2023 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3503-0718-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCV51070.2023.01027

Modality Unifying Network for Visible-Infrared Person Re-Identification

Hao Yu[1], Xu Cheng[1]*, Wei Peng[2], Weihao Liu[3], Guoying Zhao[4]
[1]School of Computer Science, Nanjing University of Information Science and Technology, China
[2]Department of Psychiatry and Behavioral Sciences, Stanford University, USA
[3]School of Computer Science and Technology, Soochow University, China
[4]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
{yuhao,xcheng}@nuist.edu.cn, wepeng@stanford.edu, whliu@stu.suda.edu.cn, guoying.zhao@oulu.fi

Abstract

Visible-infrared person re-identification (VI-ReID) is a challenging task due to large cross-modality discrepancies and intra-class variations. Existing methods mainly focus on learning modality-shared representations by embedding different modalities into the same feature space. As a result, the learned feature emphasizes the common patterns across modalities while suppressing modality-specific and identity-aware information that is valuable for Re-ID. To address these issues, we propose a novel Modality Unifying Network (MUN) to explore a robust auxiliary modality for VI-ReID. First, the auxiliary modality is generated by combining the proposed cross-modality learner and intra-modality learner, which can dynamically model the modality-specific and modality-shared representations to alleviate both cross-modality and intra-modality variations. Second, by aligning identity centres across the three modalities, an identity alignment loss function is proposed to discover the discriminative feature representations. Third, a modality alignment loss is introduced to consistently reduce the distribution distance of visible and infrared images by modality prototype modeling. Extensive experiments on multiple public datasets demonstrate that the proposed method surpasses the current state-of-the-art methods by a significant margin.

1. Introduction

Person re-identification (Re-ID) [8, 33] aims at matching pedestrian images captured from multiple non-overlapping cameras. Over the past few years, it has received increased attention due to its huge practical value in modern surveillance systems. Previous studies [10, 16, 19, 30, 40] mainly focus on matching pedestrian images captured from visible cameras and formulate the Re-ID task as a single-modality matching issue. Nevertheless, visible cameras may not provide accurate appearance information about persons in scenarios with poor illumination. To address this limitation, modern surveillance systems also employ infrared cameras, which can capture clear images in low-light conditions at night. As a result, visible-infrared person re-identification (VI-ReID) [1, 28, 29] has become a topic of growing interest in recent times, which seeks to match infrared images of the same identity when given a visible query across multiple camera views and vice versa.

VI-ReID is challenging due to the huge cross-modality discrepancy between visible and infrared images, as well as the intra-modality variation in person bodies such as pose variation and dress change. Existing methods [1, 20, 29, 31, 36, 37] primarily focus on relieving the cross-modality discrepancy by extracting modality-shared features to perform the feature-level alignment. Some studies [1, 20, 28, 31, 34] employ two-stream networks for cross-modality feature embedding, while others [3, 24, 25, 36] utilize Generative Adversarial Networks (GANs) to generate shared representations from visible and infrared images. However, these methods discard modality-specific features (such as colour and texture) that contain useful identity-aware patterns against intra-modality variations. Consequently, the learned features may not fully capture the variation of human bodies and thus lack discriminability. To address this limitation, the modality-unifying methods, *e.g.*, X-modality [9], DFM [7], SMCL [27], have been proposed to acquire the auxiliary modality by fusing visible and infrared modalities, encoding both modality-specific and modality-shared patterns to jointly relieve cross- and intra-modality discrepancies. In the SMCL [27], the authors proposed a syncretic modality generated by fusing visible and infrared pixels, which can bridge the gap between visible and infrared modalities while maintaining discriminability as the modality-specific information is preserved.

However, the existing modality-unifying works still have three weaknesses. (1) **Pixel fusion.** Previous works gener-

*Corresponding Author (Email: xcheng@nuist.edu.cn)

2380-7504/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCV51070.2023.01027

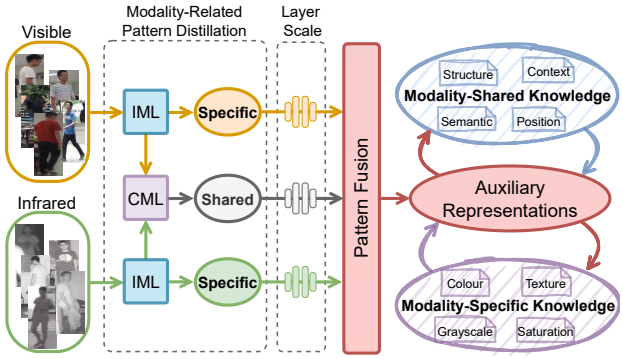

Figure 1. The main idea behind generating a strong auxiliary modality for the VI-ReID task. The IML and CML denote the intra-modality learner and cross-modality learner, respectively.

ate the auxiliary modality by fusing pixels of the raw visible and infrared images, which makes the richness of semantic patterns either equal to the original two modalities or lower in the case of pixel misalignment. In fact, the auxiliary modality is utilized to guide the learning of visible and infrared modalities, but the insufficient semantic patterns lead to a lack of identity-related information and severely limit the capacity to relieve intra-modality variations in VI-ReID.
(2) **Discrepancy bias.** During the VI training, the relative distances between visible and infrared images are constantly changing, which causes a dynamic bias toward the balance of intra- and cross-modality discrepancies. Thus, an ideal auxiliary modality should be able to dynamically control the ratio of modality-specific and modality-shared patterns it contains to model the evolving modality discrepancies. However, the existing studies simply use the global information of visible and infrared images to obtain the auxiliary representations, which are inflexible in adjusting the patterns they describe, leading to low robustness.
(3) **Inconsistency constraints.** Existing studies usually utilize features in the current batch to represent the overall distribution for distance optimization. However, this strategy suffers from randomness, as the training samples are different in each batch, which may cause a certain inconsistency in the learned feature relationships in different training stages, thus damaging the generalizability.

Inspired by the above discussions, we propose a novel Modality Unifying Network (MUN) to explore an effective and robust auxiliary modality for VI-ReID. The main idea of our auxiliary modality is illustrated in Figure 1. Specifically, we introduce an auxiliary generator comprising two intra-modality learners (IML) and one cross-modality learner (CML) to distil the modality-related patterns from visible and infrared images. Two IMLs are presented to identify the modality-specific and identity-aware patterns from visible and infrared images, respectively. They exploit multiple depth-wise convolutions with various kernel sizes to capture fine-grained semantic patterns in the human body across multiple receptive fields. Based on the outputs of two IMLs, the CML leverages spatial pyramid pooling to extract multi-scale feature representations and then fuse the modality-shared patterns learned in each feature scale. By combining IML and CML, the proposed auxiliary generator can generate a powerful auxiliary modality that is rich in modality-shared and discriminative patterns to alleviate both cross-modality and intra-modality discrepancies. In addition, the layer scale scheme is used to control the ratio of patterns learned from IML and CML, which can dynamically adjust the modality-specific and modality-shared patterns in the generated auxiliary representation.

