Base-Detail Feature Learning Framework for Visible-Infrared Person Re-Identification

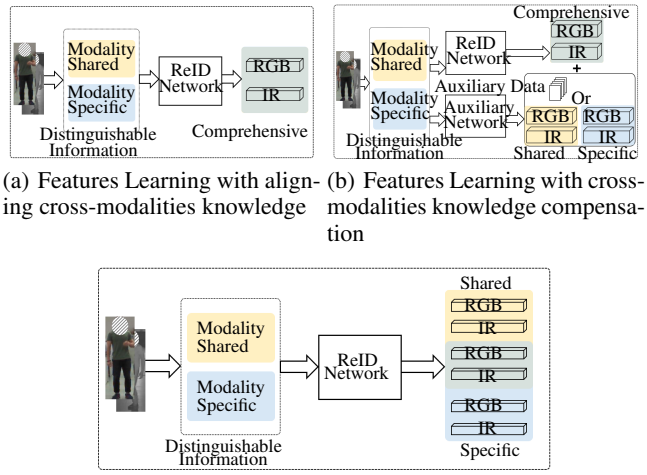
Zhihao Gong[1] , **Lian Wu**[2] , **Yong Xu**[1,*]
[1]Harbin Institute of Technology (Shenzhen)
[2]GuiZhou Education University
gongzhh888@gmail.com, wulian_best@163.com, laterfall@hit.edu.cn

Abstract

Visible-infrared person re-identification (VIReID) provides a solution for ReID tasks in 24-hour scenarios; however, significant challenges persist in achieving satisfactory performance due to the substantial discrepancies between visible (VIS) and infrared (IR) modalities. Existing methods inadequately leverage information from different modalities, primarily focusing on digging distinguishing features from modality-shared information while neglecting modality-specific details. To fully utilize differentiated minutiae, we propose a Base-Detail Feature Learning Framework (BDLF) that enhances the learning of both base and detail knowledge, thereby capitalizing on both modality-shared and modality-specific information. Specifically, the proposed BDLF mines detail and base features through a lossless detail feature extraction module and a complementary base embedding generation mechanism, respectively, supported by a novel correlation restriction method that ensures the features gained by BDLF enrich both detail and base knowledge across VIS and IR features. Comprehensive experiments conducted on the SYSU-MM01, RegDB, and LLCM datasets validate the effectiveness of BDLF.

1 Introduction

Person re-identification (ReID) aims to retrieve a target identity from gallery images captured by different cameras [Liu *et al.*, 2022] and has recently demonstrated significant advancements in the fields of security and public surveillance [Ye *et al.*, 2022a]. However, most existing methods [Cao *et al.*, 2023][Wang *et al.*, 2022][Yan *et al.*, 2021] primarily focus on utilizing RGB images captured by visible (VIS) cameras during the daytime, which are inadequate for accommodating 24-hour scenarios that involve infrared (IR) images captured by IR cameras. To address the substantial cross-modality gap and facilitate operation in all-day scenarios, visible-infrared person re-identification (VIReID) methods [Chen *et al.*, 2022][Park *et al.*, 2021] have been developed, enabling the matching of IR (RGB) images given an interest in a specific RGB (IR) pedestrian image.



(a) Features Learning with aligning cross-modalities knowledge (b) Features Learning with cross-modalities knowledge compensation

(c) Feature Learning with the proposed BDLF

Figure 1: Motivation of the proposed BDLF, which focuses on sufficiently mining the modality-shared and modality-specific knowledge simultaneously and are not applicable for additional auxiliary data.

The existing research on VIReID can generally be categorized into two principal methods: extracting distinguishing modality-shared features from VIS and IR modalities[Park *et al.*, 2021][Zhang and Wang, 2023] and compensating for modality-specific or modality-shared features [Zhang *et al.*, 2022a]. As shown in Figure 1(a), the former method aims to reduce cross-modality discrepancies by aligning comprehensive cross-modality features into a common semantic space. However, it neglects to leverage modality-specific and shared cues, which inevitably leads to performance bottlenecks. The latter approach, depicted in Figure 1(b) can be further divided into embedding-level and image-level methods. These methods generate compensatory knowledge in the embedding space and at the pixel level respectively, using auxiliary models(e.g., GANs[Goodfellow *et al.*, 2014], segmentation networks, part alignment networks, etc.). However, these methods typically introduce losses and noise into the generated features or require additional data processing by other models, making them less effective and convenient. Consequently, advancing the development of VIReID to a more

comprehensive level remains a significant challenge.

Inspired by the analyses presented above, it is essential to recognize that modality-shared information, such as the contour and movement characteristics of pedestrians, can be considered base features. In contrast, modality-specific information, including the color and texture details of the RGB modality and the thermal characteristics of the IR modality, can be regarded as detail features. Both types of them should be integrated and utilized effectively together. Therefore, in this paper, we propose a novel Base-Detail Feature Learning Framework (BDLF), as shown in Figure 1(c). This framework is designed to extract modality-shared base features and modality-specific detail features from the original images with minimal additional computational costs, while jointly optimizing modality-shared, modality-specific, and comprehensive features.

