corresponding depth features only concentrate on the whole body, such as the second column. The visualization results validate that the proposed method successfully narrows the gaps between two input modes. This is because this paper constructs an intermediary modality and fully exploits this modal to bridge the modal gaps.

4.6 Generalization Verification

As the proposed method generates an auxiliary mode without prior properties and module parameters, it can be applied to arbitrary visual cross-modal recognition tasks. To prove its generalization, this paper deploys it on the RGB-IR cross-modal re-id task. We choose a large RGB-IR dataset to realize the verification, i.e., SYSU-MM01 dataset [35]. This dataset is comprised of 395 instances with 22258 RGB images and 11909 infrared images 4 visible and 2 near-infrared cameras.

Table 8. Effeciency analysis of different components on SYSU-MM01 dataset in terms of all-search testing mode. MG denotes modal generation, MT refers to multi-modal transformer, and CCL refers to the circle contrastive learning. '√' symbol indicates that the corresponding element is adopted, and '-' denotes that the corresponding element is not included.

Method	MG	MT	CCL	Rank-1	Rank-10	mAP
baseline	-	-	-	60.3	89.5	57.6
model A	√	-	-	62.2	90.1	60.6
model B	√	√	-	67.5	92.1	63.7
IBN (ours)	√	√	√	70.8	96.1	67.1

Firstly, we carry out the ablation studies on SYSU-MM01 dataset. A *baseline* model is created, omitting all proposed components. Subsequently, the proposed components are gradually incorporated into the baseline model, resulting in the creation of three additional models: model A, model B, and model IBN. The results depicted in Table 8 demonstrate that the three key components can boost RGB-IR recognition performance. The role of multi-modal transformer is more significant in larger SYSU-MM01 dataset compared with RobutPKU dataset. This is because training data is more sufficient in SYSU-MM01 dataset. These experimental results illustrate that our method can be flexibly applied to various cross-modal pedestrian recognition, implyling its generalization.

Secondly, we compare the proposed method with the related Mixup-based methods, including MID [14] and Partmix variants [16]. They all adopt Mixup-based approaches to generate additional modalities. **Note that the proposed approach is different from these RGB-IR Mixup-based methods. Their differences and relations have been stated in Sec. 2.1.** We group these methods based on the mix pattern they adopt to ensure comparison fairness. The comparison results are displayed in Table 9. Regarding the Mixup pattern, it can be observed that the proposed "IBN+Mixup" outperforms previous methods with the same mix pattern. This superiority stems from leveraging self-supervised intermediary learning to enhance the quality of generated modalities. Besides, this paper extra puts forward a bridge network to adequately mine the bridge role of generated intermediary modal by capturing the correlations among distinct modalities. In addition, for the Mainfoid-based methods, the rank-1 performance of "IBN+Mainfoid Mixup" also demonstrates superiority. In terms of Cutmix-based approaches, "IBN+Cutmix" still offers significant performance gains, boosting the rank-1 accuracy of its *baseline* by 11.9%. However, the performance of "IBN+Cutmix" slightly lags behind that of the "Partmix Framework+Cutmix". This disparity arises because the proposed method is tailored for RGB-D tasks and is directly applied to the RGB-IR task. Moreover, it can be seen that the performances of IBN are less sensitivity to changes in mix patterns compared to the Partmix [16]. Namely, the performances of IBN vary slightly with various mix patterns. This robustness can be attributed to the additional constraints imposed on the generated modalities through self-supervised intermediary learning, ensuring the quality of the generated modalities.

5 CONCLUSION

This paper presents a novel RGB-D cross-modal person re-identification method that effectively utilizes transition information from the original two modalities without relying on networks or prior information. This approach is

Table 9. Comparisons with the most related methods on SYSU-MM01 dataset in terms of all-search testing mode.

Mix Pattern	Method	Rank-1	mAP
Mainfoid Mixup [32]	Partmix Framework [16]	71.3	67.7
	IBN (Ours)	71.5	67.8
Cutmix [41]	Partmix Framework [16]	73.4	70.7
	IBN (Ours)	72.2	68.6
Mixup [42]	Partmix Framework [16]	51.5	46.3
	MID [14]	60.3	59.4
	IBN (Ours)	70.8	67.1

simple yet effective. To fully leverage the capabilities of the intermediary mode in bridging modal discrepancies, this paper performs both decomposition and integration operations. On one hand, a multi-modal transformer is designed to integrate the information from the three modes by establishing their heterogeneous relations. This transformer applies an identification consistency constraint to enhance cross-modal associations, ensuring a more robust and reliable feature representation. On the other hand, the paper employs circle contrast learning to decompose the process of cross-modal constraints. This approach introduces an intermediate relay, suppressing modal gaps and enhancing the alignment between modalities. Extensive experiments conducted on various public datasets demonstrate that the proposed method surpasses state-of-the-art approaches in RGB-D cross-modal person re-identification. Through these experiments, the effectiveness of each component within the proposed method has been thoroughly evaluated and validated.

6 ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under grant 62302142, National Key Research and Development Programs of China under grants 2023YFC2506800 and 2019YFA0706203.

REFERENCES

[1] Emrah Basaran, Muhittin Gökmen, and Mustafa E Kamasak. 2020. An efficient framework for visible–infrared cross modality person re-identification. *Signal Processing: Image Communication* 87 (2020), 115933.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[3] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. 2020. 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*. 10257–10266.