Furthermore, to reveal the identity-related patterns in each identity set, an effective identity alignment loss (L_{ia}) is designed to optimize the distances of tri-modality identity centres. In addition, to regulate the distribution level feature relationships while relieving the inconsistency issue caused by sample variations, a novel modality alignment loss (L_{ma}) is designed to minimize the distances of three modalities, which utilizes the modality prototype to represent the learned modality information in each iteration.

In general, the major contributions of this paper can be summarized as follows.

- We propose a novel modality unifying network for the VI-ReID task by constructing a robust auxiliary modality, which contains rich semantic information from visible and infrared images to address modality discrepancies and reveal discriminative knowledge.
- A novel auxiliary generator constructed by the intra-modality and cross-modality learners is introduced to dynamically extract identity-aware and modality-shared patterns from heterogeneous images.
- The identity alignment loss and modality alignment loss are designed to jointly explore the generalized and discriminative feature relationships of the three modalities at both the identity and distribution levels.
- Extensive experiments on several public VI-ReID datasets verify the effectiveness of the proposed method and modality unifying scheme, which outperforms the current state of the arts by a large margin.

2. Related Work

Single-Modality Person Re-ID. Single-modality person re-identification [11, 17, 33] aims to match pedestrian images across different visible cameras. It presents challenges such as changes in viewpoint and human pose across camera views. Current approaches mainly focus on feature representation learning [15, 19, 39] and distance metric learning [10, 30, 35, 40]. Over the past few years, excellent performances have been achieved on several academic benchmarks. However, in practical scenarios, numerous crucial

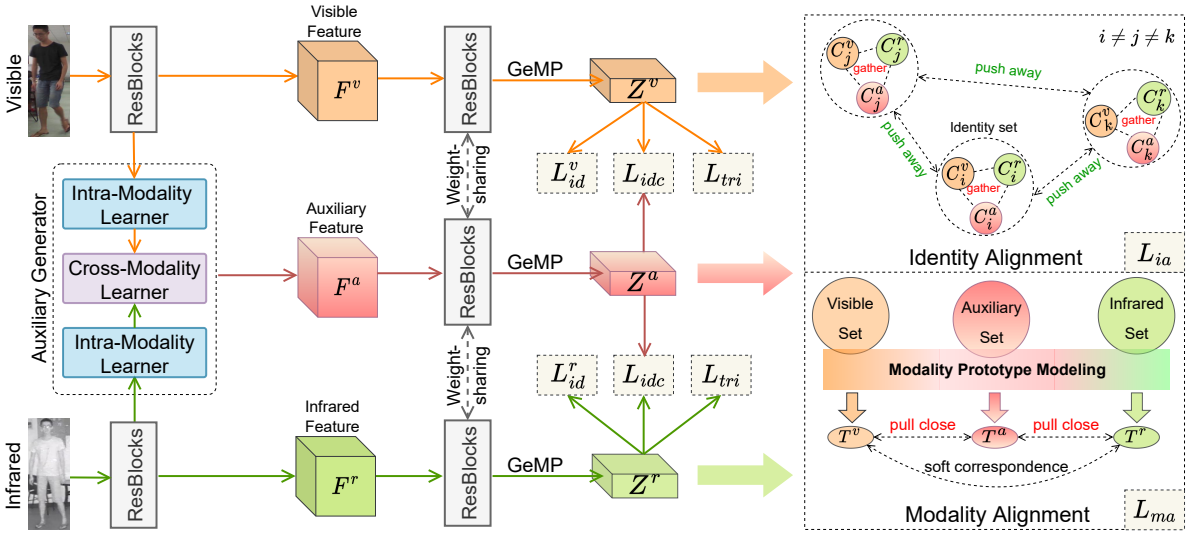

Figure 2. The overall architecture of the proposed MUN for VI-ReID. GeMP denotes the Generalized Mean Pooling [21]. The pretrained ResNet-50 [4] is adopted as the baseline network. To meet the specific requirements of VI-ReID, we initialize the first stage of the ResNet-50 twice as two independent ResBlocks to extract the low-level visible and infrared features, respectively. The remaining stages are utilized as modality-shared ResBlocks. During the inference, only visible and infrared modalities are utilized to perform cross-modality retrieval.

surveillance photos and videos are captured at night using infrared cameras. When it comes to matching pedestrians across visible and infrared modalities, the capabilities of these single-modality methods are limited due to their inability to address the huge modality gap. In contrast, we present an effective modality unifying network to bridge the modality gap and achieve precise cross-modality pedestrian matching in 24-hour monitoring scenarios.

Visible-Infrared Person Re-ID. Visible-Infrared person Re-ID [28] is a challenging task due to the cross-modality discrepancies between visible and infrared images, as well as the intra-modality variations such as pose and dress changes. Existing studies [20, 29, 31, 32, 34] mainly focus on learning the modality-shared representations to align the visible and infrared modalities. Some generation-based methods [24–26,36] have been developed to achieve modality alignment or translation by using Generative Adversarial Network (GAN). For instance, Wang *et al.* [24] proposed a dual-alignment network that used GAN to jointly learn pixel and feature level alignment. The D2RL [26] is proposed to perform image-level modality translation by adversarial training that relieves the cross-modality discrepancy. Other works [1, 20, 29, 31, 34] attempt to learn modality-shared features by designing two-stream networks to perform cross-modality feature embedding. Ye *et al.* [34] proposed a dual-constrained top-ranking method with a weight-shared two-stream network. Wu *et al.* [29] designed a cross-modality attention scheme to help the two-stream backbone discover cross-modality nuances. However, these methods usually discard modality-specific representations that help to relieve intra-modality variations, leading to low robustness and discriminability in learned features.

In order to capture both the modality-shared and identity-aware patterns from heterogeneous images, modality-unifying methods have been developed. These methods aim to obtain the auxiliary modality by combining modality-specific and modality-shared representations from both visible and infrared images. The syncretic modality [27] is proposed to guide the generation of discriminative and modality-invariant representations. The DFM [7] acquires the mixed modality by integrating visible and infrared pixels. However, these methods generate the auxiliary modality by directly fusing the raw pixels of visible and infrared images, making their auxiliary modality lack high-level semantic patterns and inflexible to adjust its representations.

