

FMCNet: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification

Qiang Zhang¹, Changzhou Lai, Jianan Liu¹, Nianchang Huan¹, Jungong Han²

¹School of Mechano-Electronic Engineering, Xidian University, China

²Computer Science Department, Aberystwyth University, U.K.

qzhang@xidian.edu.cn, fchzhilai,jianan_liu,nchuang_g@stu.xidian.edu.cn, jungonghan77@gmail.com

Abstract

For Visible-Infrared person Re-Identification (VI-ReID), existing modality-specific information compensation based models try to generate the images of missing modality from existing ones for reducing cross-modality discrepancy. However, because of the large modality discrepancy between visible and infrared images, the generated images usually have low qualities and introduce much more interfering information (e.g., color inconsistency). This greatly degrades the subsequent VI-ReID performance. Alternatively, we present a novel **Feature-level Modality Compensation Network** (FMCNet) for VI-ReID in this paper, which aims to compensate the missing modality-specific information in the feature level rather than in the image level, i.e., ~~directly generating those missing modality-specific features of one modality from existing modality-shared features of the other modality.~~ This will enable our model to mainly generate some discriminative person related modality-specific features and discard those non-discriminative ones for benefiting VI-ReID. For that, a **single-modality feature decomposition module** is first designed to decompose single-modality features into **modality-specific ones** and **modality-shared ones**. Then, a **feature-level modality compensation module** is present to generate those missing modality-specific features from existing modality-shared ones. Finally, **a shared-specific feature fusion module** is proposed to combine the existing and generated features for VI-ReID. The effectiveness of our proposed model is verified on two benchmark datasets.

1. Introduction

Person Re-Identification (ReID) aims at matching the given pedestrians from an image gallery taken by different cameras. Most existing ReID models focus on the visible-visible image matching (i.e., VV-ReID). However,

Figure 1. Illustration of the differences between our model and existing VI-ReID models. (a) Existing modality-shared feature learning based models. (b) Existing image-level compensation based models. (c) Our proposed feature-level compensation based model.

these models may have poor performance when visible cameras cannot well capture information, such as at night. Compared with visible cameras, infrared cameras can still capture clear images under those poor illumination conditions. Moreover, most cameras in modern surveillance systems support autoswitch between the visible and infrared modes under different illumination conditions. Accordingly, Visible-Infrared ReID (i.e., VI-ReID) has raised more and more attention recently.

The main challenge of VI-ReID lies in the modality discrepancy between the visible and infrared images. Meanwhile, it also suffers from large **person variations**, such as viewpoints and postures. As shown in Fig. 1(a), most existing models [1–7] try to extract the discriminative modality-shared features for VI-ReID. Although great improvements have been achieved, these models inevitably discard lots of discriminative person-related modality-specific information, which may also benefit VI-ReID. Considering that,

* Equally corresponding authors.

some works [8, 9] propose the idea of **modality-specific information compensation**, which attempts to first generate those missing modality-specific information from existing modality and then jointly uses the generated and original information for VI-ReID.

However, existing modality-specific information compensation based models usually achieve inferior results compared with those modality-shared feature learning based models. This may impute to **the image-level compensation** of existing models. That is, as shown in Fig. 1(b), existing models first generate the images of missing modality from the images of existing modality and then extract discriminative person features from the paired images for VI-ReID. However, it is very difficult to generate high-quality images of one modality from another modality, due to the large modality discrepancy between the visible and infrared images. Especially, when generating visible images from infrared images, much more noisy information (e.g., color inconsistency), instead of discriminative person features, will be introduced for VI-ReID. Besides, these existing modality-specific information compensation based models usually follow a two-stage structure and are not end-to-end trainable, where the image generation sub-networks and VI-ReID subnetworks are independent trained.

Actually, compared with the modality discrepancy between visible and infrared images, their features' discrepancy has been reduced to some extents, since some common semantics information usually coexists in the unimodal visible and infrared features. Therefore, the translation between visible data and infrared data in the feature level may be easier than that in the image level. Meanwhile, as discussed in some existing works [10–12], the single-modality features (e.g., unimodal visible features or infrared features) can be decomposed into their own modality-specific features and modality-shared features. The difficulties for cross-modality translation can be further reduced by generating those missing modality-specific features from existing modality-shared features rather than from the whole single-modality features. More importantly, compared with image-level translation, the feature-level translation allows us to explicitly control the generation of those missing modality-specific features as our requirements by designing some dedicated loss functions. For example, we can only generate some discriminative person-related modality-specific features and discard those non-discriminative ones for benefiting VI-ReID.

