Shape-Erased Feature Learning for Visible-Infrared Person Re-Identification

Jiawei Feng[1]   Ancong Wu[1*]   Wei-Shi Zheng[1,2,3]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[3]Guangdong Key Laboratory of Information Security Technology, China

fengjw3@mail2.sysu.edu.cn, wuanc@mail.sysu.edu.cn, wszheng@ieee.org

Abstract

*Due to the modality gap between visible and infrared images with high visual ambiguity, learning **diverse** modality-shared semantic concepts for visible-infrared person re-identification (VI-ReID) remains a challenging problem. Body shape is one of the significant modality-shared cues for VI-ReID. To dig more diverse modality-shared cues, we expect that erasing body-shape-related semantic concepts in the learned features can force the ReID model to extract more and other modality-shared features for identification. To this end, we propose shape-erased feature learning paradigm that decorrelates modality-shared features in two orthogonal subspaces. Jointly learning shape-related feature in one subspace and shape-erased features in the orthogonal complement achieves a conditional mutual information maximization between shape-erased feature and identity discarding body shape information, thus enhancing the diversity of the learned representation explicitly. Extensive experiments on SYSU-MM01, RegDB, and HITSZ-VCM datasets demonstrate the effectiveness of our method.*

1. Introduction

Recently, person re-identification (ReID) for pedestrian matching in non-overlapping camera views has experienced fast development. However, ReID is still challenging when people appear both in the daytime and in low-light situations where only infrared cameras can clearly capture their appearances, raising the task of visible-infrared ReID (VI-ReID). Many remarkable works [4, 5, 14, 15, 27, 30] have been witnessed in the field of VI-ReID. For realistic scenarios, discovering *rich and diverse* modality-shared semantic concepts usually helps to improve the effectiveness of VI-ReID [31, 39]. So far, *diverse* modality-shared feature learning remains challenging.
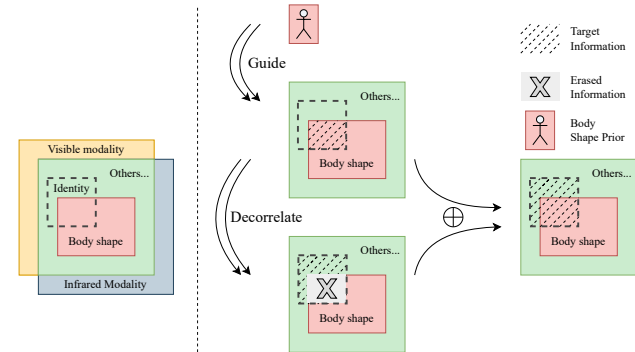
---

*Corresponding author

Figure 1. An illustration of our motivation on VI-ReID. It is assumed that body shape information and identity-related modality-shared information (presented in dashed box) are partially overlapped with each other. To make extracted features more *diverse*, we propose shape-erased feature learning paradigm that decomposes the representation into shape-related feature and shape-erased one. Learning shape-erased feature drives the model to discover richer modality-shared semantic concepts other than body shape.

Among the cues for VI-ReID, we can identify pedestrians by their body shapes in many situations, for it contains modality-invariant information and also robust to light changes. Nevertheless, body shape is not the only or a sufficient semantic concept that interprets the identity of a person. It may be hard in some situations to tell the difference only depending on the body shape, but we can still distinguish them by other semantic concepts, such as their belongings, hairstyles or face structures. Inspired by this, we illustrate an information theoretic measure between visible and infrared modality as a Venn diagram on the left of the dashed line in Fig. 1. It is assumed that body shape (presented in red) and identity-related modality-shared information (presented in dashed box) are partially overlapped with each other. Note that *partially* is also due to there exists identity-unrelated information contained in body shape map, *e.g.*, human pose. This partially overlapped assumption indicates that the target information for

VI-ReID, which is identity-related and modality-shared, can be divided into two independent components that are related and unrelated to body shape.

Based on the above observation and assumption, to dig more diverse modality-shared cues for VI-ReID, we expect to erase the body-shape-related semantic concepts in the features to force the VI-ReID model to extract more and other modality-shared features for identification. As illustrated on the right of the dashed line in Fig. 1, the *shape-erased* feature is decorrelated from the *shape-related* feature to simultaneously discover shape-unrelated knowledge, while *shape-related* feature can be explicitly guided by some given body shape prior, which is easy to obtain by existing pre-trained human parsing models [16]. In this way, both *shape-related* and *shape-erased* features are explicitly quantified while the discriminative nature of the two features can be independently maintained.

Specifically, we propose shape-erased feature learning paradigm that introduces orthogonality into representation to satisfy a relaxation of independent constraint. The representation is then decomposed into two sub-representations lying in two orthogonal subspaces for *shape-related* and *shape-erased* feature learning, respectively. By learning and covering most discriminative body shape feature in one subspace, the *shape-erased* feature is forced to discover other modality-shared discriminative semantic concepts in the the other subspace as *shape-related* feature is constrained in its orthogonal complement. Under the above assumptions, we formulate this shape-erased feature learning paradigm from a mutual information perspective, and demonstrate that jointly learning *shape-erased* and *shape-related* objectives achieves a conditional mutual information maximization between *shape-erased* feature and identity discarding body shape information, thus enhancing the diversity of the learned representation explicitly. We finally design a Shape-Guided dIverse fEature Learning (SGIEL) framework that jointly optimizes *shape-related* and *shape-erased* objectives to learn modality-shared and discriminative integrated representation. The contributions of our work are summarized as follows:

- We propose a shape-erased feature learning paradigm for VI-ReID that decorrelates *shape-erased* feature from *shape-related* one by orthogonal decomposition. *Shape-related* feature in one subspace is guided by body shape prior while *shape-erased* feature is constrained in its orthogonal complement to discover more and other modality-shared discriminative semantic concepts, thus enhancing the diversity of the learned representation explicitly.