To tackle these challenges, this paper presents the intra-modality learner and cross-modality learner to dynamically uncover substantial modality-shared and discriminative patterns from multiple receptive fields and feature scales. By integrating these learners, we introduce a powerful auxiliary modality that effectively bridges the modality discrepancy and enhances the discriminability of learned features.

3. Methodology

As shown in Figure 2, we introduce the details of the Modality Unifying Network. We first utilize two independent ResBlocks to extract low-level features from visible and infrared images, respectively. Then, the auxiliary generator is designed to generate the auxiliary features by com-

bining intra-modality and cross-modality learners. Afterwards, the visible, infrared, and auxiliary features are fed into weight-shared ResBlocks to learn high-level patterns. The auxiliary features can serve as a bridge to relieve both intra- and cross-modality discrepancies during the training. Based on the visible, infrared, and auxiliary features learned by weight-shared ResBlocks, four loss functions are developed to effectively improve cross-modality matching accuracy, including identity loss L_{id}, identity consistency loss L_{idc}, identity alignment loss L_{ia} and modality alignment loss L_{ma}.

3.1. Auxiliary Generator

The auxiliary generator contains two intra-modality learners (IML) and one cross-modality learner (CML). The two IMLs are designed to mine identity-related patterns from visible and infrared images, respectively. The CML is designed to learn modality-shared patterns based on the outcomes of two IMLs. The detailed architectures of IML and CML are shown in Figure 3.

Intra-Modality Learner. The intra-modality learner (IML) is designed to capture the discriminative and identity-aware patterns in human bodies. The visible or infrared low-level features $\mathbf{F}^m \in \mathbb{R}^{C\times H\times W}, m \in \{v, r\}$ extracted from two independent ResBlocks are regarded as the input of IML, where m denotes the visible or infrared modality.

To enrich the receptive field while keeping low computational complexity, we equally divide $\mathbf{F}^m$ into two parts along the channel dimension by matrix slice operation.

$$\mathbf{F}^m_{c_1} = \mathbf{F}^m[0 : C/2, :, :], \quad \mathbf{F}^m_{c_2} = \mathbf{F}^m[C/2 : C, :, :]. \tag{1}$$

Then, we employ 7×7 and 5×5 depth-wise convolutions (D) to operate on $\mathbf{F}^m_{c_1}$ and $\mathbf{F}^m_{c_2}$, respectively. This allows us to capture spatial patterns in different receptive fields.

$$\mathbf{R}^m = Concat\{D_{5\times5}(\mathbf{F}^m_{c_1}), D_{7\times7}(\mathbf{F}^m_{c_2})\}, \tag{2}$$

where $Concat$ denotes the concatenation on channel dimension; $\mathbf{R}^m$ indicates the visible or infrared features captured from multiple receptive fields. Then, a point-wise convolution (P) is utilized to fuse patterns with diverse receptive fields by connecting pixels in each channel.

$$\mathbf{R}^m_1 = P_{1\times1}(BatchNorm(\mathbf{R}^m)). \tag{3}$$

To integrate and encode the structural information in human bodies, another depth-wise convolution with 3×3 kernel size is introduced to remodel the learned spatial map. This layer also utilizes a residual branch to retain information from the previous layer.

$$\mathbf{R}^m_2 = D_{3\times3}(ReLU(\mathbf{R}^m_1)) + \mathbf{R}^m_1. \tag{4}$$

In addition, another point-wise convolution is utilized to fuse patterns with diverse receptive fields in $\mathbf{R}^m_2$.

$$\hat{\mathbf{F}}^m = I_{scale} * P_{1\times1}(BatchNrom(\mathbf{R}^m_2)), \tag{5}$$

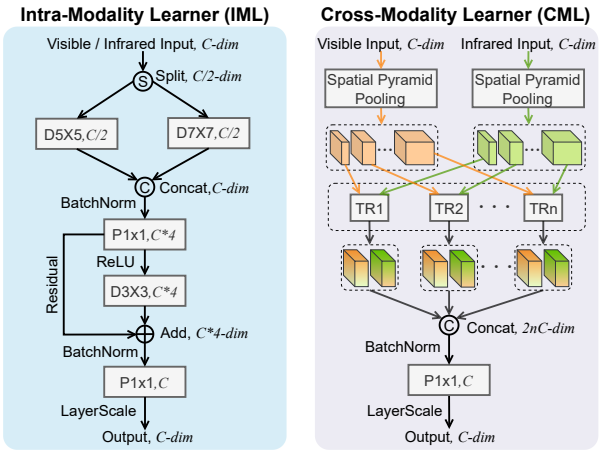

Figure 3. The detailed architecture of the proposed intra-modality learner (IML) and cross-modality learner (CML). They are designed to decouple the modeling of modality-related knowledge.

where $I_{scale} \in (0, 1]$ is the learnable layer scale factor used to control the ratio of intra-modality patterns learned by IML; $\hat{\mathbf{F}}^m$ denotes the outcomes of the two IMLs.

In the proposed intra-modality learner, three depth-wise convolutions with various kernel sizes are well combined to capture the identity-related patterns that existed in various receptive fields. Two point-wise convolutions are utilized for pattern integration and channel relation reasoning based on the inverted residual architecture [12]. The first point-wise convolution increases the channel dimension from C to $C * 4$ and the last point-wise convolution reduces the channel dimension from $C * 4$ to C.