The proposed BDLF comprises a modality-specific detail feature extraction (DFE) module and a modality-shared base embedding generation (BEG) block, which ultimately combine the optimized features collected. Inspired by [Zhao *et al.*, 2023], we designed the DFE module to mine the modality-specific detail information losslessly. Subsequently, the BEG block derives modality-shared base features. To fully capture both specific and shared information, we proposed a novel specific-shared knowledge distillation(SKD) loss. It encourages the detail (base) features to effectively incorporate modality-specific (modality-shared) knowledge by imposing a constraint on the correlation that the cross-modality detail and base features should exhibit. Specifically, it ensures that the correlations across RGB and IR modalities are indistinct and notable, respectively. Perspectives in [Feng *et al.*, 2023] explain that the independent decomposition of features can maximize the mutual information of sub-features; therefore, we introduced an independence constraint in the semantic space between the derived detail and base features. This indicates that the base feature exclusively encompasses modality-shared knowledge, while the detail feature contains modality-specific information. In summary, the main contributions of our work are as follows:

- A novel correlation optimization method is proposed that effectively generates both modality-shared and modality-specific features using a non-parametric approach, rather than relying on classifiers.
- We propose an end-to-end Base-Detail Feature Learning Framework (BDLF) for VIReID that integrates extracts of modality-shared base knowledge and modality-specific detail knowledge.
- Extensive experiments have demonstrated that the proposed BDLF outperforms other state-of-the-art methods for the VIReID task on the SYSU-MM01, RegDB, and LLCM datasets.

2 Related Work

The main idea for solution VI-ReID task is decreasing the notable discrepence across VIS and IR modalities, thereby the existing methods consist of aligning the cross-modality features and utilizing the auxiliary data or features generated by other models.

The alignment of feature representation methods seeks to convert cross-modality features into a unified semantic space through either metric learning techniques [Liu *et al.*, 2022] [Park *et al.*, 2021] [Luo *et al.*, 2019] or by enhancing networks with more effective feature extraction components [Zhang and Wang, 2023] [Sarker and Zhao, 2024]. However, these approaches ultimately encounter performance bottlenecks due to the loss of modality-specific information.

The methods for utilizing auxiliary information produced by other models are proposed to enhance identifiable knowledge. GAN-based methods [Zhang *et al.*, 2022a]d[Wang *et al.*, 2020] generate compensatory features at either the image level or the embedding level to simulate features from another modality. XIV [Li *et al.*, 2020] introduces the X-modality generated by a lightweight auxiliary network to decrease discrepancies between the two modalities. LUPI [Alehdaghi *et al.*, 2022] establishes an intermediate domain between VIS and IR modalities. Furthermore, it generates images that belong to this intermediate domain to guide the network in acquiring more discernible information. SGIEL [Feng *et al.*, 2023] innovatively adopts the shape knowledge of identity generated by segmentation models to enrich supplementary information. TMD [Lu *et al.*, 2024] generates style-aligned images to minimize differences at the image level, subsequently aligning cross-modality features to eliminate discrepancies in feature distribution and instance features. However, this remains a challenging field of research because these methods either inevitably introduce information distortion during the generation process or fail to completely capture modality-specific and modality-shared information.

3 Methodology

3.1 Overall Framework

The pipeline of our proposed method, referred to as BDLF, is illustrated in Figure 2. This method utilizes a single-stream ResNet-50 network[He *et al.*, 2016a] as its backbone. The intermediate features Z^M, which pass through a portion of the backbone, are fed into the proposed detail feature extraction (DFE) module to yield detail features Z^D. Additionally, the base feature Z^B is generated by inputting the output Z from the backbone into the proposed base embedding generation (BEG) block. A novel specific-shared knowledge distillation (SKD) loss is proposed to ensure that the generated detail(base) features contain as much modality-specific (modality-shared) knowledge as possible, thereby effectively leveraging modality-specific and shared information. Furthermore, we construct a modality-shared feature Z^F using a cross-modality feature fusion method to optimally supplement the base features. During the inference phase, only the comprehensive feature Z yielded by the backbone is used for performance evaluation. This is because the proposed DFE and BEG modules effectively enhance the comprehensive feature by incorporating additional detail and base information.

Given an identity image from either the visible or infrared modality, VIReID intends to identify the most similar sequence of that identity in another modality. Let