Considering that, we will present a novel end-to-end feature-level modality-specific information compensation based model, e., the **Feature-level Modality Compensation Network (FMCNet)**, for VI-ReID in this paper. As shown in Fig. 1(c), our proposed FMCNet aims to compensate those missing modality-specific information in the feature level rather than in the image level, i.e., directly generat-

ing those missing modality-specific features of one modality from existing modality-shared features of other modality. To this end, a **Single-modality Feature Decomposition (SFD) module** is first utilized to decompose the input single-modality features into their own modality-specific and modality-shared features, respectively. Meanwhile, a **modality decomposition loss** is designed to facilitate the decomposition of those single-modality features. Then, a **Feature-level Modality Compensation (FMC) module** is designed to generate the missing modality-specific features of one modality from the existing modality-shared ones of the other modality for each sample image. Finally, a **Shared-speci c Feature Fusion (SFF) module** is designed to jointly use the existing modality-shared and modality-specific features as well as the generated modality-specific features for VI-ReID.

Similarly, cm-SSFT [13] also tries to simultaneously exploit those modality-shared and modality-specific features for VI-ReID. It achieves shared-speci c feature transfer by modeling the affinities among different samples. Specially, those missing modality-specific features in the cm-SSFT are transferred from all the samples of the other modality in the gallery. This may also introduce more modality-specific information of other identities, thus easily leading to sub-optimal results. Different from cm-SSFT, our proposed model does not rely on other samples and is able to directly and explicitly generate those missing modality-specific features from its own modality-shared features.

In summary, the main contributions of this work are as follows:

- (1) A novel FMCNet is presented, which proposes feature-level rather than image-level modality-specific information compensation for VI-ReID. This enables our model to focus on generating some required missing modality-specific features (e.g., discriminative person-related ones) for VI-ReID.
- (2) Our proposed FMCNet provides an unified end-to-end framework, achieving unimodal feature decomposition, modality-specific feature compensation and modality shared-speci c feature fusion for VI-ReID via the proposed SFD, FMC and SFF modules, respectively.
- (3) Our model significantly outperforms those image-level compensation based models and obtains competitive and even better results than some state-of-the-art modality-shared feature learning based ones.

2. Related work

VV-ReID has been well studied for many years and has achieved significant progress. Summarizing the vast amount of existing works on VV-ReID is beyond the scope of this paper and we refer those interested readers to [14–16] for recent surveys.

Recently, VI-ReID has raised more and more attention

Figure 2. Framework of the proposed Feature-level Modality Compensation Network (FMCNet).

due to its potential in real-life applications [5–7]. Most existing VI-ReID models can be divided into two categories, i.e., modality-shared feature learning based ones and modality-specific information compensation based ones. Modality-shared feature learning based models aim to embed the features from different modalities into the same feature space and reduce the cross-modality discrepancy by using some feature-level constraints [1, 4, 7, 12, 17]. For example, [4] proposed a dual-path network to extract an adversarial way. Finally, the original modality-specific modality-shared features from the input images by using features and modality-shared features as well as their combined a shared network and designed a new bi-directional dual-modality-specific features will be combined in the constrained top-ranking loss to learn person discriminative features for VI-ReID. Differently, modality-specific information compensation based models try to make up the missing modality-specific information from existing modalities [8, 9, 18–20]. For example, [8] first designed two cross-modality image translation sub-networks to transfer an infrared image into its visible counterpart and transferred a visible image to its infrared version, respectively. Then, a ReID network was presented to reduce the appearance discrepancy by introducing some feature-level constraints.

In this paper, our proposed model follows the idea of modality-specific information compensation. However, different from existing models which compensate missing modality-specific information in the image level, our proposed model adopts the feature-level compensation.

3. Method

As shown in Fig. 2, the proposed model, Feature-level Modality Compensation Network (FMCNet), mainly consists of three parts, i.e., a Single-modality Feature Decomposition (SFD) module, a Feature-level Modality Com-

pensation (FMC) module and a Shared-specific Feature Fusion (SFF) module. Concretely, the proposed SFD module first extracts single-modality features from the input modality-specific information and then decomposes them into their own modality-specific and modality-shared ones. Then, the proposed module generates the missing (or compensated) visible (infrared) modality-specific features from those existing decomposed infrared (visible) modality-shared features in an adversarial way. Finally, the original modality-specific features and modality-shared features as well as their compensated modality-specific features will be combined in the proposed SFF module for VI-ReID. Details about these modules will be discussed in the following contents. Suppose that the training set $(X_V; X_I)$ contains P identities and each identity contains K samples. $X_V = \{x_V^{k;p}; k = 1; \dots; K; p = 1; \dots; P\}$ denotes visible sample images, and $X_I = \{x_I^{k;p}; k = 1; \dots; K; p = 1; \dots; P\}$ denotes infrared sample images.

3.1. SFD Module

As shown in Fig. 2, given the input visible images X_V or the infrared images X_I , the proposed SFD module first extracts their single-modality features and then decomposes those extracted single-modality visible (infrared) features into their own modality-specific features and modality-shared features. Here, the ways of extracting and decomposing those single-modality visible and infrared features are the same. Therefore, we take the input visible images X_V as the example to detail the corresponding process.