- Based on the proposed shape-erased feature learning paradigm, we design a Shape-Guided dIverse fEature Learning framework that jointly optimizes *shape-*

related and *shape-erased* objectives to learn modality-shared and discriminative integrated representation.

- Extensive experiments on SYSU-MM01, RegDB, and HITSZ-VCM datasets demonstrate the effectiveness of our method.

2. Related Work

2.1. Visible-Infrared Person Re-Identification

To alleviate visible-infrared modality discrepancy and discover modality-shared discriminative features, researchers contributed numerous significant works in different levels of VI-ReID framework. Specifically, in the feature learning level, Kansal *et al*. [14] designed a model to disentangle spectrum information and extract identity discriminative features to make cross-modal learning more efficient. Wu *et al*. [29] exploited the same-modality similarity as a constraint to guide the learning of cross-modality similarity along with the alleviation of modality-specific information. Zhao *et al*. [38] designed a model which both learned the color-irrelevant features and aligned the identity-level feature distributions. Zhang *et al*. [35] proposed FMCNet to compensate the missing modality-specific information in the feature level.In the input level, Ye *et al*. [32] proposed a channel augmented joint learning strategy to improve the robustness against cross-modal variations. Wei *et al*. [28] introduced syncretic modality generative module to produce a new modality incorporating cross-modal features. In the model architecture level, most researchers implemented their frameworks in a one-stream [6] or two-stream [5, 24] manner, while some other researchers Li *et al*. [15] claimed that an auxiliary modality was required as an assistant to bridge the huge gap between the two modalities. Two existing related works consider a similar challenge including PAENet [39] and MPANet [31]. PAENet introduced eight attributes as annotations to learn the fine-grained semantic attribute information, which is a labeling-intensive journey in normal situations. Wu *et al*. [31] proposed a joint Modality and Pattern Alignment Network to discover cross-modality nuances in different local patterns, however, they only add diverse constraints on the different activation maps and directly combine them together, which would fall into trivial solutions if one feature is enough to deal with the training set.

Most existing VI-ReID approaches concentrated on reducing modality-specific features or converting them to modality-shared ones, without explicitly discovering diverse modality-shared features. Compared to previous works, we decorrelates *shape-erased* feature from *shape-related* one by orthogonal decomposition and by joint learning *shape-erased* and *shape-related* objectives, the diversity of the learned representation is explicitly enhanced.
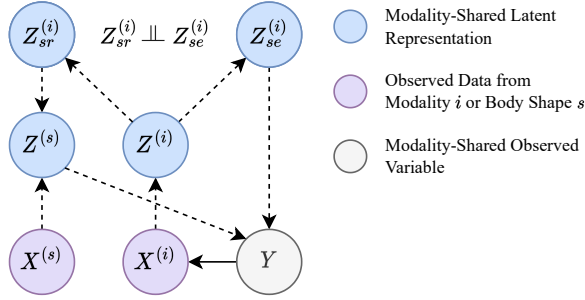
Figure 2. Graphical model of our method. The superscript $^{(s)}$ denotes body shape, and $^{(i)}$, where $i = 1, 2$, denotes different modalities (1(2) for visible(infrared)). $Z_{sr}^{(i)}$ and $Z_{se}^{(i)}$ are assumed to be independent to explicitly quantify shape-related and shape-erased features.

2.2. Semantic Parsing for Person Re-Identification

Kalayeh *et al.* [13] proposed the first work in person Re-ID community to adopt human semantic parsing to precisely localize arbitrary contours of various body parts. They claimed that compared to detection based method, semantic segmentation based method exhibited better pixel-level accuracy and capability of modeling arbitrary contours. Song *et al.* [25] first introduced the binary segmentation masks to construct synthetic RGB-Mask pairs as inputs, and designed a mask-guided contrastive attention model to learn features separately from the body and background regions. Guo *et al.* [8] applied a human parsing model to extract the binary human part masks and a self-attention mechanism to capture the soft latent (non-human) part masks. Zhu *et al.* [40] proposed the Identity-Guided Human Semantic Parsing approach to locate both the human body parts and personal belongings at pixel-level for person re-ID only with person identity labels. Hong *et al.* [11] proposed a fine-grained shape-appearance mutual learning framework to learn fine-grained discriminative body shape knowledge to complement the cloth-unrelated knowledge in the appearance features.

For VI-ReID problem, one related work was contributed by Huang *et al.* [12]. They considered using person mask prediction as an auxiliary task with the help of a pre-trained human parsing model. In comparison to their work, we aim to discover richer modality-shared discriminative features in each subspace by erasing the body shape information to decorrelate shape-erased and shape-related features.

3. Shape-Guided Diverse Feature Learning

In this section, we first present preliminaries of our method, and then describe our proposed shape-erased feature learning paradigm, and finally introduce Shape-Guided dIverse fEature Learning (SGIEL) framework.

3.1. Preliminary

VI-ReID Setup. Consider random variables $X^{(i)}$ and Y representing data and label of VI-ReID, where $i = 1$ for visible modality and $i = 2$ for infrared modality. The observed values of $X^{(i)}$ and Y are used to build a dataset $D = \{D^{(i)}\}_{i=1}^2$, where $D^{(i)} = \{x_j^{(i)}, y_j\}_{j=1}^{N_i}$. Samples of each modality are collected from the same group of C persons, but the number of each identity's samples for each modality may arbitrary. Let f and g denote image encoder and classifier, the goal of VI-ReID is to learn an f to extract representation $z^{(i)} = f(x^{(i)}) \in \mathbb{R}^n$ invariant to different modalities and different camera views.