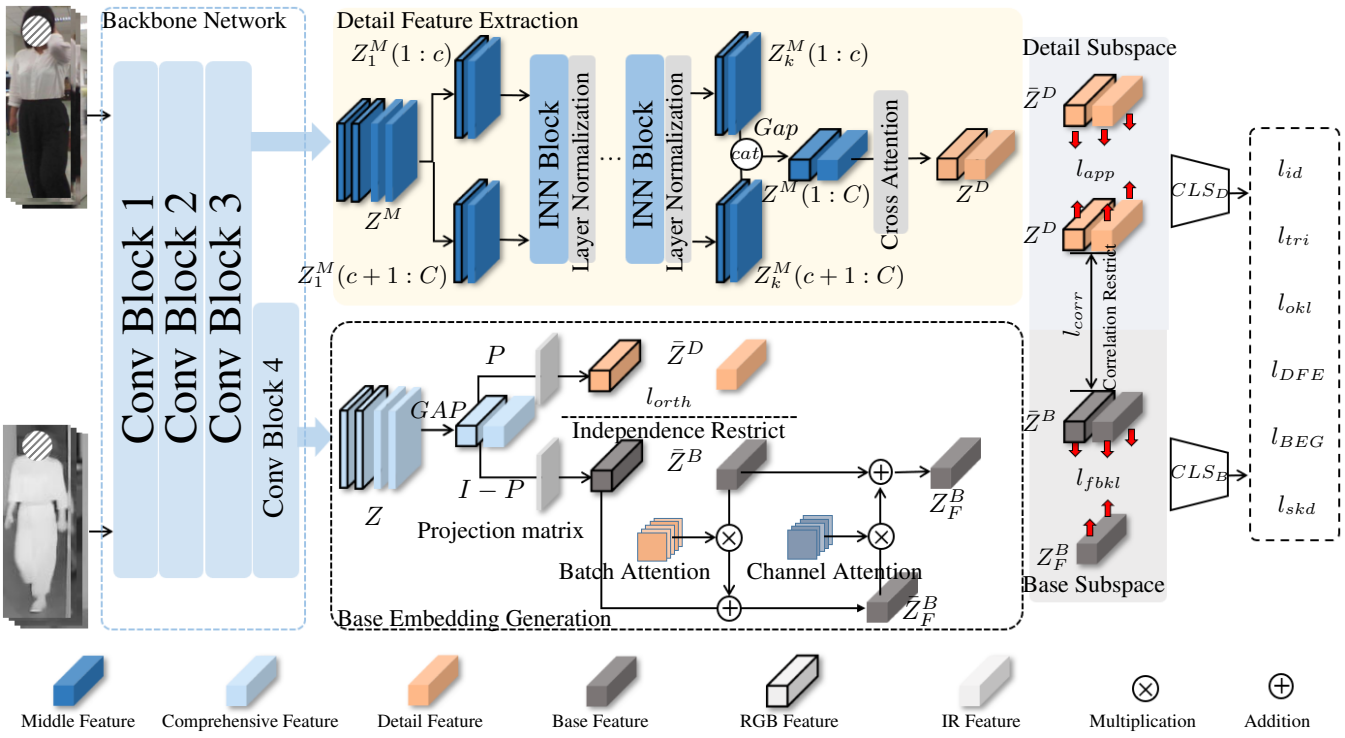


Figure 2: The pipeline of the proposed Base-Detail Feature Learning Framework (BDLF), which consists of a Detail Feature Extraction (DFE) module and a Base Embedding Generation (BEG) block, and jointly optimizes the extracted detail, base, and comprehensive features.

the training set $\{X_V, X_I\}$ consist of B identities, with each identity including P samples. Therefore, $X_V = \{x_V{}^{b,p}, b = 1, ..., B; p = 1, ..., P\}$ symbolizes the set of visible images, while $X_I = \{x_I{}^{b,p}, b = 1, ..., B; p = 1, ..., P\}$ denotes the set of infrared images. As illustrated in Figure 2, the VIS and IR images are processed through the backbone network, i.e,

$$\begin{aligned} Z_{V/I}^M &= E^{fore}(X_{V/I}) \\ Z_{V/I} &= E^{rear}(Z_{V/I}^M) \\ Z = cat(Z_V, Z_I), &\quad Z^M = cat(Z_V^M, Z_I^M) \end{aligned} \tag{1}$$

where $E^{fore}(\cdot)$ and $E^{rear}(\cdot)$ are the former and latter blocks of the backbone network, the embeddings $Z^M \in \mathrm{R}^{B\times C'\times H'\times W'}$ and $Z \in \mathrm{R}^{B\times C\times H\times W}$ denote the intermediate and complete outputs from the backbone for the VIS and IR modalities, $cat(\cdot)$ refers to the concatenation operation along the batch dimension.

3.2 Specific-shared Knowledge Distillation

We observe that the similarity of base information, such as contours and movements, between the VIS and IR modalities is noticeable. In contrast, the similarity of detail information including color, texture, and thermal details between the two modalities is suppressed. Inspired by [Zhao *et al.*, 2023], as shown in Figure 3, the base and detail features can be generated by increasing and reducing the correlation between the two modalities respectively. Based on this, we propose a novel specific-shared knowledge distillation (SKD) loss, which is numerically smoother and easier to optimize, formulated as follows:

$$l_{skd} = \frac{log[Corr(Z_V^B, Z_I^B)]}{\sqrt[3]{log[Corr(Z_V^D, Z_I^D)] + \gamma}} \tag{2}$$

in which $Z_{V/I}^B$ denotes the base features generated by the proposed BEG block, and $Z_{V/I}^D$ denotes the detail features extracted from the proposed DFE module. $Corr(\cdot)$ is the Pearson correlation coefficient operation, while γ represents a constant that ensures the denominator remains non-zero. According to optimize the SKD loss, the correlation between the VIS and IR modalities of both base and detail features(i.e, $Corr(Z_V^B, Z_I^B)$ and $Corr(Z_V^D, Z_I^D)$ in formula (2)) is simultaneously increased and decreased. This approach allows the proposed DFE module to extract embeddings rich in detailed knowledge. Consequently, the proposed BEG block is capable of generating base embeddings that contain a greater amount of modality-shared knowledge.

3.3 Detail Feature Extraction

The proposed DFE module aims to acquire detail features that imply modality-specific information from the intermediate embedding $Z_{V/I}^M$ by utilizing a series of invertible neural network (INN) blocks[Zhao *et al.*, 2023][Dinh *et al.*, 2017][Zhou *et al.*, 2022], which can effectively preserves detailed characteristics and mitigates information loss during feature extraction by making its input and output embeddings

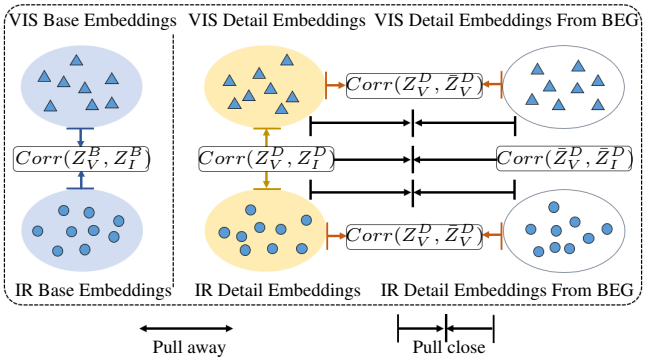


Figure 3: Illustration of correlation instruct to learn modality specific and shared information.