Cross-Modality Learner. The cross-modality learner is designed to mine the modality-shared patterns from multiple feature scales based on the outcomes of two IMLs. Specifically, the spatial pyramid features are mined by applying n average pooling layers with various ratios.

$$\begin{aligned} \mathbf{S}^m_1 &= Avgpool_1(\hat{\mathbf{F}}^m) \quad , \\ \mathbf{S}^m_2 &= Avgpool_2(\hat{\mathbf{F}}^m) \quad , \\ &.... \quad , \\ \mathbf{S}^m_n &= Avgpool_n(\hat{\mathbf{F}}^m) \quad , \end{aligned} \tag{6}$$

where $\{\mathbf{S}^m_1, \mathbf{S}^m_2, ..., \mathbf{S}^m_n\}, m \in \{v, r\}$ denote the spatial pyramid features with various feature scales. Afterwards, we obtain the modality-shared spatial patterns from each pair of cross-modality spatial pyramid features $\{\mathbf{S}^v_i, \mathbf{S}^r_i\}^n_{i=1}$ with the same feature scale by using a group of learnable transposed convolutions $\{TR_1, TR_2, ..., TR_n\}$.

$$\hat{\mathbf{S}}^v_i = TR_i(\mathbf{S}^v_i), \quad \hat{\mathbf{S}}^r_i = TR_i(\mathbf{S}^r_i), \quad i = 1, 2, ..., n. \tag{7}$$

Here, the spatial dimensions of each cross-modality feature pair $\hat{\mathbf{S}}^v_i$ and $\hat{\mathbf{S}}^r_i$ are rebuilt to $H \times W$ by the corresponding transposed convolution TR_i, respectively. In this manner, the significant patterns of visible and infrared features

on each feature scale are embedded together, which helps to discover and amplify the abundant modality-shared information on multiple feature scales.

Further, all the embedded features are concatenated on the channel dimension via Eq.(8).

$$\hat{\mathbf{S}} = Concat\{\hat{\mathbf{S}}_i^v, \hat{\mathbf{S}}_2^v, ..., \hat{\mathbf{S}}_n^v, \hat{\mathbf{S}}_1^r, \hat{\mathbf{S}}_2^r, ..., \hat{\mathbf{S}}_n^r\}. \tag{8}$$

Then, the auxiliary feature is obtained by fusing patterns captured from multiple feature scales.

$$\mathbf{F}^a = C_{scale} * P_{1\times1}(BatchNomr(\hat{\mathbf{S}})), \tag{9}$$

where $\mathbf{F}^a$ denotes the auxiliary feature generated by our method. It contains substantial modality-shared and identity-aware information captured by three learners; $C_{scale} \in (0, 1]$ denotes the learnable layer scale factor used to control the ratio of modality-shared representations in the learned auxiliary feature $\mathbf{F}^a$; $P_{1\times1}$ is a point-wise convolution used to fuse patterns across different channels.

The CML mines significant patterns from multiple feature scales and amplifies the modality-shared parts from them using transposed convolutions. It makes our auxiliary feature a powerful tool to handle cross-modality variations.

3.2. Classification Constraint

To ensure the learned visible and infrared features are identity-related, the identity loss (L_{id}) implemented with the cross-entropy term is introduced as follows.

$$L_{id}^m = -\frac{1}{k}\sum_{i=1}^{k} logP(y_i|C_m(\mathbf{Z}_i^m)), \quad s.t. \quad m \in \{v, r\}, \tag{10}$$

where $\mathbf{Z}_i^v$ and $\mathbf{Z}_i^r$ denote the generalized mean pooled visible and infrared features in the i-th identity, respectively. k is the number of visible or infrared images in each batch; y_i is the *i*-th identity label; $C_v(\cdot)$ and $C_r(\cdot)$ are the predictions of visible and infrared classifiers, respectively.

The learned presentations are modality-shared if two classifiers can give consistent predictions for features from any modalities. However, if we directly apply features in one modality to the classifier in another modality (*e.g.*, $C_r(\mathbf{Z}^v)$), it may impose the classifier to learn modality-specific patterns rather than the modality-shared patterns, as the former is typically more discriminative. To solve this issue, we present an identity-consistency loss L_{idc} to update the parameters of both visible and infrared classifiers with the aid of auxiliary features. It can be defined as follows.

$$L_{idc} = -\frac{1}{k}\sum_{i=1}^{k}[logP(y_i|C_v(\mathbf{Z}_i^a)) + logP(y_i|C_r(\mathbf{Z}_i^a))], \tag{11}$$

where $\mathbf{Z}_i^a$ denotes the pooled auxiliary feature in the i-th identity. The auxiliary features effectively integrate visible and infrared patterns, facilitating the transfer of identity-related knowledge between modalities without compromising the original intra-modality learning.

3.3. Identity Alignment Loss

To relieve the class-level modality discrepancies and learn discriminative feature relationships, the identity alignment loss L_{ia} is designed to align the visible and infrared features of each identity with the aid of auxiliary features.

$$L_{ia} = \sum_{i=1}^{P}\Big[\alpha + \max_{m_1\in\{v,r\}} ||\mathbf{C}_i^a - \mathbf{C}_i^{m_1}||_2 - \min_{\substack{m_2\in\{v,r\}\\ q\neq i}} ||\mathbf{C}_i^a - \mathbf{C}_q^{m_2}||_2\Big], \tag{12}$$

where α is the margin parameter; P denotes the number of person identities; N is the number of images in the i-th identity; $\mathbf{C}_i^a = \frac{1}{N}\sum_{j=1}^{N}\mathbf{Z}_{i,j}^a$, $\mathbf{C}_i^v = \frac{1}{N}\sum_{j=1}^{N}\mathbf{Z}_{i,j}^v$, $\mathbf{C}_i^r = \frac{1}{N}\sum_{j=1}^{N}\mathbf{Z}_{i,j}^r$ are the auxiliary, visible, and infrared centres in the i-th identity, respectively; $\mathbf{Z}_{i,j}^v$, $\mathbf{Z}_{i,j}^r$ and $\mathbf{Z}_{i,j}^a$ denote the j-th visible, infrared, and auxiliary features in the i-th identity set.

In this paper, identity alignment loss is proposed to optimize the hardest cross-modality positive and negative centre pairs in a triplet-metric manner. It regulates discriminative and robust feature relationships by forcing all identities to form a tight intra-class space and pushing centres of different identities away across the three modalities.

3.4. Modality Alignment Loss

Previous works [7, 20, 29] typically align the two modalities by constraining visible and infrared features in each iteration. This scheme suffers from inconsistencies in the learned cross-modality feature relationships because the training samples are different in each iteration. To overcome this issue, we propose a modality alignment strategy that consistently aligns visible and infrared modalities by modeling the prototypes from features in each iteration.

Specifically, we first introduce three modality prototypes to represent the global information of visible, infrared, and auxiliary modalities, respectively. They can be denoted as $\mathbf{T}^v = \{\mathbf{t}_1^v, \mathbf{t}_2^v, ..., \mathbf{t}_B^v\}$, $\mathbf{T}^r = \{\mathbf{t}_1^r, \mathbf{t}_2^r, ..., \mathbf{t}_B^r\}$ and $\mathbf{T}^a = \{\mathbf{t}_1^a, \mathbf{t}_2^a, ..., \mathbf{t}_B^a\} \in \mathbb{R}^{B\times C}$, where $\mathbf{t}_i^v$, $\mathbf{t}_i^r$ and $\mathbf{t}_i^a$ are the modality prototypes for the i-th visible, infrared, and auxiliary features in each training batch (B), respectively.