Specifically, the single-modality features F_V are first extracted from X_V by using a visible feature extraction sub-network $E_V(\cdot)$. Then, a visible modality-specific feature

The decorrelation loss L_{dc} aims to push the modality-specific features away from the modality-shared features. For that, it first computes the modality-specific feature center as well as the modality-shared feature center for each identity by

$$C_{sp;m}^p = \frac{1}{K} \sum_{k=1}^K F_{sp;m}^{k;p}; C_{sh;m}^p = \frac{1}{K} \sum_{k=1}^K F_{sh;m}^{k;p}; \quad (3)$$

Figure 3. Illustration of the proposed MD loss.

extraction sub-network $E_{sp}^V(\cdot)$ and a modality-shared feature extraction sub-network $E_{sh}(\cdot)$ are performed on F_V to decompose them into their corresponding visible modality-specific features $F_{sp;V}$ and visible modality-shared features $F_{sh;V}$, respectively, i.e.,

$$F_V = E_V(X_V); F_{sp;V} = E_{sp}^V(F_V); F_{sh;V} = E_{sh}(F_V); \quad (1)$$

Finally, a specific visible identity class $eP_{sp}^V(\cdot)$ is performed on $F_{sp;V}$ to predict the corresponding identity score $S_{sp;V}$. Meanwhile, a shared identity class $eP_{sh}(\cdot)$ is performed on $F_{sh;V}$, which outputs their predicted identity score $S_{sh;V}$. Mathematically, these processes can be expressed by

$$S_{sp;V} = P_{sp}^V(F_{sp;V}); S_{sh;V} = P_{sh}(F_{sh;V}); \quad (2)$$

Similarly, we may obtain the single-modality features F_I , infrared modality-specific features $F_{sp;I}$ and infrared modality-shared features $F_{sh;I}$ from X_I by using the $E_I(\cdot)$, $E_{sp}^I(\cdot)$ and $E_{sh}(\cdot)$, respectively. The corresponding identity scores $S_{sp;I}$ and $S_{sh;I}$ are thus obtained by using the specific infrared identity class $eP_{sp}^I(\cdot)$ and the shared identity class $eP_{sh}(\cdot)$, respectively.

Here, $E_V(\cdot)$ and $E_I(\cdot)$ follow the same structure with the first three blocks of ResNet-50 [21]. Similarly, $E_{sp}^V(\cdot)$ and $E_{sp}^I(\cdot)$ follow the same structure with the last two blocks of ResNet-50 and further attach an extra global average pooling layer, respectively. Moreover, these subnetworks' parameters are independent to each other. $E_{sh}(\cdot)$ has the same network structure with $E_{sp}^V(\cdot)$. However, its parameters are shared for single-modality visible and infrared features to extract their modality-shared features.

Loss function: To facilitate decomposing the single-modality features F_m ($m \in \{V, I\}$) into modality-specific features $F_{sp;m}$ and modality-shared features $F_{sh;m}$, a novel **Modality Decomposition (MD) loss** is further designed. As shown in Fig. 3, MD loss aims to separate modality-shared features away from those modality-specific features, and make the decomposed modality-shared features be identity-meanwhile makes those decomposed modality-specific features and modality-shared features identity-distinguishable. Therefore, the proposed MD loss consists of three items, including a decorrelation loss L_{dc} , a modality-specific feature separation loss L_{sps} and a modality-shared feature separation loss L_{shs} .

Here, $C_{sp;m}^p$ denotes the center of the p -th identity's modality-specific features. Similarly, $C_{sh;m}^p$ denotes the center of the p -th identity's modality-shared features. Then, to push the modality-specific features away from the modality-shared features, it constraints that the maximum distances among different modality-specific feature centers (e.g., l_1 in Fig. 3) should be smaller than the minimum distances from the modality-specific feature centers to the modality-shared feature centers (e.g., l_2 in Fig. 3), i.e.,

$$L_{dc} = \sum_{p=1}^P \max_d \max_k k C_{sp;V}^p - C_{sp;V}^d; k_2 + 1; 0 + \min_j k C_{sp;V}^p - C_{sh;V}^j; k_2 + 1; 0 + \max_d \max_k k C_{sp;I}^p - C_{sp;I}^d; k_2 + \min_j k C_{sp;I}^p - C_{sh;I}^j; k_2 + 1; 0; \quad (4)$$

Here, $d; j = 1; 2; \dots; P$. P_1 denotes the corresponding margin and is empirically set to 1.

As shown in the right part of Fig. 3, the modality-specific feature separation loss L_{sps} tries to separate the decomposed modality-specific features according to their identities. To this end, it enlarges the distances among the visible (infrared) modality-specific feature centers of different identities (e.g., l_1 in Fig. 3), i.e.,

$$L_{sps} = \sum_{p=1}^P \max_{j \in P} \min_{j \in P} k C_{sp;V}^p - C_{sp;V}^j; k_2; 0 + \max_{d \in P} \min_{d \in P} k C_{sp;I}^p - C_{sp;I}^d; k_2; 0; \quad (5)$$

Here, $j; d = 1; 2; \dots; P$. P_2 denotes the corresponding margin, which is empirically set to 0.7.