[4] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-modality person re-identification with generative adversarial training.. In *IJCAI*, Vol. 1. 2.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[6] Javier Domínguez-Martín, María J Gómez-Silva, and Arturo De la Escalera. 2023. Neural Architectures for Feature Embedding in Person Re-Identification: A Comparative View. *ACM Transactions on Intelligent Systems and Technology* 14, 5 (2023), 1–21.

[7] Jiahao Gong, Sanyuan Zhao, Kin-Man Lam, Xin Gao, and Jianbing Shen. 2023. Spectrum-irrelevant fine-grained representation for visible–infrared person re-identification. *Computer Vision and Image Understanding* 232 (2023), 103703.

[8] Frank M Hafner, Amran Bhuiyan, Julian FP Kooij, and Eric Granger. 2019. RGB-depth cross-modal person re-identification. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–8.

[9] Frank M Hafner, Amran Bhuyian, Julian FP Kooij, and Eric Granger. 2022. Cross-modal distillation for RGB-depth person re-identification. *Computer Vision and Image Understanding* 216 (2022), 103352.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).
[14] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. 2022. Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1034–1042.
[15] Vijay John, Gwenn Englebienne, and Ben Krose. 2013. Person re-identification using height-based gait in colour depth camera. In *2013 IEEE International Conference on Image Processing*. IEEE, 3345–3349.
[16] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. 2023. PartMix: Regularization Strategy to Learn Part Discovery for Visible-Infrared Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18621–18632.
[17] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. 2020. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4610–4617.
[18] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2197–2206.
[19] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo. 2014. Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence* 37, 8 (2014), 1629–1642.
[20] Hong Liu, Liang Hu, and Liqian Ma. 2017. Online RGB-D person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology* 2, 1 (2017), 48–55.
[21] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. 2022. Learning Memory-Augmented Unidirectional Metrics for Cross-Modality Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19366–19375.
[22] Wenhe Liu, Xiaojun Chang, Ling Chen, Dinh Phung, Xiaoqin Zhang, Yi Yang, and Alexander G Hauptmann. 2020. Pair-based uncertainty and diversity promoting early active learning for person re-identification. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 2 (2020), 1–15.
[23] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks.. In *ICML*, Vol. 2. 7.
[24] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. 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*. 0–0.
[25] Andreas Møgelmose, Thomas B Moeslund, and Kamal Nasrollahi. 2013. Multimodal person re-identification using RGB-D sensors and a transient identification database. In *2013 International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–4.
[26] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. 2014. One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*. Springer, 161–181.
[27] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Springer, 69–84.
[28] Federico Pala, Riccardo Satta, Giorgio Fumera, and Fabio Roli. 2015. Multimodal person reidentification using RGB-D cameras. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 4 (2015), 788–799.
[29] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. 2020. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2149–2158.
[30] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).
[31] Jiajie Tian, Qihao Tang, Rui Li, Zhu Teng, Baopeng Zhang, and Jianping Fan. 2021. A camera identity-guided distribution consistency method for unsupervised multi-target domain person re-identification. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 4 (2021), 1–18.
[32] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*. PMLR, 6438–6447.
[33] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. 2019. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3623–3632.
[34] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. 2019. 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*. 618–626.
[35] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*. 5380–5389.

[36] Jingjing Wu, Jianguo Jiang, Meibin Qi, Cuiqun Chen, and Jingjing Zhang. 2022. An End-to-end Heterogeneous Restraint Network for RGB-D Cross-modal Person Re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 4 (2022), 1–22.
[37] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3733–3742.
[38] Xinxing Xu, Wen Li, and Dong Xu. 2015. Distance metric learning using privileged information for face verification and person re-identification. *IEEE transactions on neural networks and learning systems* 26, 12 (2015), 3150–3162.
[39] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 229–247.
[40] Mang Ye, Jianbing Shen, and Ling Shao. 2020. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security* 16 (2020), 728–739.
[41] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.
[42] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).
[43] Peng Zhang, Jingsong Xu, Qiang Wu, Yan Huang, and Jian Zhang. 2019. Top-push constrained modality-adaptive dictionary learning for cross-modality person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 12 (2019), 4554–4566.
[44] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 649–666.
[45] Yukang Zhang and Hanzi Wang. 2023. Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2153–2162.
[46] Yukang Zhang, Yan Yan, Jie Li, and Hanzi Wang. 2023. MRCN: A Novel Modality Restitution and Compensation Network for Visible-Infrared Person Re-identification. *arXiv preprint arXiv:2303.14626* (2023).
[47] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. 2021. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 788–796.
[48] Zhenghui Zhao, Rui Sun, Zi Yang, and Jun Gao. 2021. Visible-Infrared Person Re-Identification Based on Frequency-Domain Simulated Multispectral Modality for Dual-Mode Cameras. *IEEE Sensors Journal* 22, 1 (2021), 989–1002.
[49] Jiaxuan Zhuo, Junyong Zhu, Jianhuang Lai, and Xiaohua Xie. 2017. Person re-identification on heterogeneous camera network. In *CCF Chinese Conference on Computer Vision*. Springer, 280–291.

Received 10 November 2023; revised 22 March 2024; accepted 18 July 2024