The initial prototypes of the three modalities are obtained based on the pooled features $[\mathbf{Z}^v]^0$, $[\mathbf{Z}^r]^0$ and $[\mathbf{Z}^a]^0 \in R^{B\times C}$ in the 0-th iteration.

$$[\mathbf{T}^m]^0 = \mathbf{W}_p^m[\mathbf{Z}^m]^0, \quad s.t. \quad m \in \{v, r, a\}, \tag{13}$$

where $\mathbf{W}_p^m$ are learnable matrices to distil modality-related patterns from the *m*-th modality. $[\mathbf{T}^m]^0$ denotes the prototype of the *m*-th modality calculated in the 0-th iteration.

Further, to dynamically model the modality information during the training, we develop a temporal accumulation

strategy to update the modality prototype by the learned features in each iteration, which can be defined as follows.

$$[\mathbf{T}^m]^i = [\beta]^i * \mathbf{W}_p^m[\mathbf{Z}^m]^i + (1-[\beta]^i) * [\mathbf{T}^m]^{i-1}, \quad s.t. \quad m \in \{v, r, a\}, \tag{14}$$

where $[\mathbf{T}^m]^i$ is the m-th modality prototype calculated in the i-th iteration; β is the updating ratio, which gradually increases from $1e^{-8}$ to 1 with the training goes on. The temporal accumulation strategy ensures that the modality information in each iteration is considered, thereby synchronizing the cross-modality alignment during the training.

Based on the modality prototypes, the modality alignment loss (L_{ma}) is designed as:

$$L_{ma} = \frac{1}{P}\sum_{p=1}^{P}[mmd(\mathbf{T}_p^v, \mathbf{T}_p^a) + mmd(\mathbf{T}_p^a, \mathbf{T}_p^r)], \tag{15}$$

where $\mathbf{T}_p^v$, $\mathbf{T}_p^r$ and $\mathbf{T}_p^a$ denote visible, infrared, and auxiliary prototypes of the p-th identity, respectively. $mmd(\cdot,\cdot)$ is the MMD loss [5] implemented the modality level alignment by constraining the distance of modality prototypes in each iteration. In Equation 15, the $mmd(\mathbf{T}_p^v, \mathbf{T}_p^a)$ is defined as:

$$mmd(\mathbf{T}_p^v, \mathbf{T}_p^a) = ||\frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{T}_{p,i}^v) - \frac{1}{M}\sum_{j=1}^{M}\phi(\mathbf{T}_{p,j}^a)||_H^2, \tag{16}$$

where $\mathbf{T}_{p,i}^v$ and $\mathbf{T}_{p,j}^a$ denote the i-th visible prototype and the j-th auxiliary prototype in the p-th identity, respectively; $||\cdot||_H$ denotes the distribution measured by Gaussian kernel function $\phi(\cdot)$ which projects prototypes into the reproducing kernel Hilbert space. The $mmd(\mathbf{T}_p^a, \mathbf{T}_p^r)$ term can also be obtained in a similar way.

The modality alignment loss is used to constrain the identity-guided distribution distances of visible, infrared, and auxiliary modalities through the prototypes. It can effectively reduce the modality discrepancy and relieve the inconsistency issue in learned feature relationships. Meanwhile, the auxiliary modality can act as a bridge to decrease the relative distance between visible and infrared modalities in the common feature space, thereby significantly reducing the optimization difficulty of cross-modality alignment.

3.5. Overall Loss Function

Following the previous works, we employ the identity loss ($L_{id} = L_{id}^v + L_{id}^r$) and hard-mining triplet loss (L_{tri}) [26,32] as our baseline loss functions. The overall loss function of the proposed MUN can be summarized as:

$$L_{total} = L_{id} + L_{tri} + \gamma * L_{idc} + \theta * L_{ia} + \sigma * L_{ma}, \tag{17}$$

where γ, θ, and σ are parameters to balance the contribution of each proposed loss term during the training.

4. Experiments

4.1. Datasets and Evaluation Settings

SYSU-MM01 [28] is the largest dataset for VI-ReID, which comprises six cameras, including four visible and two infrared cameras. It encompasses a total of 491 individuals, with 287,628 visible images and 15,792 infrared images. The training set is composed of 395 individuals, with 22,258 visible images and 11,909 infrared images. The test set consists of 96 individuals, with 3,803 infrared images for queries and a gallery selected from 301 visible images. The dataset offers two testing settings: all-search mode and indoor-search mode. For both modes, we employ the hardest single-shot setting to perform the evaluation.

RegDB [18] comprises 412 identities and a total of 8,240 images, with 206 identities allocated for training and another 206 identities for testing. Each identity is represented by 10 visible and 10 infrared images. The testing phase in RegDB involves two modes: Visible to Infrared, where visible images are searched using an infrared image, and Infrared to Visible, which entails the reverse scenario. For both modes, we repeat the testing process 10 times and average the results to report the mean values.

Evaluation Settings. We utilize the standard cumulative matching characteristics (CMC) and mean average precision(mAP) as the evaluation metrics.

4.2. Implementation Details

We implement all the experiments on the PyTorch framework with an NVIDIA RTX-3090 GPU. To ensure repeatability and facilitate fair comparisons with existing methods, we adopt the pretrained ResNet-50 [4] as our backbone network, where the first stage is initialized twice as two modality-specific ResBlocks, and the rest stages are used as the modality-shared ResBlocks.

At the training stage, all images are resized to 288×144. Data augmentations, including random horizontal flipping, erasing, and channel augmentations [31], are utilized against overfitting. Our model is trained with the AdamW optimizer [13] for 90 epochs with a weight decay of 0.01. The learning rate is gradually increased from 10^{-8} to 0.002 in the first 15 epochs and then decays by 0.1 at the 30^{th} and 60^{th} epochs. We find the optimal settings of all the hyper-parameters by grid search and repeated ablation experiments. Specifically, the pooling ratios of the spatial pyramid pooling in CML are set to be $\{2, 4, 6, 12\}$; The margin parameter α is set to 0.55; the loss balance parameters γ, θ, and σ are set to 0.25, 0.5, and 0.008, respectively.