Body Shape Data. We borrowed pre-trained Self-Correction Human Parsing (SCHP) model proposed in [16] to segment body shape from background. Given a pixel of an image, we directly summed the probabilities of being a part of the head, torso, or limbs, predicted by SCHP, to create the body-shape map. Specifically, for each sample $x^{(i)}$ from dataset D, either visible or infrared, we used SCHP to produce its paired body shape map $x^{(s)}$ with the same image size and label, *i.e.*, it is a one-to-one mapping between D and its corresponding body shape data. Let f_s and g_s denote body shape map encoder and classifier, the latent representation of $x^{(s)}$ is $z^{(s)} = f_s(x^{(s)}) \in \mathbb{R}^m$, $m < n$.

3.2. Shape-Erased Feature Learning Paradigm

In this section, we first explain the key independent assumption for explicitly quantifying *shape-related* and *shape-erased* features, and a relaxation to approximate it. Based on this relaxed independent constraint, we introduce the proposed Shape-Erased Feature Learning.

3.2.1 Independence between $Z_{sr}^{(i)}$ and $Z_{se}^{(i)}$

We first formulate the main design of our shape-erased feature learning paradigm as a graphical model illustrated in Fig. 2. It is assumed that modality-shared *shape-related* feature, $Z_{sr}^{(i)}$, and modality-shared *shape-erased* feature, $Z_{se}^{(i)}$, are independent from each other, and derived from an integrated representation $Z^{(i)}$ extracted in $X^{(i)}$, *i.e.*, $Z^{(i)} \rightarrow Z_{sr}^{(i)}$, $Z^{(i)} \rightarrow Z_{se}^{(i)}$. The independence between the two components, $Z_{sr}^{(i)} \perp\!\!\!\perp Z_{se}^{(i)}$, is necessary for learning any two features simultaneously without affecting each other. We formulate this independence as the following Eq. (1),

$$I(Z_{sr}^{(i)}; Z_{se}^{(i)}) = 0, \quad (1)$$

where $I(\cdot; \cdot)$ denotes mutual information. As the mutual information estimation is complex and time-consuming, we relax the independence as an orthogonal constraint, and per-
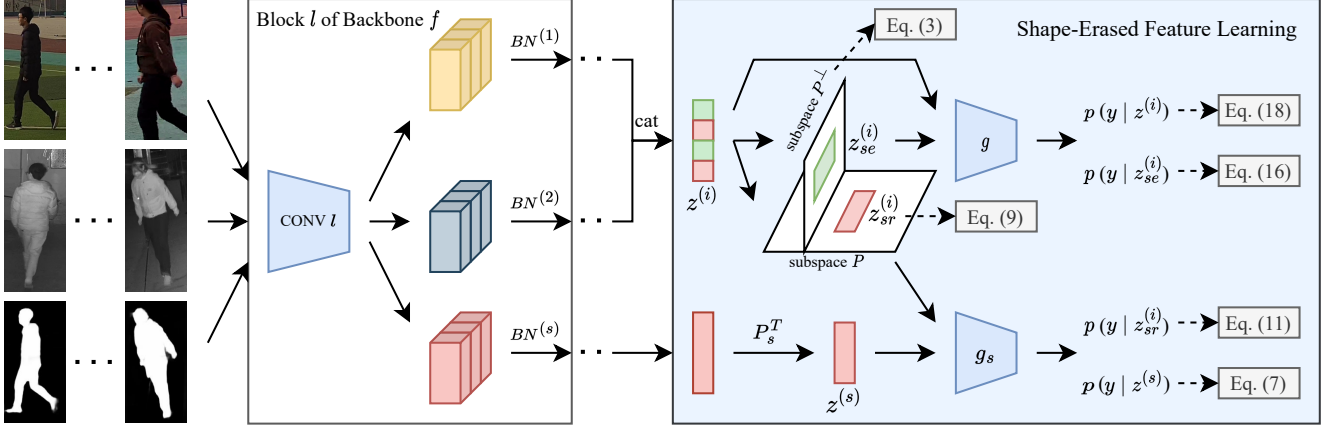
Figure 3. Shape-Guided Diverse Feature Learning. We utilize one shared backbone for visible, infrared images, and their body shape maps, while only BN layers are view-specific; "cat" refers to concatenating $z^{(i)}$ along batch dimension, where $i = 1(2)$ for visible(infrared); In shape-erased feature learning paradigm, by regularizing P to be semi-orthogonal (Eq. (3)), we decompose $z^{(i)}$ into *shape-related* $z_{sr}^{(i)}$ and *shape-erased* $z_{se}^{(i)}$. $z_{sr}^{(i)}$ is learned to imitate and cover discriminative body shape features in subspace P (Eq. (9) and (11)), while $z_{se}^{(i)}$ is decorrelated to mine other modality-shared discriminative features in subspace $P^\perp$ (Eq. (16)).