As shown in the left part of Fig. 3, the modality-shared feature separation loss L_{shs} tries to simultaneously distinguishable and modality-invariant. To this end, it tries to shrink the distances between the visible modality-shared feature centers and the infrared modality-shared feature centers from the same identities (e.g., l_4 in Fig. 3), and meanwhile enlarge the distances between the visible (infrared) modality-shared feature centers and the both visible

Figure 4. Distributions of the modality-shared features and modality-specific features. (a) FMCNet without MD loss. (b) FMCNet with MD loss.

and infrared modality-shared feature centers from different identities (e.g., I_3 in Fig. 3), i.e.,

$$L_{shs} = \sum_{p=1}^P \left(\kappa C_{sh;V}^p - C_{sh;I}^p \right) k_2 + \max \left(\sum_{j \in p} \min_{d \in p} \kappa C_{sh;V}^j - C_{sh;m}^j, 0 \right) + \max \left(\sum_{d \in p} \min_{j \in p} \kappa C_{sh;I}^d - C_{sh;m}^d, 0 \right) \quad (6)$$

Here, $j, d = 1; 2; \dots; P$. κ denotes the corresponding margin and is also set to 0.7. α is a predefined tradeoff parameters to balance the different losses and is set to 2.

Accordingly, the proposed MD loss is totally expressed by

$$L_{MD} = L_{shs} + \alpha_1 L_{dc} + \alpha_2 L_{sps}; \quad (7)$$

where α_1 and α_2 are the predefined tradeoff parameters to balance different losses and are both set to 0.5.

Besides, an identity classification (ID) loss is employed to facilitate extracting those person-related features and discarding those background information.

$$L_{ID} = L_{CE}(S_{sh;V}; Y_V) + L_{CE}(S_{sh;I}; Y_I) + L_{CE}(S_{sp;V}; Y_V) + L_{CE}(S_{sp;I}; Y_I); \quad (8)$$

where Y_V and Y_I denote the ground truths. Here, the ID loss is constructed by using cross-entropy loss, i.e.,

$$L_{CE}(X; Y) = -\frac{1}{N} \sum_{i=1}^N y_i \log(x_i); \quad (9)$$

where, $X = [x_1; \dots; x_N]^T$ and $Y = [y_1; \dots; y_N]^T$. x_i denotes the predicted classification score for the i -th sample, and y_i denotes the corresponding ground truth. Here, N is the total numbers of samples contained in the dataset. Therefore, the total loss L_{SFD} for training the SFD module is

$$L_{SFD} = L_{MD} + L_{ID}; \quad (10)$$

As shown in Fig. 4(a), without using the proposed MD loss, the modality-shared and modality-specific features of different identities are mixed together. While, by virtue of the proposed MD loss, those modality-shared and modality-specific features are effectively separated from each other

(e.g., Fig. 4(b)). This means that, with the collaboration of the designed network structure and the MD loss, the input single-modality features will be successfully decomposed into the modality-shared ones and modality-specific ones.

3.2. FMC Module

As discussed in the earlier part of this section, the next step in FMCNet is to directly generate those missing modality-specific information in the feature level rather than image level via the proposed FMC module. Meanwhile, as shown in Fig. 2, the process of generating the missing infrared modality-specific features $F_{sp;I}^0$ from the existing visible modality-shared features $F_{sh;V}$ is similar to that of generating the missing visible modality-specific features $F_{sh;I}^0$ from the existing infrared modality-shared features $F_{sh;I}$. We take the process of generating $F_{sp;I}^0$ from $F_{sh;V}$ as an example to detail our proposed FMC module.

Specifically, the proposed FMC module consists of a **feature-level generator** $G_{V \rightarrow I}(\cdot)$ and a **feature-level modality discriminator** $D_{V \rightarrow I}(\cdot)$. The visible modality-shared features $F_{sh;V}$ are first fed into the feature-level generator $G_{V \rightarrow I}(\cdot)$ to generate the missing (or compensated) infrared modality-specific features $F_{sp;I}^0$, i.e.,

$$F_{sp;I}^0 = G_{V \rightarrow I}(F_{sh;V}); \quad (11)$$

Here, the feature-level generator $G_{V \rightarrow I}(\cdot)$ is constructed by using three stacked fully connected layers.

Then, given the generated infrared modality-specific features $F_{sp;I}^0$ and the existing real infrared modality-specific features $F_{sp;I}$, the feature-level modality discriminator $D_{V \rightarrow I}(\cdot)$ aims to accurately distinguish the two types of modality-specific features. It is constructed by using one layer fully connected layer stacked with a Sigmoid function and outputs a classification score s for distinguishing the two types of features. The higher values of s indicates that the input features are more likely to be corresponding real infrared modality-specific features.