4.3. Ablation Study

We evaluate the effectiveness of each proposed component on SYSY-MM01 and RegDB datasets, as shown in Table 1. Compared with the baseline (B) which only learns

from visible and infrared modalities, the leveraging of auxiliary modality (Aux.) can effectively relieve both cross-modality and intra-modality discrepancies, thus greatly improving all the metrics on two datasets. Further, when applying the identity consistency loss (L_{idc}) to refine the modality-shared discriminative patterns, the performance is further improved. Meanwhile, the proposed identity alignment loss (L_{ia}) or modality alignment loss (L_{ma}) can enhance the cross-modality matching accuracy by aligning the visible and infrared features at the identity level or distribution level. By combining them, we can regulate a more robust cross-modality feature relationship, achieving a rank-1 of 76.24% and mAP of 73.81% on the SYSU-MM01 dataset. The results demonstrate that all the proposed components contribute consistently to the accuracy gain.

It is worth noting that when adding only the auxiliary modality to the baseline in Table 1, we employed a simple identity loss (like L_{id}^{a}) to supervise the auxiliary modality. This loss is deprecated when employing the proposed L_{idc} to jointly supervise the three modalities.

Table 1. Evaluation of different components of the proposed method on SYSU-MM01 and RegDB datasets. CMC (%) at rank 1 and mAP (%). The Red bold font and blue bold front denote the best and second best performances, respectively.

B	Aux.	L_{idc}	L_{ia}	L_{ma}	SYSU-MM01		RegDB	
					r=1	mAP	r=1	mAP
✓					57.49	55.83	75.42	71.93
✓	✓				62.55	58.42	79.26	73.81
✓	✓	✓			66.58	61.29	83.51	79.79
✓	✓	✓	✓		**71.35**	65.54	89.42	**84.66**
✓	✓	✓		✓	69.77	**66.96**	**89.93**	84.02
✓	✓	✓	✓	✓	**76.24**	**73.81**	**95.19**	**87.15**

Table 2. Performance of using different intermediate modalities in our MUN on two datasets. CMC (%) at rank 1 and mAP (%).

Intermediate Modality	SYSU-MM01		RegDB	
	r=1	mAP	r=1	mAP
X-modality [9]	66.17	63.06	79.95	74.28
Mixed modality [7]	66.42	62.85	79.43	73.09
Syncretic modality [27]	**72.95**	**68.74**	**84.59**	**79.11**
Auxiliary modality (Ours)	**76.24**	**73.81**	**95.19**	**87.15**

Effectiveness of auxiliary modality. We conduct the ablations in the MUN by replacing our auxiliary modality with intermediate modalities designed by previous works, namely X [9], mixed [7], and syncretic [27]. As shown in Table 2, the X-modality is generated from visible images only, ignoring the impact of infrared modality and achieving relatively low performances. Although syncretic and mixed modalities combine both visible and infrared patterns, they only utilize pixel-level information, lacking the ability to discover fine-grained and semantic patterns.

When switching back to our auxiliary modality, we observed a significant improvement in performance with a boost of 3.29%/5.07% in terms of rank-1/mAP on the SYSU-MM01 dataset. These results further prove that our auxiliary modality is superior to other intermediate modalities. In summary, our auxiliary modality can effectively integrate modality-related information from both visible and infrared images, while preserving strong discriminability. This promotes robust representation learning in VI-ReID.

Effectiveness of loss design schemes. In the modality alignment loss (L_{ia}), we design the modality prototype scheme to relieve the inconsistency issue and the auxiliary bridge scheme to reduce the optimization difficulty. To validate the effectiveness of these two schemes, we compare the performance of our method with or without using these two proposed schemes. As shown in Table 3.

Table 3. Performance comparison of with or without using the modality prototype and auxiliary (Aux.) bridge schemes in the modality alignment loss. CMC (%) at rank 1 and mAP (%).

Schemes in L_{ma}		SYSU-MM01		RegDB	
Prototype	Aux. bridge	r=1	mAP	r=1	mAP
		69.02	65.83	80.09	73.97
✓		**74.15**	**72.10**	**86.73**	**77.24**
	✓	71.66	68.35	83.92	75.54
✓	✓	**76.24**	**73.81**	**95.19**	**87.15**

In Table 3, it is evident that both the modality prototype scheme and the auxiliary bridge scheme contribute to the VI-ReID accuracy gain. The modality prototype consistently captures global modality information, enhancing the robustness of learned modality relationships by synchronizing the alignment in each iteration; The auxiliary bridge scheme reduces the relative distance between visible and infrared features, effectively alleviating the difficulty of cross-modality distance optimization.

4.4. Comparison with State-of-the-art Methods

In this section, we compare our MUN with state-of-the-art works on two public datasets, as shown in Table 4.

Comparison on SYSU-MM01 dataset. As illustrated in Table 4, the proposed MUN achieves impressive results with 76.24% rank-1 and 73.81% mAP on the all-search mode of the SYSU-MM01 dataset. Compared to traditional visible-infrared representation learning methods (AGW [33], DELN-ViT [38], SPOT [1], PMT [14], CMT [6], MPANet [29]), our MUN outperforms them by a margin of at least 4.36% rank-1 and 5.24% mAP on the all-search mode. The reason can be attributed that the proposed visible-auxiliary-infrared learning framework can capture more identity-related knowledge across modalities and regulate the discriminative feature relationships well. Additionally, the overall performance of our approach is also

Table 4. Comparison with state-of-the-art methods on SYSU-MM01 and RegDB datasets. CMC (%) at rank r and mAP (%).