form orthogonal decomposition to achieve the relaxed version of Eq. (1) as:

$$z_{sr}^{(i)} = P^T z^{(i)},$$
$$z_{se}^{(i)} = (I_n - PP^T) z^{(i)}, \qquad (2)$$

where $P \in \mathbb{R}^{n \times m} (m < n)$ denotes a semi-orthogonal matrix and PP^T forms an orthogonal projector. In this way, *shape-related* feature is learned in subspace P while *shape-erased* features is learned in the orthogonal complement $P^\perp$, approximately satisfying the independent constraint. In practice, as P is usually initialized by standard normal distribution, if $n \to \infty$, the probability that P becomes a semi-orthogonal matrix goes to 1. To further enhance this orthogonality, we regularize P by L^1-norm on the difference of each dimension between $P^T P$ and identity matrix I_m by Eq. (3):

$$\mathcal{L}_{ortho} = \frac{1}{m} \sum_{j=1}^{m} \|(P^T P)_j - (I_m)_j\|_1. \qquad (3)$$

3.2.2 Shape-Erased Feature Learning

As introduced in Section 1, we aim to explicitly quantify $Z_{sr}^{(i)}$ and $Z_{se}^{(i)}$ so that $Z_{se}^{(i)}$ can infer identity Y when discarding information used to describe $X^{(s)}$. This can be formulated as maximizing conditional mutual information between $Z_{se}^{(i)}$ and Y given body shape $X^{(s)}$, *i.e.*, $I(Z_{se}^{(i)}; Y | X^{(s)})$:

$$\max I(Z_{se}^{(i)}; Y | X^{(s)}) = I(Z_{se}^{(i)}; Y) - I(Z_{se}^{(i)}; Y; X^{(s)}), \qquad (4)$$

where the first term represents mutual information between $Z_{se}^{(i)}$ and Y, and the second represents mutual information between $Z_{se}^{(i)}, Y$ and $X^{(s)}$.

Maximize $I(Z_{se}^{(i)}; Y)$. To optimize Eq. (4), we can maximize the first term $I(Z_{se}^{(i)}; Y)$ by minimizing cross-entropy $(l_{ce}(q, p) = -\sum_{k=1}^{C} p_k \log q_k)$ as Eq. (5),

$$\mathcal{L}_{seid} = \mathbb{E}_{(z_{se}^{(i)}, y) \sim (Z_{se}^{(i)}, Y)} l_{ce}(g(z_{se}^{(i)}), y), \qquad (5)$$

In the following, we will discuss how to estimate and minimize the second term $I(Z_{se}^{(i)}; Y; X^{(s)})$.

Minimize $I(Z_{se}^{(i)}; Y; X^{(s)})$. Since $I(Y; X^{(s)})$ is intractable, we approximate it by the following two steps.

(1) Approximate $I(Z_{se}^{(i)}; Y; X^{(s)})$ by $I(Z_{se}^{(i)}; Y; Z^{(s)})$

Firstly, we consider a requirement that a representation Z of X can describe Y at least as well as using the original data X instead. This requirement is known as *sufficiency* [1] that can be defined as follows:

Definition 1 (Sufficiency). *A representation Z of X is sufficient for Y if and only if:*

$$I(X; Y | Z) = 0 \iff I(X; Y) = I(Z; Y). \qquad (6)$$

For $Z^{(s)}$, if the classification loss $\mathcal{L}_{sid}$ is minimized,

$$\mathcal{L}_{sid} = \mathbb{E}_{(z^{(s)}, y) \sim (Z^{(s)}, Y)} l_{ce}(g_s(z^{(s)}), y), \qquad (7)$$

then following [2], we can assume $Z^{(s)}$ of $X^{(s)}$ for Y is *sufficient*, and thus we replace $I(Z_{se}^{(i)}; Y; X^{(s)})$ with

$I(Z_{se}^{(i)}; Y; Z^{(s)})$ in Eq. (4) combining Eq. (6). This replacement can be formulated as the following **Theorem 1**.

Theorem 1 *If representation $Z^{(s)}$ of $X^{(s)}$ is **sufficient** for Y, then $I(Z_{se}^{(i)}; Y; X^{(s)}) = I(Z_{se}^{(i)}; Y; Z^{(s)})$.*

The proof can be found in supplementary material.

(2) Approximate $I(Z_{se}^{(i)}; Y; Z^{(s)})$ by $I(Z_{se}^{(i)}; Y; Z_{sr}^{(i)})$

Secondly, we hope shape-related feature $Z_{sr}^{(i)}$ can fully represent real body shape feature $Z^{(s)}$, so that if $Z_{sr}^{(i)} \equiv Z^{(s)}$, then $I(Z_{se}^{(i)}; Y; Z^{(s)})$ can also be approximated by $I(Z_{se}^{(i)}; Y; Z_{sr}^{(i)})$:

$$
\begin{aligned}
I(Z_{se}^{(i)}; Y; X^{(s)}) &= I(Z_{se}^{(i)}; Y; Z^{(s)}) \\
&= I(Z_{se}^{(i)}; Y; Z_{sr}^{(i)}) \le I(Z_{se}^{(i)}; Z_{sr}^{(i)}) = 0,
\end{aligned}
\tag{8}
$$

which is upper-bounded by Eq. (1). To achieve a $Z_{sr}^{(i)}$ fully representing $Z^{(s)}$, as there exists a one-to-one mapping between $Z_{sr}^{(i)}$ and $Z^{(s)}$, we maximize $I(Z_{sr}^{(i)}; Z^{(s)})$ by minimizing element-wise mean squared error (MSE) to guide $Z_{sr}^{(i)}$ to imitate $Z^{(s)}$ as Eq. (9),

$$
\mathcal{L}_{srmse} = \mathbb{E}_{(z_{sr}^{(i)}, z^{(s)}) \sim (Z_{sr}^{(i)}, Z^{(s)})} \frac{\|z_{sr}^{(i)} - z^{(s)}\|_2^2}{m}, \tag{9}
$$

where $\| \cdot \|_2$ denotes l^2-norm.