The feature-level generator $G_{V \rightarrow I}(\cdot)$ and the feature-level modality discriminator $D_{V \rightarrow I}(\cdot)$ are trained in an adversarial way. Concretely, $G_{V \rightarrow I}(\cdot)$ tries to fool the discriminators $D_{V \rightarrow I}(\cdot)$ by generating the missing infrared modality-specific features that approximate real infrared modality-specific features as closely as possible. While, the discriminators $D_{V \rightarrow I}(\cdot)$ tries to distinguish the generated modality-specific features and the real ones as accurately as possible. Accordingly, the generated infrared modality-specific features $F_{sp;I}^0$ will be eventually close to the real ones $F_{sp;I}$. Mathematically, the adversarial loss is defined by

$$\min_{G_{V \rightarrow I}} \max_{D_{V \rightarrow I}} L_{GAN}^{V \rightarrow I} = \frac{1}{PK} \sum_{p=1}^P \sum_{k=1}^K \left(-\log D_{V \rightarrow I}(F_{sp;I}^{k;p}) + \log [1 - D_{V \rightarrow I}(G_{V \rightarrow I}(F_{sh;V}^{k;p}))] \right); \quad (12)$$

Figure 5. Distributions of the existing modality-specific features, modality-shared features and those generated modality-specific features of two modalities.

Besides Eq. (12), the generator $G_{I_1}(\cdot)$ is also supervised by a **feature consistency loss** L_{FC}^V and an **identity consistency loss** L_{IC}^V . L_{FC}^V aims to make the generated features be close to the infrared modality-specific feature's centers of the same identities, which is expressed by

$$L_{FC}^V = \frac{1}{PK} \sum_{p=1}^P \sum_{k=1}^K \|F_{sp,l}^{(k;p)} - C_{sp,l}^p\|_1; \quad (13)$$

where $\|\cdot\|_1$ denotes the l_1 -norm of a vector or matrix.

While, L_{IC}^V enforces the generated features to be discriminative for person identification, i.e.,

$$L_{IC}^V = L_{CE}(S_{sp,l}^0; Y_I); \quad (14)$$

where $S_{sp,l}^0$ denotes the set of the predicted identity scores by feeding $F_{sp,l}^0$ into the specific infrared classifier $P_{sp,l}^I(\cdot)$.

Similarly, given the infrared modality-shared features $F_{sh,l}$, their corresponding visible modality-specific features $F_{sp,V}^0$ can be obtained in the same way by using a feature-level generator $G_{I_1,V}(\cdot)$ and a feature-level discriminator $D_{I_1,V}(\cdot)$.

Fig. 5 shows that the distributions of those visible (infrared) modality-specific features generated by FMC module are very close to those of existing visible (infrared) modality-specific features. Moreover, both the existing and the generated modality-specific features are identity-discriminable. This means that, by virtue of the proposed FMC module, the missing modality-specific information will be effectively compensated in the feature level.

3.3. SFF Module

After that, a Shared-specific Feature Fusion (SFF) module is further designed to mine the original modality-specific features and modality-shared features as well as those generated modality-specific features for VI-ReID. Here, we take the fusion of the visible modality-shared features $F_{sh,V}$, the visible modality-specific features $F_{sp,V}$ and the generated infrared modality-specific features $F_{sp,I}$ as an example to describe its steps.

In our proposed SFF module, the modality-shared features $F_{sh,V}$ are considered as the primary information for VI-ReID, while those modality-specific features $F_{sp,V}$ and

$F_{sp,I}^0$ serve as the auxiliary information. Therefore, we first combine $F_{sp,V}$ and $F_{sp,I}^0$ to obtain **fused modality-specific features** $F_{fu,V}$ via a weighted fusion way, i.e.,

$$F_{fu,V} = \lambda_1 F_{sp,V} + \lambda_2 F_{sp,I}^0; \quad (15)$$

Here, λ_1 and λ_2 are weights for $F_{sp,V}$ and $F_{sp,I}^0$, respectively, which are also learnable parameters.

Then, the modality-shared features $F_{sh,V}$ and the fused modality-specific features $F_{fu,V}$ are concatenated to obtain the final fused person features $F_{fp,V}$ of the visible images, i.e.,

$$F_{fp,V} = \text{Cat}(F_{sh,V}; F_{fu,V}); \quad (16)$$

where $\text{Cat}(\cdot)$ denotes the concatenation operation. The corresponding identity score $S_{fp,V}$ is thus obtained by feeding the fused features $F_{fp,V}$ into a shared identity classifier $P_{fp}(\cdot)$.

Similarly, the final fused person features $F_{fp,I}$ of the infrared images are obtained by fusing $F_{sh,I}$, $F_{sp,I}$ and $F_{sp,V}^0$ in the same way. The corresponding identity score $S_{fp,I}$ is thus obtained by feeding $F_{fp,I}$ into the shared identity classifier $P_{fp}(\cdot)$.