Methods	Ref.	SYSU-MM01						RegDB					
		All-Search			Indoor-Search			Visible to Infrared			Infrared to Visible		
		r=1	r=10	mAP	r=1	r=10	mAP	r=1	r=10	mAP	r=1	r=10	mAP
Zero-padding [28]	ICCV17	14.80	54.12	15.95	20.58	68.38	26.92	17.75	34.21	18.90	16.63	34.68	17.82
JSIA-ReID [25]	AAAI20	38.10	80.70	36.90	43.80	86.20	52.90	48.50	–	49.30	48.10	–	48.90
AlignGAN [24]	ICCV19	42.40	85.00	40.70	45.90	87.60	54.30	57.90	–	53.60	56.30	–	53.40
AGW [33]	TPAMI21	47.50	84.39	47.65	54.17	91.14	62.97	70.05	86.21	67.64	70.49	87.21	65.90
X-Modality [9]	AAAI20	49.92	89.79	50.73	–	–	–	62.21	83.13	60.18	–	–	–
DFLN-ViT [38]	TMM22	59.84	92.49	57.70	62.13	94.83	69.03	92.10	97.97	82.11	91.21	**98.20**	81.62
SPOT [1]	TIP22	65.34	92.73	62.25	69.42	96.22	74.63	80.35	93.48	72.46	79.37	92.79	72.26
FMCNet [36]	CVPR22	66.34	–	62.51	68.15	–	63.82	89.12	–	84.43	88.23	–	83.86
SMCL [27]	ICCV21	67.39	92.87	61.78	68.84	96.55	75.56	83.93	–	79.83	83.05	–	78.57
PMT [14]	AAAI23	67.53	95.36	64.98	71.66	96.73	76.52	84.83	–	76.55	84.16	–	75.13
AGW+J [31]	ICCV21	69.88	95.71	66.89	76.26	97.88	80.37	85.03	95.49	79.14	84.75	95.33	77.82
MPANet [29]	CVPR21	70.58	96.10	68.24	76.74	98.21	**80.95**	83.70	–	80.90	82.80	–	80.70
CMT [6]	ECCV22	**71.88**	**96.45**	**68.57**	**76.90**	**97.68**	79.91	**95.17**	**98.82**	**87.30**	**91.97**	97.92	**84.46**
MUN (Ours)	ICCV23	**76.24**	**97.84**	**73.81**	**79.42**	**98.09**	**82.06**	**95.19**	**98.93**	**87.15**	**91.86**	**97.99**	**85.01**

superior to GAN-based methods (JSIA-ReID [25], AlignGAN [24], FMCNet [36]) thanks to the powerful auxiliary modality which dynamically combines the information of visible and infrared images without introducing extra noise.

Furthermore, the proposed method significantly outperforms existing modality-unifying methods (SMCL [27], X-Modality [9]) by at least 8.85% on the rank-1 metric. This can be attributed to the fact that we not only mine the fine-grained semantic representations to generate the auxiliary modality but also decouple the extraction of specific and shared patterns in two modalities, which contribute to the dynamic generation of the auxiliary modality for relieving the changeable modality discrepancies during the training.

Comparison on RegDB dataset. The results on RegDB are also listed in Table 4. In this dataset, image samples are spatially aligned and present less intra-class variations. Thus, the accuracy of all methods is higher than that on SYSU-MM01. The proposed MUN achieves Rank-1 of 95.19% and mAP of 87.15% in visible to infrared mode. Similar improvement can also be observed in the infrared to visible mode, which shows that our method obtains Rank-1 of 91.86% and mAP of 85.01%. This improvement can be attributed to the capacity of our method to generate a robust auxiliary modality, effectively mitigating both cross-modality and intra-modality discrepancies.

4.5. Evaluation on Generalizability

To verify the generalizability of MUN, we conduct experiments on two corrupted VI Re-ID datasets [2], namely SYSU-MM01-C and RegDB-C. We utilize the same corruption settings as [2], which only performs corruptions during the testing stage and randomly selects one corruption type (*e.g.*, elastic, snow, frosted glass, etc.) and one severity level for each image in the visible gallery set. The results reported for all the compared methods are obtained using the official best settings as provided in their respective papers.

Table 5. Evaluations on corrupted datasets. Each evaluation is performed 10 times to obtain the mean value. L_{id}^{a} denotes an individual identity loss used to supervise the auxiliary modality.

Index	Method	SYSU-MM01-C		RegDB-C	
		r=1	mAP	r=1	mAP
1	X-Modality [9]	31.98	26.20	37.26	35.97
2	CIL [2]	36.95	35.92	**52.25**	**49.76**
3	SMCL [27]	37.08	36.12	51.93	49.22
4	AGW+J [31]	**40.09**	**37.86**	51.53	49.04
5	B	25.92	23.13	32.05	29.64
6	+Aux.+L_{id}^{a}	30.85	26.54	36.40	31.21
7	+Aux.+L_{idc}	31.12	27.44	38.13	32.29
8	+Aux.+L_{idc}+L_{ia}	36.78	32.32	45.44	42.89
9	+Aux.+L_{idc}+L_{ia}+L_{ma}	**41.17**	**38.63**	**52.69**	**50.18**

Table 5 shows that the performance of baseline (B) is relatively lower than that of existing SOTAs under data corruption scenarios. However, by introducing the auxiliary modality to bridge the gap between visible and infrared modalities (Index 6), the accuracy is significantly improved. Furthermore, by incorporating the proposed identity alignment loss (L_{ia}) and modality alignment loss (L_{ma}) to refine the learned cross-modality feature relationships, the rank-1 and mAP accuracies experience a substantial enhancement of 15.25% and 15.5% respectively on the SYSU-MM01-C dataset. This outperforms the current SOTAs by a remarkable margin. Specifically, our MUN method exceeds the highly robust AGW+J [31] by 1.08% in rank-1 and 0.77% in mAP. The experiments on two corrupted datasets verify the strong generalizability and robustness of the proposed MUN, which can consistently learn modality-shared patterns and regulate stable feature relationships under corrupted data scenarios.

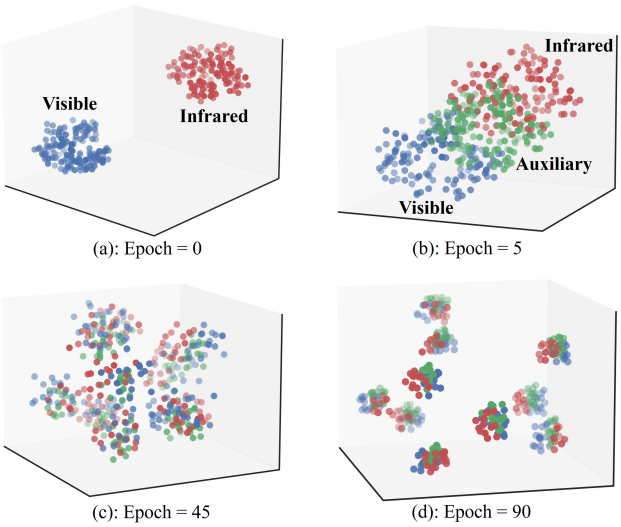

Figure 4. Distributions of learned visible, infrared and auxiliary features during the training. All the samples are randomly selected in the SYSU-MM01 dataset. The 3D visualization is made by T-SNE [23]. Please view in colour and zoom in.