Moreover, to reduce cross-view discrepancy between $Z_{sr}^{(i)}$ of $X^{(i)}$ and $Z^{(s)}$ of $X^{(s)}$, we aim to minimize the following conditional mutual information $I(X^{(i)}; Z_{sr}^{(i)} | X^{(s)})$:

$$
\min I(X^{(i)}; Z_{sr}^{(i)} | X^{(s)}), \tag{10}
$$

denoting the remaining information in $Z_{sr}^{(i)}$ given the view of $X^{(s)}$. To minimize Eq. (10), we follow [26] to approximate an upper bound of it as a Kullback–Leibler (KL) divergence between $p(y|z_{sr}^{(i)})$ and $p(y|z^{(s)})$. The proof of that can be found in supplementary material. For simplicity, we directly minimize cross-entropy loss as Eq. (11) for the remaining information entropy term in KL divergence only depending on target distribution $p(y|z^{(s)})$,

$$
\mathcal{L}_{srkl} = \mathbb{E}_{(z_{sr}^{(i)}, z^{(s)}) \sim (Z_{sr}^{(i)}, Z^{(s)})} l_{ce}(g_s(z_{sr}^{(i)}), g_s(z^{(s)})). \tag{11}
$$

Combining Eq. (9) and (11), the final shape-related objective becomes:

$$
\mathcal{L}_{sr} = \mathcal{L}_{srmse} + \mathcal{L}_{srkl}. \tag{12}
$$

Minimizing Eq. (12), we can represent $Z^{(s)}$ by $Z_{sr}^{(i)}$ approximately. If both *sufficiency* of $Z^{(s)}$ is achieved and Eq. (12) is minimized, then,

$$
I(Z_{se}^{(i)}; Y | X^{(s)}) \ge I(Z_{se}^{(i)}; Y) \tag{13}
$$

will hold by Eq. (8). In this way, *shape-erased* features can be learned by minimizing classification loss in Eq. (5) as information used to describe discriminative body shape feature are approximately discarded by orthogonal decomposition in Eq. (2).

Eliminate Modality-Specific Information. It is to be noted that both $Z_{sr}^{(i)}$ and $Z_{se}^{(i)}$ are assumed to be shared features between the two modalities. $Z_{sr}^{(i)}$ is learned to imitate body shape representation $Z^{(s)}$ to be modality-shared naturally; For $Z_{se}^{(i)}$, we eliminate modality-specific information in a mutual manner as follows:

$$
\min I(X^{(1)}; Z_{se}^{(1)} | X^{(2)}) + I(X^{(2)}; Z_{se}^{(2)} | X^{(1)}). \tag{14}
$$

Similar to Eq. (11), we approximated it as a cross-modal cross-entropy between $p(y|z_{se}^{(1)})$ and $p(y|z_{se}^{(2)})$ and vice versa:

$$
\mathcal{L}_{sekl} = \mathbb{E}_{(z_{se}^{(i)}, z_{se}^{(3-i)}) \sim (Z_{se}^{(i)}, Z_{se}^{(3-i)})} l_{ce}(g(z_{se}^{(i)}), g(z_{se}^{(3-i)})), \tag{15}
$$

where $i = 1, 2$.

Combining Eq. (5) and (15), the shape-erased objective can be formulated as:

$$
\mathcal{L}_{se} = \mathcal{L}_{seid} + \mathcal{L}_{sekl}. \tag{16}
$$

3.3. Overall Framework

In Section 3.2, we decompose representation $z^{(i)}$ into two orthogonal components named *shape-related* $z_{sr}^{(i)}$ and *shape-erased* $z_{se}^{(i)}$. To further enhance the discriminative and modality-shared natures of $z^{(i)}$, we apply commonly used classification loss $\mathcal{L}_{id}$ and triplet loss [10] $\mathcal{L}_{triplet}$ on $z^{(i)}$. For triplet pairs, we find the hardest positive and negative pairs among all samples in a mini-batch, consisting of both visible and infrared samples.

Similar to $\mathcal{L}_{sekl}$, we apply the following Eq. (17) for eliminating cross-modal discrepancy in a mutual way:

$$
\mathcal{L}_{kl} = \mathbb{E}_{(z^{(i)}, z^{(3-i)}) \sim (Z^{(i)}, Z^{(3-i)})} l_{ce}(g(z^{(i)}), g(z^{(3-i)})). \tag{17}
$$

Combining $\mathcal{L}_{id}$, $\mathcal{L}_{triplet}$ and (17), the integrated representation objective can be formulated as:

$$
\mathcal{L}_{int} = \mathcal{L}_{id} + \mathcal{L}_{triplet} + \mathcal{L}_{kl}. \tag{18}
$$