Loss function: Similar to that in Eq. (6), a **cross-modality center (MC) loss** is also employed to make the learned person features $F_{sp,V}$ and $F_{sp,I}$ be discriminative and modality-invariant, i.e.,

$$L_{MC} = \sum_{p=1}^P \|C_{fp,V}^p - C_{fp,I}^p\|_2^2 + \max_{j \in P} \min_{j \neq p} \|C_{fp,V}^p - C_{fp,m}^j\|_2^2 + \max_{d \in P} \min_{d \neq p} \|C_{fp,I}^p - C_{fp,m}^d\|_2^2; \quad (17)$$

Here, $j, d = 1, 2, \dots, P$, and $C_{fp,m}^p = \frac{1}{K} \sum_{k=1}^K F_{fp,m}^{(k;p)}$ denotes the corresponding margin and the predefined tradeoff parameter to balance different losses, which are set to 0.7 and 2, respectively. Besides, the identity classification loss is also performed on the fused features by

$$L_{FID} = L_{CE}(S_{fp,V}; Y_V) + L_{CE}(S_{fp,I}; Y_I); \quad (18)$$

Therefore, total loss for training our proposed SFF module is

$$L_{SFF} = L_{MC} + L_{FID}; \quad (19)$$

4. Experiments

4.1. Datasets and Evaluation Protocols

Two public available cross-modality VI-ReID datasets (SYSU-MM01 [1] and RegDB [31]) are employed to evaluate our model. In SYSU-MM01, 22,258 visible images and 11,909 infrared images of 395 identities are employed

Methods	SYSU-MM01								RegDB			
	All-search				Indoor-search				Visible-to-Infrared		Infrared-to-Visible	
	Single-shot		Multi-shot		Single-shot		Multi-shot					
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
HAT [17]	55.29	53.89	-	-	62.10	69.37	-	-	71.83	67.56	70.02	66.30
X-Modal [22]	49.92	50.73	-	-	-	-	-	-	62.21	50.18	-	-
MICT [23]	61.71	58.06	-	-	-	-	-	-	61.71	58.06	-	-
cm-SSFT [13]	61.60	63.20	63.40	62.00	70.50	72.60	73.00	72.40	71.00	71.30	72.30	72.90
GECNet [24]	53.37	51.83	-	-	60.60	62.89	-	-	82.33	78.45	78.93	75.58
FMI [25]	60.02	58.80	-	-	66.05	72.98	-	-	73.20	71.60	71.80	70.10
DGTL [26]	57.34	55.13	-	-	63.11	69.20	-	-	83.92	73.78	81.59	71.65
NFS [27]	56.91	55.45	63.51	48.56	62.79	69.79	70.03	61.45	80.54	72.10	77.95	69.79
CM-NAS [28]	61.99	60.02	68.68	53.45	67.01	72.95	76.48	65.11	84.54	80.32	82.57	78.31
MLC [29]	62.22	59.56	65.83	53.34	69.68	71.00	76.27	64.07	81.02	78.73	-	-
SFANet [7]	65.74	60.83	-	-	71.60	80.05	-	-	76.31	68.00	70.15	63.77
D2RL* [8]	28.90	29.20	-	-	-	-	-	-	43.40	44.10	-	-
AlignGAN* [18]	42.40	47.40	51.50	33.90	45.90	54.30	57.10	45.30	56.30	53.40	57.90	53.60
JSIA* [30]	38.10	36.90	45.10	29.50	43.80	52.90	52.70	42.70	48.10	48.90	48.50	49.30
TS-GAN* [19]	58.30	55.10	55.90	39.70	62.10	71.30	59.70	50.90	-	-	-	-
mtGAN* [9]	41.10	40.50	-	-	-	-	-	-	65.60	60.00	65.80	59.60
FMCNet*(ours)	66.34	62.51	73.44	56.06	68.15	74.09	78.86	63.82	89.12	84.43	88.38	83.86

Table 1. Comparisons with some state-of-the-art models on SYSU-MM01 dataset and RegDB dataset.

for training. While, 3803 infrared images and 301 randomly selected visible images are employed for testing. Following [13, 22], in SYSU-MM01, our model is tested under two different settings, i.e., the all-search mode and the indoor-search mode. RegDB contains 412 identities, which randomly selects 206 identities for training and the remaining 206 identities for testing. RegDB also has two test modes, including Visible-to-Infrared setting, which retrieves infrared images from visible images, and Infrared-to-Visible setting, which retrieves visible images from infrared images. The cumulative matching characteristics (CMC) [32] and mean average precision (mAP) [33] are used as evaluation metrics.

4.2. Implementation Details

We implement our model on PyTorch framework by using a single NVIDIA GeForce 2080Ti GPU. In each batch, we randomly sample 4 identities and 8 images for each identity. The SGD optimizer is adopted for training, where the momentum is set to 0.9. We set the total training epochs to 80 and set the initial learning rate to 0.1 with a warm-up strategy [34]. The learning rate decays by 0.1 at the 20th epoch and 0.01 at the 50th epoch. All the parameters in the feature extraction subnetworks are pre-trained on ImageNet [35], while other parameters are initialized by using the Kaiming initialization [36]. All the images in the training set and those in the testing set are resized into 288.

In the training stage, the input images will be augmented by randomly flipping and erasing [37].

4.3. Comparison with SOTA Methods

Our proposed FMCNet is compared with some SOTA models, including X-Modal [22], MICT [23], cm-SSFT [13], HAT [17], FMI [25], SFANet [7], NFS [27], MLC [29], GECNet [24], DGTL [38] and CM-NAS [28].