are mutually generated. Taking the VIS case as an example, we obtain the input for the DFE module, ${Z_V(1:c)}^M$ and ${Z_V(c+1:C)}^M \in \mathrm{R}^{\frac{B}{2}\times\frac{C'}{2}\times H'\times W'}$ by splitting ${Z_V}^M$ in half along the channel dimension. The transformations in each block can be denoted as follows:

$$\begin{aligned} Z^M_{V,k+1}(c+1:C) &= Z^M_{V,k}(c+1:C) + F_1[Z^M_{V,k}(1:c)] \\ Z^M_{V,k+1}(1:c) &= F_2[Z^M_{V,k+1}(c+1:C)] \\ &\quad + Z^M_{V,k}(1:c) \bullet exp\{F_3[Z^M_{V,k+1}(c+1:C)]\} \\ Z^M_{V,k+1} &= LN\{cat[Z^M_{V,k+1}(1:c), Z^M_{V,k+1}(c+1:C)]\} \end{aligned} \quad (3)$$

Here, $Z^M_{V,k}$ is the input of the kth ($k \in 1,...,K$) block, $F_i(\cdot)(i \in 1,2,3)$ denotes the convolution blocks. The symbol $\bullet$ indicates element-wise multiplication of matrices, $LN(\cdot)$ represents layer normalization in the Lite Transformer[Wu *et al.*, 2020] and $cat(\cdot)$ is the channel concatenation operation. The IR situation can be easily derived by substituting I for the subscript V in the aforementioned formulas.

At the final stage of DFE, we consider the detail features of both modalities integrally, as concatenating the extracted detail embeddings from the two modalities can help reduce computational complexity. Therefore, we feed the extracted detail embeddings into a cross-attention-based transformer to facilitate cross-modality reasoning and information exchange. This process enables the detail feature integration of knowledge from various modalities and allows for a more effective focus on distinguishable information, thereby enhancing the robustness and efficacy of semantic representation. Inspired by[Li *et al.*, 2022][Xu *et al.*, 2024], the transformer can be denoted as follows:

$$\begin{aligned} Z^M_{V/I} &= GAP(Z^M_{V/I,K}) \\ Z^D_V &= softmax[(Z^M_V W_q)(Z^M_I W_k)^T](Z^M_I W_v) + Z^M_V \\ Z^D_I &= softmax[(Z^M_I W_q)(Z^M_V W_k)^T](Z^M_V W_v) + Z^M_I \end{aligned} \quad (4)$$

where $GAP(\cdot)$ is the global average pooling operation, $Z^M_{V/I} \in \mathrm{R}^{\frac{B}{2}\times C'}$ denotes the embeddings after pooling. W_q, W_k , and W_v are the learnable parameters for DFE, and $softmax(\cdot)$ indicates the calculation of the softmax by row.

Ultimately, the detail feature $Z^D \in \mathrm{R}^{B\times C'}$ produced by the proposed DFE module is obtained by concatenating the VIS and IR detail embeddings along the batch dimension:

$$Z^D = cat(Z^D_V, Z^D_I) \quad (5)$$

With the proposed SKD loss formulated in formula (2), the extracted detail feature Z^D can significantly enrich modality-specific detail knowledge. Thus a private classifier CLS_D that is specially designed for the detail feature Z^D is constructed, alongside a communal classifier CLS_B that processes the base embeddings and the comprehensive feature Z obtained from formula (1), as illustrated in Figure 2. Furthermore, the commonly used id loss[Luo *et al.*, 2019] driven by cross-entropy ($ce(p,q) = -\sum_{i=1}^{n} q_i log(p_i)$) was applied to strengthen the distinguishable information of detail feature Z^D, i.e,

$$l^D_{id} = \mathrm{E}_{(z^D \sim Z^D)}\, ce(CLS_D(z^D), Y) \quad (6)$$

Since there are differences in the distribution of classification results between the detail feature Z^D and the comprehensive feature Z, this misalignment may impede our goal of enhancing the representation ability of Z leveraging detailed knowledge. Therefore, we constrain the probability distribution predicted from Z^D to align with the distribution from Z, ensuring that their semantic representations are consistent. This process can be expressed as follows:

$$l_{odkl} = \mathrm{E}_{(z,z^D \sim Z,Z^D)}\, ce(CLS_D(z^D), CLS_B(z)) \quad (7)$$

The total loss of the proposed DFE module can be obtained by combining formulas (6) and (7):

$$l_{DFE} = l^D_{id} + l_{odkl} \quad (8)$$

3.4 Base Embedding Generation

The proposed BEG block is designed to produce the base embeddings from Z utilizing the acquired detail feature Z^D. Take notice that there are significant semantic differences between modality-specific detail information such as color and texture and modality-shared base information, which includes movements, contours, and so on. For this reason, inspired by [Feng *et al.*, 2023], we have developed a method to ensure that the detail(base) features can only contain modality-specific(modality-shared) distinguishable knowledge, thereby maximizing the collection of both modality-specific and modality-shared information. Furthermore, the proposed DFE and BEG blocks can learn these two categories of knowledge simultaneously without interfering with each other. Based on this premise, we consider the detail and base embeddings to be independent of each other, i.e, $Z^D \perp Z^B$. According to the approach of making $\bar{Z}^D$ comprehensively converge to Z^D and impose the independence restriction between the detail and base embedding, the proposed BEG block can then generate modality-shared base embedding by excluding detailed knowledge from Z in the semantic space,i.e,

$$\begin{cases} Z \times P = \bar{Z}^D \\ Z \times (I - P) = Z^B \end{cases}, \bar{Z}^D \to Z^D \quad (9)$$