Moreover, we implement a re-weighting mechanism to focus on more difficult objective between $\mathcal{L}_{sr}$ and $\mathcal{L}_{se}$ during the whole training process. Let θ_t denote parameters to be optimized at training iteration t-th, we measure this difficulties by comparing the norms of $\frac{\partial \mathcal{L}_{sr}(\theta_t)}{\partial \theta_t}$ and $\frac{\partial \mathcal{L}_{se}(\theta_t)}{\partial \theta_t}$. The objective with a larger gradient norm is regarded to be more difficult at iteration t-th. To save computation cost,

following [18, 20], we approximate the actual parameter-level gradients by the representation-level gradients, *i.e.*, replacing $\frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t}$ with $\frac{\partial \mathcal{L}(\theta_t)}{\partial z^{(i)}}$. Our final re-weighting mechanism is as follows:

$$
\begin{aligned}
\alpha_t^{sr} &= \|\frac{\partial \mathcal{L}_{sr}(\theta_t)}{\partial z^{(i)}}\|_2^2 / (\|\frac{\partial \mathcal{L}_{sr}(\theta_t)}{\partial z^{(i)}}\|_2^2 + \|\frac{\partial \mathcal{L}_{se}(\theta_t)}{\partial z^{(i)}}\|_2^2), \\
\alpha_t^{se} &= \|\frac{\partial \mathcal{L}_{se}(\theta_t)}{\partial z^{(i)}}\|_2^2 / (\|\frac{\partial \mathcal{L}_{sr}(\theta_t)}{\partial z^{(i)}}\|_2^2 + \|\frac{\partial \mathcal{L}_{se}(\theta_t)}{\partial z^{(i)}}\|_2^2),
\end{aligned} \quad (19)
$$

The overall training loss can be summarized as the following Eq. (20) :

$$
\mathcal{L}_{train} = \mathcal{L}_{int} + \alpha_t^{sr}\mathcal{L}_{sr} + \alpha_t^{se}\mathcal{L}_{se} + \mathcal{L}_{ortho} + \mathcal{L}_{sid}. \quad (20)
$$

The overall framework of our method is illustrated in Fig. 3. To reduce the computation and GPU memory consumption, one modeling backbone is shared for three types of data (two modalities and body shape map). Considering the distribution gap among them, similar to AdaBN [17], for each Batch Normalization (BN) layer in the backbone, we implement three new parameter-specific BNs to replace it as a normalization across different distributions. Following the design of BNNeck by [21], we perform three parameter-specific BNNecks after the backbone.

4. Experiments

To validate the effectiveness of our method, we conducted experiments on two benchmark datasets, SYSU-MM01 [30] and RegDB [22]. We also evaluated our method on a large-scale video-based VI-ReID dataset named HITSZ-VCM [19]. Furthermore, we performed ablation study to validate the effectiveness of each component in our method.

4.1. Dataset and Evaluation Protocol

SYSU-MM01 [30] contains four(two) cameras for capturing visible(infrared) images. The query and gallery set are collected from another 96 identities containing 3,803 infrared images and 301 randomly sampled visible images under single-shot setting. RegDB [22] is captrured by one pair of visible and thermal cameras. It is collected from 412 identities. The identities are randomly and equally split into a training set and a testing set. RegDB has two testing protocols: infrared to visible, which means infrared serves as query, and visible to infrared, meaning visible as query. HITSZ-VCM [19] is the first large-scale video-based VI-ReID dataset captured by 6 visible and 6 infrared cameras. There are 251,452 visible images and 211,807 infrared images. Every 24 consecutive images are regarded as a tracklet. 927 identities are divided into 500 for training and 427 for testing. The testing set contains 5,159 infrared and 5,643 visible tracklets. The testing protocol is similar to RegDB's but in a *tracklet-to-tracklet* way.

To quantitatively evaluate the performance of our proposed model, Cumulative Matching Characteristic curve (CMC) and mean Average Precision(mAP) are adopted as the evaluation metrics on all three datasets.

4.2. Implementation Details

Following [32, 34], we adopted an ImageNet pre-trained ResNet50 [9] as our backbone, and replaced the average pooling layer with GEM-pooling similar to [34]. We adopted Random Channel Exchangeable Augmentation and Channel-Level Random Erasing proposed in [32] and exponential moving average (EMA) model, similar to [7], for our method and baseline. Only the EMA model is used for testing. We replaced $z^{(s)}$ in Eq. (11) with a temporal ensemble $z^{(s)}$ produced by EMA model. The dimension $m(n)$ of P and P_s is set to $512(2048)$. The three sets of BNs were initialized to be identical to each other.

We used SGD optimizer with initial learning rate of 0.1 for randomly initialized parameters and 0.01 for pre-trained parameters. For SYSU-MM01 and RegDB dataset, we trained the model 100 epochs and decreased the learning rate by a factor of 10 at the 20-th and 50-th epoch. We randomly sampled 8 identities, each of which involved 4 visible and 4 infrared images to build a mini-batch. For HITSZ-VCM dataset, we trained the model 200 epochs and decreased the learning rate by a factor of 10 at the 35-th and 80-th epoch. We randomly sampled 8 identities, each of which involved 2 visible and 2 infrared tacklets, and randomly sampled 3 frames per tracklet to build a mini-batch.

4.3. Comparison with State-of-the-Art Methods

Comparison on SYSU-MM01 and RegDB. We compared our method with existing state-of-the-art methods for VI-ReID, including CM-NAS [6], CAJL [32], MPANet [31], MMN [36], MTL [12], PAENet [39], MSCLNet [37]. For SYSU-MM01 in Table 1, "Params" indicates which times of ResNet50 parameter size the backbone is. To achieve a comparable size with "≈2x params" methods, we directly concatenated the output representations of two checkpoints of our model. Our method outperformed most models with comparable size of parameters according to Rank-1 and mAP metrics. Note that PAENet introduced eight extra fine-grained attributes as annotations for each sample. For RegDB in Table 3, our method also achieved comparable results with previous methods.