As shown in Table 1, our proposed FMCNet outperforms most existing SOTA models on SYSU-MM01. While, for RegDB, our proposed model outperforms all the SOTA models mentioned here, and meanwhile achieves close results in the infrared-to-visible mode and in the visible-to-infrared mode. This indicates the effectiveness of our proposed FMCNet. Especially, compared with those modality-shared feature learning based models, those image-level information compensation based models (e.g., D2RL [8], AlignGAN [18], JSIA [30], mtGAN [9] and TS-GAN [19]) achieve inferior results. While, our proposed feature-level information compensation based model, FMCNet, achieves significant improvements over those image-level compensation based models. This further indicates the validities of our proposed feature-level information compensation based model for VI-ReID.

Setting	Rank-1	mAP
Base	57.09	53.11
Base+SFD	63.16	58.83
Base+SFD+FMC	65.50	62.32
Base+SFD+FMC+SFF	66.34	62.51

Table 2. Evaluation of each component in our proposed model.

Setting	Rank-1	mAP
FMCNet+ L_{ID}	58.68	54.33
FMCNet+ L_{ID} + L_{shs}	63.96	58.64
FMCNet+ L_{ID} + L_{shs} + L_{sps}	64.68	59.34
FMCNet+ L_{ID} + L_{shs} + L_{sps} + L_{dc}	66.34	62.51

Table 3. Effectiveness of different items in our proposed MD loss.

Setting	Rank-1	mAP
w/o FMC	63.16	58.83
FMC+ L_{IC} + L_{FC}	57.78	55.56
FMC+ L_{GAN}	62.56	58.40
FMC+ L_{GAN} + L_{IC}	65.61	62.17
FMC+ L_{GAN} + L_{IC} + L_{FC}	66.34	62.51

Table 4. Evaluation results of different losses in FMC module.

4.4. Ablation Study

In this subsection, we evaluate each component of our proposed model on SYSU-MM01 dataset.

Effectiveness of each module: As shown in Table 2, we first remove SFD and FMC from our model as the 'Base', which just consists of $E_V(\cdot)$, $E_I(\cdot)$ and $E_{sh}(\cdot)$ in Fig. 2. Moreover, 'Base' is trained by only using ID loss. 'Base+SFD' denotes the model that employs the proposed SFD module, and is jointly trained by using MD loss and ID loss. As well, 'Base+SFD' only uses the decomposed modality-shared features for VI-ReID. 'Base+SFD+FMC' further employs the proposed FMC module for VI-ReID, where the existing modality-shared and modality-specific features and the generated modality-specific features are simply concatenated. 'Base+SFD+FMC+SFF' then attaches the proposed SFF module as the final model.

It can be seen that, compared with 'Base', 'Base+SFD' can significantly increase the performance. This indicates that the modality-shared features are well separated from the unimodal features via SFD, which greatly reduces the modality discrepancy between visible and infrared images, thus benefiting VI-ReID. The results of 'Base+SFD+FMC' indicates that the modality-specific features generated by using FMC contain much more discriminative person-related information for VI-ReID. Finally, the results of 'Base+SFD+FMC+SFF' shows that the proper

exploitation of those existing and generated features further boosts the performance of VI-ReID.

Verifying the effectiveness of each item in the proposed MD loss: As shown in Table 3, the modality-shared feature separation loss L_{shs} can significantly improve our model's performance. This indicates that L_{shs} can effectively reduce the modality discrepancy between the modality-shared visible and infrared features. Similarly, the modality-specific feature separation loss L_{sps} can further boost our model's performance, by separating different modality-specific features from each other according to their identities. Moreover, the decorrelation loss L_{dc} also increases our model's performance. This means that the modality-shared and modality-specific features are well separated from each other with L_{dc} , which further benefits the subsequent missing modality-specific feature compensation and VI-ReID.

Verifying the effectiveness of different loss functions in the proposed FMC module: In Table 4, 'w/o FMC' means only using the decomposed modality-shared features for VI-ReID. 'FMC' means using FMC module to compensate missing modality-specific features. 'FMC+ L_{FC} ' degrades rather than increases the performance. This indicates that, without using the adversarial loss, the missing modality-specific features cannot be well generated, thus leading to performance drops. On the contrary, with the adversarial loss L_{GAN} (i.e., 'FMC+ L_{GAN} '), FMC ensures the similarity between the generated modality-specific features and real ones. With the proposed identity consistency loss L_{IC} and feature consistency loss L_{FC} , the discriminability of those generated modality-specific features is further enhanced, thus leading to performance improvements.

5. Conclusion

In this paper, our proposed FMCNet invests the feature-level rather than image-level modality-specific information compensation for VI-ReID, which is achieved by using the proposed SFD, FMC and SFF modules. Compared with that in image level, the proposed modality-specific information compensation in feature level avoids the introduction of interfering information, and meanwhile is able to explicitly generate more discriminative person-related modality-specific features, thus effectively boosting the performance of VI-ReID. The experimental results demonstrate that our approach significantly outperforms existing image-level modality-specific information compensation based models. Moreover, it even achieves better results than some SOTA modality-shared feature learning based models.