in which Z is the output of backbone network, I is the identity matrix, $\rightarrow$ denotes approximating, $\bar{Z}^D$ and $Z^B \in \mathrm{R}^{B\times C}$ are the gained detail and base embeddings by using a projection matrix $P \in \mathrm{R}^{C\times C}$ to decompose Z into mutually orthogonal subspaces. By the properties of orthogonal projection matrix, P should be a conjugate symmetric idempotent matrix and must satisfy the following constraints in the real number case:

$$P^2 = P, P^T = P \tag{10}$$

The process of approaching can be divided into three components: approximating in the feature space, semantic representation, and the correlation between Z^D and $\bar{Z}^D$. In the case of approximating on feature space, we first calculate the distances between all embeddings in a mini-batch for $\bar{Z}^D$ and Z^D respectively, and obtain the difference map M by:

$$M = ||softmax[\bar{Z}^D(\bar{Z}^D)^T - Z^D(Z^D)^T]||^2 \tag{11}$$

Then we enforce the distance distribution of $\bar{Z}^D$ to converge to that of Z^D by optimizing the following loss:

$$l_{fkl} = \mathrm{E}_{(a_{i,j}\sim M)}\, a_{i,j} \tag{12}$$

Furthermore, we aligned the semantic representation between Z^D and $\bar{Z}^D$ by adjusting the predicted probability distribution of $\bar{Z}^D$ closer to that of Z^D. By drawing an analogy with formula (7), we have:

$$l_{dkl} = \mathrm{E}_{(\bar{z}^D, z^D\sim \bar{Z}^D, Z^D)}\, ce(CLS_D(\bar{z}^D), CLS_D(z^D)) \tag{13}$$

Considering that the detail feature $\bar{Z}^D$ generated by the BEG block should exhibit the same correlation properties as Z^D. As illustrated on the right side of the dashed line in Figure 3, we achieved consistency in correlation between $\bar{Z}^D$ and Z^D by pulling close their cross-modalities correlations denoted as $Corr(Z_V^p, Z_I^p), p \in \{D, /\bar{D}\}$ and by reducing the discrepancy in correlation within the same modality, represented as $Corr(Z_m^D, Z_m^{\bar{D}}), m \in \{V, I\}$. This is accomplished by optimizing the follows loss:

$$l_{dcorr} = \frac{(Corr(\bar{Z}_V^D, \bar{Z}_I^D) - Corr(Z_V^D, Z_I^D))^2}{Corr(\bar{Z}_V^D, Z_V^D)^2 + Corr(\bar{Z}_I^D, Z_I^D)^2 + \gamma} \tag{14}$$

Thereby, the total approaching function for $\bar{Z}$ is:

$$l_{app} = l_{fkl} + l_{dkl} + l_{dcorr} \tag{15}$$

After the description provided above, we generated the base feature Z^B by eliminating the detail feature Z^D from Z. Given that the base information across modalities, such as contours and movements, should exhibit significant similarities, we constructed a cross-modality feature fusion method that integrates the base feature Z_V^B and Z_I^B to generate an auxiliary feature Z_F^B. Inspired by [Li *et al.*, 2022] [Wang *et al.*, 2018], the fusion method can be formulated as follows:

$$\begin{aligned}\bar{Z}_F^B &= \frac{1}{C}[(Z_V^B P_q)^T (Z_I^B P_k)](Z_I^B P_v) + Z_V^B \\ Z_F^B &= \frac{2}{B}[(Z_I^B Q_q)(\bar{Z}_F^B Q_k)^T](\bar{Z}_F^B Q_v) + Z_I^B\end{aligned} \tag{16}$$

Here, $Z_{V/I}^B \in \mathrm{R}^{\frac{B}{2}\times C}$ represents the cross-modality base embedding, P,Q are the learnable parameters. The fused Z_F^B aggregates the base knowledge from VIS and IR modalities, employing attention mechanisms across both channel and batch dimensions. We then enhance the similarity between Z_V^B and Z_I^B by aligning them with Z_F^B:

$$l_{fbkl} = \mathrm{E}_{(z_F^B, z_{V/I}^B \sim Z_F^B, Z_{V/I}^B)}\, ce(CLS_B(z_{V/I}^B), CLS_B(z_F^B)) \tag{17}$$

This approach ensures that Z^B contains only the knowledge shared between the modalities. In addition, we also utilize cross-modality semantic alignment for $Z_{V/I}^B$ to strengthen the collection of modality-shared knowledge:

$$l_{bkl} = \mathrm{E}_{(z_{V/I}^B \sim Z_{V/I}^B)}\, ce(CLS_B(z_V^B), CLS_B(z_I^B)) \tag{18}$$

The id loss for both was also employed to enhance the distinguishable information of Z^B and Z_F^B, and the loss for cross-modality feature fusion method is:

$$l_{cmf} = l_{id}^F + l_{fbkl} \tag{19}$$

Consequently, the total loss for the BEG block can be summarized as follows:

$$l_{BEG} = l_{id}^B + l_{app} + l_{bkl} + l_{cmf} + l_{orth} \tag{20}$$

where l_{orth} represents the constraint in formula (10) for parameter P to achieve the decomposition of orthogonal subspaces.