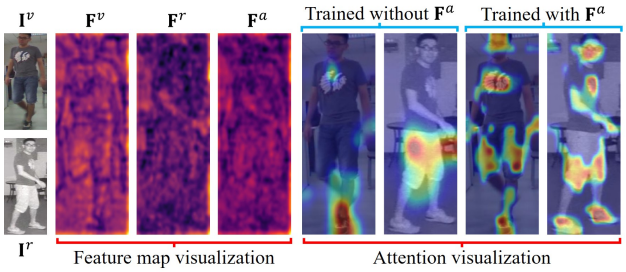

Figure 5. Visualization of learned feature maps and attention maps. Given the input visible image $\mathbf{I}^v$ and infrared image $\mathbf{I}^r$, we visualize the corresponding learned visible feature $\mathbf{F}^v$, infrared feature $\mathbf{F}^r$, and the generated auxiliary feature $\mathbf{F}^a$. It is obvious that the auxiliary feature $\mathbf{F}^a$ can preserve most of the modality-shared patterns, including body shape and structure. The attention visualizations on the input images are made by Grad-CAM [22]. Please view in colour and zoom in.

4.6. Visualization

Distribution Visualization. To visually demonstrate the effectiveness of MUN, we randomly select 10 identities from the SYSU-MM01 dataset and visualize their feature distributions during the training. The visualization results are presented in Figure 4.

At the onset of training (epoch 0 in Figure 4 (a)), significant modality disparities arise between visible (depicted by blue dots) and infrared images (depicted by red dots), rendering cross-modality matching unfeasible. As the training progresses, the proposed auxiliary modality (depicted by green dots) serves as a bridge to connect the visible and infrared modalities in the common feature space. We can observe that the learned visible and infrared features show a convergence trend, and the modality discrepancies are gradually eliminated at epoch 5 in Figure 4 (b). Afterwards, the network learns identity-aware patterns by regulating smaller intra-class distances and larger inter-class distances at epoch 45. Finally, in Figure 4 (d), all the learned features are well grouped into their respective identity centers, demonstrating powerful discriminability under cross-modality scenes, which proves the effectiveness of MUN in learning robustness and identity-aware features.

Pattern Visualization. In order to further illustrate the effectiveness of our auxiliary modality, we select one visible and one infrared image with the same identity in the SYSU-MM01 dataset to visualize the corresponding learned feature maps and attention maps via Grad-CAM [22].

The visualization results are shown in Figure 5. It is clear that the visible ($\mathbf{F}^v$) and infrared ($\mathbf{F}^r$) feature maps extracted from the backbone have different patterns and structures, which makes it difficult to perform cross-modality matching. Notably, the proposed auxiliary generator can dynamically reconstruct and align the modality-shared spatial patterns using multiple transposed convolutions. This makes our auxiliary feature ($\mathbf{F}^a$ in Figure 5) preserves most of the shared patterns between $\mathbf{F}^v$ and $\mathbf{F}^r$ without corrupting the structure information of person bodies. In addition, the attention visualization in Figure 5 further indicates that the proposed auxiliary modality plays a critical role in helping the network learn modality-shared representations.

5. Conclusion

This paper proposes a novel Modality Unifying Network to jointly explore the robust auxiliary modality and generalize cross-modality feature relationships for VI-ReID. The auxiliary modality is generated by combining the cross-modality learner and the intra-modality learner, enabling the dynamic extraction of modality-specific patterns from multiple receptive fields and feature scales. This approach empowers our auxiliary modality to effectively alleviate both cross-modality and intra-modality discrepancies. Moreover, we propose identity alignment loss and modality alignment loss to regulate discriminative feature relationships in multi-modality tasks. Extensive experiments on public datasets demonstrate the effectiveness and generalizability of our MUN as well as each proposed component.

Acknowledgements. This work is supported by the Academy of Finland for Academy Professor project EmotionAI (Grants No. 336116, 345122), the University of Oulu & The Academy of Finland Profi 7 (Grant No. 352788), and the National Natural Science Foundation of China (Grant No. 61802058, 61911530397). We appreciate the professional and cost-effective GPU computing service provided by www.AutoDL.com.

References

[1] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31:2352–2364, 2022.

[2] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification. *arXiv preprint arXiv:2111.00880*, 2021.

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

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

[5] Chaitra Jambigi, Ruchit Rawal, and Anirban Chakraborty. Mmd-reid: A simple but effective solution for visible-thermal person reid. *arXiv preprint arXiv:2111.05059*, 2021.

[6] Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 480–496. Springer, 2022.

[7] Jun Kong, Qibin He, Min Jiang, and Tianshan Liu. Dynamic center aggregation loss with mixed modality for visible-infrared person re-identification. *IEEE Signal Processing Letters*, 28:2003–2007, 2021.

[8] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2019.

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

[10] Shengcai Liao and Ling Shao. Graph sampling based deep metric learning for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7359–7368, 2022.

[11] Minjie Liu, Jiaqi Zhao, Yong Zhou, Hancheng Zhu, Rui Yao, and Ying Chen. Survey for person re-identification based on coarse-to-fine feature learning. *Multimedia Tools and Applications*, pages 1–35, 2022.

[12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[13] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[14] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1835–1843, 2023.

[15] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 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*, pages 0–0, 2019.

[16] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019.

[17] Neha Mathur, Shruti Mathur, Divya Mathur, and Pankaj Dadheech. A brief survey of deep learning techniques for person re-identification. In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pages 129–138. IEEE, 2020.

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

[19] Xin Ning, Ke Gong, Weijun Li, Liping Zhang, Xiao Bai, and Shengwei Tian. Feature refinement and filter network for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3391–3402, 2020.

[20] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12046–12055, 2021.

[21] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[23] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[24] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3623–3632, 2019.

[25] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12144–12151, 2020.

[26] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019.

[27] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 225–234, 2021.

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

[29] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021.

[30] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Xin Ning, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Transactions on Multimedia*, 24:1665–1677, 2021.

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

[32] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 229–247. Springer, 2020.

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

[34] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, page 2, 2018.

[35] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13657–13665, 2020.

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

[37] Sen Zhang, Zhaowei Shang, Mingliang Zhou, Yingxin Wang, and Guoliang Sun. Cross-modal identity correlation mining for visible-thermal person re-identification. *Multimedia Tools and Applications*, pages 1–14, 2022.

[38] Jiaqi Zhao, Hanzheng Wang, Yong Zhou, Rui Yao, Silin Chen, and Abdulmotaleb El Saddik. Spatial-channel enhanced transformer for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 2022.

[39] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147, 2019.

[40] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Xing Sun, and Weishi Zheng. Aware loss with angular regularization for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13114–13121, 2020.