Comparison on HITSZ-VCM. There are two testing strategies named video-based and image-based. Following [19], we conducted an average pooling layer on frame-level representations to form a tracklet-level representation during testing. The results are shown in Table 2. "Baseline" refers to only optimizing with one classification and triplet loss. Our method outperformed all existing methods,

Table 1. Comparison with state-of-the-art methods in SYSU-MM01 under single-shot setting. The performance is shown by Rank-k accuracy (%) and mAP (%). The best results and the second are **bold** and <u>underlined</u> marked, respectively. "C" refers to concatenated the output representations of two checkpoints of our model. "2x" refers to 2 times of size of parameters contained in ResNet50.

Params	Method	Venue	All Search				Indoor Search			
			Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
≈1x	CM-NAS [6]	CVPR'21	61.99	92.87	97.25	60.02	67.01	97.02	99.32	72.95
≈1x	CAJL [32]	ICCV'21	69.88	95.71	98.46	66.89	76.30	97.90	99.50	80.40
≈1x	MPANet [31]	CVPR'21	70.58	96.21	98.8	68.24	76.64	98.21	99.57	80.95
≈1x	MMN [36]	ACMMM'21	70.60	96.20	99.00	66.90	76.20	97.20	99.30	79.60
≈1.75x	MTL [12]	PR'22	67.25	95.38	98.46	64.29	69.58	96.66	99.03	74.37
≈1.25x	PAENet [39]	ACMMM'22	74.22	**99.03**	**99.97**	**73.90**	78.04	**99.58**	**100.00**	**83.54**
≈2x	MSCLNet [37]	ECCV'22	<u>76.99</u>	<u>97.93</u>	<u>99.18</u>	71.64	<u>78.49</u>	<u>99.32</u>	<u>99.91</u>	81.17
≈1x	Ours	-	75.18	96.87	99.13	70.12	78.40	97.46	98.91	81.20
≈2x	Ours (C)	-	**77.12**	97.03	99.08	<u>72.33</u>	**82.07**	97.42	98.87	<u>82.95</u>

Table 2. Comparisons of our method with state-of-the-art methods on HITSZ-VCM dataset. Rank-k accuracy (%) and mAP (%) are reported. The best results and the second are in **bold** and <u>underlined</u>, respectively.

Strategy	Method	Venue	Infrared to Visible				Visible to Infrared			
			Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Video	MITML [19]	CVPR'22	<u>63.74</u>	<u>76.88</u>	<u>81.72</u>	45.31	64.54	<u>78.96</u>	82.98	47.69
Image	LbA [23]	ICCV'21	46.38	65.29	72.23	30.69	49.30	69.27	75.90	32.38
	MPANet [31]	CVPR'21	46.51	63.07	70.51	35.26	50.32	67.31	73.56	37.80
	DDAG [33]	ECCV'20	54.62	69.79	76.05	39.26	59.03	74.64	79.53	41.50
	VSD [26]	CVPR'21	54.53	70.01	76.28	41.18	57.52	73.66	79.38	43.45
	CAJL [32]	ICCV'21	56.59	73.49	79.52	41.49	60.13	74.62	79.86	42.81
	Baseline	-	62.02	75.35	81.35	<u>47.05</u>	64.90	78.64	<u>83.68</u>	<u>48.21</u>
	Ours	-	**67.65**	**80.32**	**84.73**	**52.30**	**70.23**	**82.19**	**86.11**	**52.54**

Table 3. Comparison with state-of-the-art methods in RegDB. The performance is shown by Rank-1 (%) and mAP (%). The best results and the second are **bold** and <u>underlined</u> marked, respectively.

Method	Infrared to Visible		Visible to Infrared	
	Rank-1	mAP	Rank-1	mAP
CM-NAS [6]	82.57	78.31	84.54	80.32
CAJL [32]	84.75	77.82	85.03	79.14
MPANet [31]	82.8	80.7	83.7	80.9
MMN [36]	87.5	80.5	91.6	84.1
MTL [12]	88.34	84.06	89.91	85.64
PAENet [39]	**95.35**	**89.98**	**97.57**	**91.41**
MSCLNet [37]	83.86	78.31	84.17	80.09
Ours	<u>91.07</u>	<u>85.23</u>	<u>92.18</u>	<u>86.59</u>

and compared to the baseline, our method achieved 5% improvements on most metrics.

4.4. Ablation Study and Analysis

In this section, we conducted ablation study to evaluate the contribution of each component in our proposed Shape-Guided dIverse fEature Learning (SGIEL). All experiments were performed on SYSU-MM01 under all search, single-shot setting using the same baseline. We tuned the hyper-parameters of each experiment carefully. The results are summarized in Table 4 and 5 and described as follows.

Effectiveness of Erasing Body Shape. We first introduce the baseline (Exp 1) of SGIEL, which using the same backbone but only optimized with one classification and triplet loss on representation $z^{(i)}$. In Exp 2, we add $\mathcal{L}_{gkl}$ to form the objective Eq. (18).

To evaluate the effectiveness of erasing body shape information, in Exp 3, we only applied *shape-related* objective, $\mathcal{L}_{sr}$ in Eq. (12), which indicated the model was forced to only concentrate on *shape-related* feature without considering others, resulting in degeneration of performance compared to Exp 2. However, when applying shape erased feature learning paradigm with $\mathcal{L}_{se}$ in Eq. (16) (Exp 4), we achieved 3% in Rank-1 and 4% in mAP improvement compared to Exp 3. Moreover, in Exp 5, adding an additional orthogonal constraint ($\mathcal{L}_{ortho}$) to strengthen the orthogonality of *shape-related* and *shape-erased* features also boosted the performance. The above ablation study demonstrated the effectiveness of erasing body shape in our SGIEL.