3.5 Optimization

In the preceding section, the proposed DFE module extracted detailed knowledge from the intermediate feature Z^M and subsequently produced the detail feature Z^D, the proposed BEG block produced the base feature by eliminating the detailed knowledge from the comprehensive feature Z, the proposed SKD loss ensures that both the detail and base features effectively capture modality-specific and shared information. We also incorporated the commonly used id and triplet loss [Hermans *et al.*, 2017] l_{tri} for Z into our method. Similar to (18), we enforce cross-modality consistency for Z by:

$$l_{okl} = \mathrm{E}_{(z_{V/I}\sim Z_{V/I})}\, ce(CLS_B(z_V), CLS_B(z_I)) \tag{21}$$

Eventually, the total loss of BDLF is defined as:

$$l_{totel} = l_{id} + l_{tri} + l_{okl} + l_{DFE} + l_{BEG} + l_{skd} \tag{22}$$

4 Experiments

In this section, we validate the effectiveness of our BDLF by conducting experiments on the widely recognized SYSU-MM01, RegDB and LLCM benchmarks.

4.1 Datasets and Evaluation Protocol

SYSU-MM01 dataset [Wu *et al.*, 2017] comprises 287,628 VIS and 15,792 IR images from 491 identities captured by 4 RGB and 2 IR cameras. It features both All-Search and Indoor-Search modes for evaluation. RegDB [Nguyen *et al.*, 2017] contains 412 identities, each represented by 10 VIS and

Methods	Venue	SYSU-MM01				RegDB				LLCM			
		All-Search		Indoor-Search		VIS to IR		IR to VIS		VIS to IR		IR to VIS	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
CAJ[Ye *et al.*, 2021a]	ICCV'21	69.9	66.9	76.3	80.4	85.0	79.1	84.8	77.8	56.5	59.8	48.8	56.6
MMN[Zhang *et al.*, 2021]	ACMMM'21	70.6	66.9	76.2	79.6	91.6	84.1	87.5	80.5	59.9	62.7	52.5	58.9
FMCNet[Zhang *et al.*, 2022a]	CVPR'22	66.3	62.5	68.2	74.1	89.1	84.4	88.4	83.9	-	-	-	-
LUPI[Alehdaghi *et al.*, 2022]	ECCV'22	71.1	67.6	82.4	82.7	88.0	82.7	86.8	81.3	-	-	-	-
MSCLNet[Zhang *et al.*, 2022b]	ECCV'22	77.0	71.6	78.5	81.2	84.2	81.0	83.7	78.3	-	-	-	-
DEEN[Zhang and Wang, 2023]	CVPR'23	74.7	71.8	80.3	83.3	91.1	85.1	89.5	83.4	62.5	65.8	54.9	62.9
SGIEL[Feng *et al.*, 2023]	CVPR'23	75.2	70.1	78.4	81.2	92.2	86.6	91.1	85.2	-	-	-	-
TMD[Lu *et al.*, 2024]	TMM'23	73.9	67.8	81.2	78.9	93.0	84.3	87.4	81.3	-	-	-	-
AGCC[Yu *et al.*, 2024]	PR'24	75.9	73.0	79.3	84.6	92.6	86.2	91.4	84.9	-	-	-	-
ReViT[Sarker and Zhao, 2024]	PR'24	68.1	65.1	72.4	77.6	91.7	86.0	93.0	86.1	-	-	-	-
STAR[Wu *et al.*, 2024]	TMM'24	76.1	72.7	83.5	85.8	94.1	88.8	93.3	88.2	-	-	-	-
BDLF(ours)	-	76.8	74.6	84.2	85.8	94.4	90.1	94.5	89.6	67.0	68.9	58.1	64.5

Table 1: Comparisons between the proposed BDLF and several state-of-the-art methods on the SYSU-MM01, RegDB, and LLCM datasets.

Settings				SYSU-MM01	
DFE	l_{app}	l_{orth}	l_{skd}	R-1	mAP
				72.7	68.1
✓				73.7	69.6
✓	✓			75.3	72.1
✓			✓	73.7	69.0
	✓	✓	✓	74.0	70.9
✓	✓		✓	75.5	72.3
✓	✓	✓	✓	76.8	74.6

Table 2: Effectiveness of each component for the proposed BDLF.

10 IR images captured from a pair of cameras. We adhere to the evaluation protocol outlined in [Ye *et al.*, 2022b] to randomly split the identities into training and testing sets of equal size. LLCM [Zhang and Wang, 2023]is a challenging large-scale low-light dataset for VI-ReID task, which contains 713 identities with 25,626 VIS and 21,141 IR images, all captured by 9 cameras in both RGB and IR modalities

The Cumulative Matching Characteristic curve (CMC) and mean Average Precision(mAP) are adopted as standard evaluation metrics in our experiments to comprehensively assess the performance of our framework.

4.2 Implementation Details

The entire framework is implemented using PyTorch and runs on a single NVIDIA RTX3090 GPU with 24GB VRAM. We employed a pre-trained ResNet-50[He *et al.*, 2016b] as the backbone network and incorporated INN blocks with affine coupling layers[Dinh *et al.*, 2017][Zhou *et al.*, 2022] to construct the DFE module, setting the number of INN blocks to 6. All images are resized to $3 \times 384 \times 144$, and we adopted the Random Channel Exchangeable Augmentation and Channel-Level Random Erasing techniques proposed in [Ye *et al.*, 2021b] during the training phase. The SGD optimizer was used, with the initial learning rate set to 1×10^{-2}, which was warmed up to 1×10^{-1} during the first 10 epochs, then we decayed the learning rate to 1×10^{-2} and 1×10^{-3} at epochs 20 and 95 for SYSU-MM01, and at epochs 70 and 130 for RegDB and LLCM, respectively. The learning rate was further decayed to 1×10^{-4} at 180 epoch, with a total of 220 epochs. For each mini-batch, we randomly sampled 8 identities, each consisting of 4 VIS and 4 IR images for training. Additionally, the exponential moving average (EMA) model [Ge *et al.*, 2020] also employed in our method.