Limitation and Societal Impact: The missing modality-specific features compensated by our model still lack spatial structure information, which may be further beneficial for VI-ReID. Moreover, all used datasets are publicly available and involve no ethical issues.

References

- [1] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5380–5389. **1, 3, 6**
- [2] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 1092–1099. **1**
- [3] Y. Hao, N. Wang, L. Jie, and X. Gao, "HSME: Hyper-sphere manifold embedding for visible thermal person re-identification," pp. 8385–8392, 2019. **1**
- [4] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 407–419, 2020. **1, 3**
- [5] Z. Wei, X. Yang, N. Wang, and X. Gao, "Flexible body partition-based adversarial learning for visible infrared person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021. **1, 3**
- [6] J. Sun, Y. Li, H. Chen, Y. Peng, X. Zhu, and J. Zhu, "Visible-infrared cross-modality person re-identification based on whole-individual training," *Neurocomputing*, vol. 440, pp. 1–11, 2021. **1, 3**
- [7] H. Liu, S. Ma, D. Xia, and S. Li, "Sfanet: A spectrum-aware feature augmentation network for visible-infrared person re-identification," *arXiv preprint arXiv:2102.12137*, 2021. **1, 3, 7**
- [8] Z. Wang, Z. Wang, Y. Zheng, Y. Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, pp. 618–626. **2, 3, 7**
- [9] X. Fan, W. Jiang, H. Luo, and W. Mao, "Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal person re-identification," *The Visual Computer*, pp. 1–16, 2020. **2, 3, 7**
- [10] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," pp. 10 257–10 266, 2020. **2**
- [11] K. Kansal, A. Subramanyam, Z. Wang, and S. Satoh, "Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3422–3432, 2020. **2**
- [12] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 579–590, 2019. **2, 3**
- [13] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, "Cross-modality person re-identification with shared-specific transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 379–13 389. **2, 7**
- [14] M. Ye, J. Shen, G. Lin, T. Xiang, and S. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2021. **2**
- [15] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1092–1108, 2019. **2**
- [16] D. Wu, S. J. Zheng, X. P. Zhang, C. A. Yuan, F. Cheng, Y. Zhao, Y. J. Lin, Z. Q. Zhao, Y. L. Jiang, and D. S. Huang, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, no. APR.14, pp. 354–371, 2019. **2**
- [17] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 728–739, 2020. **3, 7**
- [18] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3623–3632. **3, 7**
- [19] Z. Zhang, S. Jiang, C. Huang, Y. Li, and R. Y. Da Xu, "Rgb-ir cross-modality person reid based on teacher-student gan model," *arXiv preprint arXiv:2007.07452*, 2020. **3, 7**
- [20] Y. Yang, T. Zhang, J. Cheng, Z. Hou, P. Tiwari, H. M. Pandey et al., "Cross-modality paired-images generation and augmentation for RGB-Infrared person re-identification," *Neural Networks*, vol. 128, pp. 294–304, 2020. **3**
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 770–778. **4**
- [22] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020, pp. 4610–4617. **7**
- [23] X. Hu and Y. Zhou, "Cross-modality person reid with maximum intra-class triplet loss," in *Pattern Recognition and Computer Vision*, 2020, pp. 557–568. **7**
- [24] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C.-W. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021. **7**
- [25] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, "Farewell to mutual information variational distillation for cross-modal person re-identification," in *Computer Vision and Pattern Recognition*, 2021, pp. 1522–1531. **7**
- [26] H. Liu, Y. Chai, X. Tan, D. Li, and X. Zhou, "Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification," *IEEE Signal Processing Letters*, vol. 28, pp. 653–657, 2021. **7**
- [27] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, "Neural feature search for rgb-infrared person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 587–597. **7**

- [28] C. Fu, Y. Hu, X. Wu, H. Shi, T. Mei, and R. He, "Cm-nas: Rethinking cross-modality neural architectures for visible-infrared person re-identification," *ArXiv*, vol. abs/2101.08467, 2021. 7
- [29] Z. Sun, Y. Zhu, S. Song, J. Hou, S. Du, and Y. Song, "The multi-layer constrained loss for cross-modality person re-identification," *Proceedings of the International Conference on Artificial Intelligence and Signal Processing*, pp. 1–6, 2020. 7
- [30] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z. Hou, "Cross-modality paired-images generation for rgb-infrared person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 144–12 151. 7
- [31] N. Dat, H. Hong, K. Ki, and P. Kang, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017. 6
- [32] H. Moon and P. J. Phillips, "Computational and performance aspects of pca-based face-recognition algorithm," *Recognition*, vol. 30, no. 3, pp. 303–21, 2001. 7
- [33] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) December 2015*. 7
- [34] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2020. 7
- [35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 1–42, 2014. 7
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 7
- [37] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, 2017. 7
- [38] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling and a strong convolutional baseline," *Proceedings of the European Conference on Computer Vision*, 2018, pp. 501–518. 7