4.3 Comparison with State-of-the-art Methods

We demonstrate the superiority of our BDLF by comparing performance with several existing state-of-the-art methods on the SYSU-MM01, RegDB, and LLCM datasets. The performance of these methods is presented in Table 1, with optimal performances annotated by underlining.

Comparison on SYSU-MM01 and RegDB. Table 1 presents the results of our BDLF alongside selected outstanding methods, confirming the superiority of our BDLF, which almost outperforms all other state-of-the-art methods. In the All-Search mode of SYSU-MM01, our method achieved a rank-1 accuracy of 76.8% and a mAP of 74.6%, in the Indoor-Search mode, BDLF achieved a rank-1 accuracy of 84.2% and a mAP of 85.8%. On the RegDB dataset, our method achieved a rank-1 accuracy of 94.4% and a mAP of 90.1% for the VIS to IR search, and attained a rank-1 accuracy of 94.5% and a mAP of 89.6% for the IR to VIS search. These results validate the effectiveness of BDLF that independently learns the detail and base information and sufficiently utilizes cross-modalities knowledge.

Comparison on LLCM. According to Table 1, our method outperformed other approaches. Specifically, BDLF achieved a rank-1 accuracy of 67.0% and a mAP of 68.9% in VIS to IR search, as well as a rank-1 accuracy of 58.1% and a mAP of 64.5% in IR to VIS search. It is evident that our BDLF is well-equipped to handle challenging scenarios.

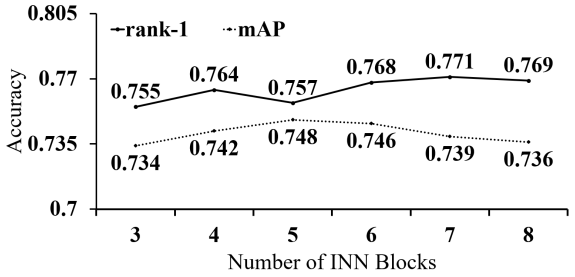


Figure 4: Effectiveness of how many INN blocks are more favorable for the proposed DFE.

Location of DFE	SYSU-MM01	
	R-1	mAP
After stage-1	59.9	55.3
After stage-2	71.1	67.1
After stage-3	76.8	74.6
After stage-4	73.4	71.1

Table 3: Effectiveness of which stage of ResNet-50 to combine the proposed DFE.

4.4 Ablation Studies

Effectiveness of each component. In this section, we designed an ablation experiment to validate the effectiveness of certain components of BDLF. Specifically, we removed the DFE, l_{app}, l_{orth} and l_{skd} modules from BDLF, while retaining the backbone with the BEG block as the baseline. All experiments adopted the same training settings, and we evaluated their performance in the All-Search mode of SYSU-MM01. The results are presented in Table 2, Notably, the removal of the DFE module resulted in poor precision, demonstrating the effective detail extraction capability of DFE. The experiments also indicated that the l_{app} loss enhances the model's distinguishing performance by effectively aiding in the generation of base embeddings, eliminating detailed knowledge from the comprehensive feature. Although the DFE module significantly promotes the mining of detail information, its performance remains suboptimal, as the model cannot extract all modality-specific and shared information without interference each other due to the absence of correlation constraint l_{corr} and independent constraint l_{orth} .

Effectiveness of how many INN blocks are more favorable for DFE. The proposed DFE module consists of a series of INN blocks with an LN layer to extract detail information non-destructively. We conducted experiments to determine the optimal number of blocks for our framework. As shown in Figure 4, we modified the number of INN blocks and evaluated performance in the All-Search mode of SYSU-MM01. The results indicate accuracy gradually improves as the number of INN blocks increases, reaching a plateau when the count is 6. This confirm that a balance exists between accuracy and computational complexity when the number of INN blocks is set to 6.

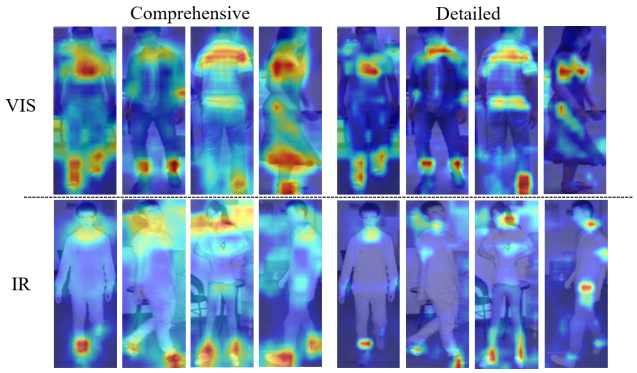


Figure 5: Visualization of the comprehensive and detailed features.

Effectiveness of which stage of ResNet-50 to combine DFE module. In this section, we implement experiments to assess which stage of ResNet-50 is most suitable for serving as the input to the proposed DFE module. All experiments maintain consistent settings, except for the locations of the DFE module within ResNet-50. The results are presented in Table 3, we observed that connecting the DFE module to stage-3 of ResNet-50 yielded the best accuracy in the All-Search mode of SYSU-MM01. This can be attributed to the fact that modality-shared information is more prominent in the high-level features produced by stages-4, which impedes the extraction of modality-specific detail information. Furthermore, the low-level features generated by stages 1 and 2 are inadequate for effectively expressing the semantics necessary to distinguish between different identities. These findings elucidate why the best accuracy is achieved when the DFE module is connected to stage-3 of ResNet-50.

4.5 Visualization

To investigate the detail information extraction capabilities of the proposed DFE, we visualize the comprehensive and detailed features of several identities produced by BDLF. As illustrated in Figure 5, a comparison of the images of comprehensive and detailed features reveals that the attention regions of the comprehensive features is broader and more dispersed than that of the detailed features. This observation indicates that the DFE module has the capacity to focus on subtly distinguishable characteristics.

5 Conclusion

In this paper, we propose a novel base-detail feature learning framework(BDLF) that learns detail and base features from a correlation and mutual information maximization for the VI-ReID task. The proposed BDLF consists of a DFE module and a BEG block. The DFE module non-destructively extracts detail information, while the BEG block generates base features by eliminating detail information from the output of the backbone network, constrained by independence and correlation requirements on the detail and base embeddings. Extensive experiments on the SYSU-MM01, RegDB, and LLCM datasets have demonstrated the superiority of BDLF.

Acknowledgments

This work was partially supported by the Shenzhen Science and Technology Program (KJZD20230923114600002), and the Guangdong Major Project of Basic and Applied Basic Research (2023B0303000010).

References

[Alehdaghi *et al.*, 2022] Mahdi Alehdaghi, Arthur Josi, Rafael M. O. Cruz, and Eric Granger. Visible-infrared person re-identification using privileged intermediate information. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V*, volume 13805 of *Lecture Notes in Computer Science*, pages 720–737. Springer, 2022.

[Cao *et al.*, 2023] Chengzhi Cao, Xueyang Fu, Hongjian Liu, Yukun Huang, Kunyu Wang, Jiebo Luo, and Zheng-Jun Zha. Event-guided person re-identification via sparse-dense complementary learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 17990–17999. IEEE, 2023.

[Chen *et al.*, 2022] 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.

[Dinh *et al.*, 2017] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[Feng *et al.*, 2023] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22752–22761. IEEE, 2023.

[Ge *et al.*, 2020] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[He *et al.*, 2016a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[He *et al.*, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[Li *et al.*, 2020] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an X modality. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 4610–4617. AAAI Press, 2020.

[Li *et al.*, 2022] Wuyang Li, Xinyu Liu, and Yixuan Yuan. SIGMA: semantic-complete graph matching for domain adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5281–5290. IEEE, 2022.

[Liu *et al.*, 2022] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19344–19353. IEEE, 2022.

[Lu *et al.*, 2024] Zefeng Lu, Ronghao Lin, and Haifeng Hu. Tri-level modality-information disentanglement for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 26:2700–2714, 2024.

[Luo *et al.*, 2019] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1487–1495. Computer Vision Foundation / IEEE, 2019.

[Nguyen *et al.*, 2017] 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.

[Park *et al.*, 2021] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12026–12035, 2021.

[Sarker and Zhao, 2024] Prodip Kumar Sarker and Qingjie Zhao. Enhanced visible-infrared person re-identification

based on cross-attention multiscale residual vision transformer. *Pattern Recognit.*, 149:110288, 2024.

[Wang *et al.*, 2018] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7794–7803. Computer Vision Foundation / IEEE Computer Society, 2018.

[Wang *et al.*, 2020] Guan'an Wang, Yang Yang, Tianzhu Zhang, Jian Cheng, Zengguang Hou, Prayag Tiwari, and Hari Mohan Pandey. Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification. *Neural Networks*, 128:294–304, 2020.

[Wang *et al.*, 2022] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7287–7297. IEEE, 2022.

[Wu *et al.*, 2017] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5390–5399. IEEE Computer Society, 2017.

[Wu *et al.*, 2020] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[Wu *et al.*, 2024] Jianbing Wu, Hong Liu, Wei Shi, Mengyuan Liu, and Wenhao Li. Style-agnostic representation learning for visible-infrared person re-identification. *IEEE Trans. Multim.*, 26:2263–2275, 2024.

[Xu *et al.*, 2024] QiHao Xu, Xiaoling Luo, Chao Huang, Chengliang Liu, Jie Wen, Jialei Wang, and Yong Xu. Hacdr-net: Heterogeneous-aware convolutional network for diabetic retinopathy multi-lesion segmentation. In *AAAI Conference on Artificial Intelligence*, 2024.

[Yan *et al.*, 2021] Cheng Yan, Guansong Pang, Lei Wang, Jile Jiao, Xuetao Feng, Chunhua Shen, and Jingjing Li. Bv-person: A large-scale dataset for bird-view person re-identification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10923–10932. IEEE, 2021.

[Ye *et al.*, 2021a] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13547–13556, 2021.

[Ye *et al.*, 2021b] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13547–13556, 2021.

[Ye *et al.*, 2022a] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6):2872–2893, 2022.

[Ye *et al.*, 2022b] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.

[Yu *et al.*, 2024] Hao Yu, Xu Cheng, Kevin Ho Man Cheng, Wei Peng, Zitong Yu, and Guoying Zhao. Discovering attention-guided cross-modality correlation for visible-infrared person re-identification. *Pattern Recognit.*, 155:110643, 2024.

[Zhang and Wang, 2023] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2153–2162. IEEE, 2023.

[Zhang *et al.*, 2021] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 788–796. ACM, 2021.

[Zhang *et al.*, 2022a] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7339–7348. IEEE, 2022.

[Zhang *et al.*, 2022b] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, page 462–479, Berlin, Heidelberg, 2022. Springer-Verlag.

[Zhao *et al.*, 2023] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5906–5916. IEEE, 2023.

[Zhou *et al.*, 2022] Man Zhou, Xueyang Fu, Jie Huang, Feng Zhao, Aiping Liu, and Rujing Wang. Effective pan-sharpening with transformer and invertible neural network. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–15, 2022.