Effectiveness of Orthogonal Constraint. In Section 3.2, we demonstrated the necessity of independent constraint on *shape-related* and *shape-erased* features, and introduced orthogonal decomposition on representation as a relaxation of independent constraint. We evaluate this necessity and our design on orthogonal constraint experimentally in Table 5.

Table 4. The individual improvements of our method performed on SYSU-MM01 with all search, single shot setting. Compared to Exp 2, the baseline (Exp 1) only contains one classification and triplet loss. "SEFEL" denotes our Shape-Erased FEature Learning paradigm. "α" denotes our re-weighting mechanism.

| Exp | $\mathcal{L}_{kl}$ | Component | | | α | Rank-1 | mAP |
| | | SEFEL | | | | | |
		$\mathcal{L}_{ortho}$	$\mathcal{L}_{sr}$	$\mathcal{L}_{se}$			
1(base)						68.55	65.32
2	✓					70.73	66.44
3	✓		✓			69.60	64.06
4	✓		✓	✓		72.00	68.57
5	✓	✓	✓	✓		73.52	69.05
6(full)	✓	✓	✓	✓	✓	75.18	70.12

Table 5. Variants of orthogonal constraint. "Num of Proj" denotes the number of linear projectors.

Num of Proj	$\mathcal{L}_{ortho}$	Rank-1	mAP
2		70.18	66.36
2	✓	72.76	68.53
1		73.49	68.97
1	✓	75.18	70.12

- Necessity of Orthogonal Constraint. To evaluate the necessity of orthogonal constraint, we designed a fully unconstrained model as follows. We only replaced orthogonal projector P in our method with two unconstrained projectors $P_1 \in \mathbb{R}^{n \times m_1}, P_2 \in \mathbb{R}^{n \times m_2}$ independently learned by $\mathcal{L}_{sr}$ and $\mathcal{L}_{se}$, respectively, (*i.e.*, $z_{sr}^{(i)} = P_1^T z^{(i)}$, $z_{se}^{(i)} = P_2^T z^{(i)}$). We reported this experiment as the 1-st line in Table 5, which paid a huge loss of performance.

To further evaluate this necessity, similar to Exp 4 in Table 4, we directly discarded $\mathcal{L}_{ortho}$ in our method (but kept the decomposition process), and also resulted in a performance degeneration. We reported this experiment as the 3-rd line in Table 5. We observed across datasets that cosine similarity between individual dimensions in P remains about 0.015 in average after initialization, and increased to 0.03 in average when convergence, which would be harmful to discover body-shape-unrelated features.

- Design of Orthogonal Constraint. To evaluate the design of orthogonal constraint in shape erased feature learning, based on the above mentioned fully unconstrained model, we added an orthogonal constraint loss similar to $\mathcal{L}_{ortho}$ but between the two projectors, P_1 and P_2. We carefully tuned the hyper-parameters including $m_1 = 512$ and $m_2 = 1024$, and reported the final results as the 2-nd line in Table 5, achieving a lower performance compared to ours.

Visualization on Feature Maps. We visualized *shape-related* objective $\mathcal{L}_{sr}$ and *shape-erased* objective $\mathcal{L}_{se}$ sepa-
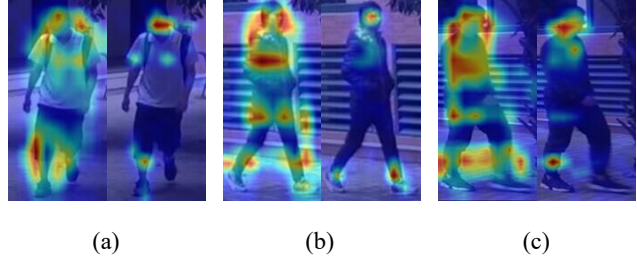

(a) (b) (c)

Figure 4. Visualization of *shape-related* objective $\mathcal{L}_{sr}$ and *shape-erased* objective $\mathcal{L}_{se}$ separately on the last feature map of the backbone through Grad-CAM++. For each pair of images, the left/right images visualized the gradient CAM produced by $\mathcal{L}_{sr}$/$\mathcal{L}_{se}$.

rately on the last feature map of the backbone through Grad-CAM++ [3]. As illustrated in Fig. 4, On the left of each pair of gradient CAM images produced by $\mathcal{L}_{sr}$, the activation mostly focused on the contour of their figures, while the gradient CAM of the right one produced by $\mathcal{L}_{se}$ is more centralized in complementary parts, such as head and legs.

5. Conclusion

In this work, we investigate how to learn diverse features in VI-ReID and try to erase body-shape-related semantic concepts in the learned features to force the ReID model to extract more and other modality-shared discriminative features. We propose shape-erased feature learning paradigm that decorrelates features in two orthogonal subspaces of shape-related and shape-erased ones. By learning discriminative body shape feature in one subspace, the shape-erased feature is forced to discover other modality-shared discriminative semantic concepts in the the other subspace as shape-related feature is constrained in its orthogonal complement. We demonstrate that jointly learning the two objectives achieves a conditional mutual information maximization between shape-erased feature and identity discarding body shape information, thus enhancing the diversity of the learned representation explicitly. Extensive experiments on SYSU-MM01, RegDB, and HITSZ-VCM datasets demonstrate the effectiveness of our method.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. 4

[2] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897–2905, 2018. 4

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 8

[4] 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. 1

[5] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2020. 1, 2

[6] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11823–11832, October 2021. 2, 6, 7

[7] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020. 6

[8] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[9] 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)*, pages 770–778, 2016. 6

[10] Alexander Hermans, Lucas Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *ArXiv*, abs/1703.07737, 2017. 5

[11] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10508–10517, 2021. 3

[12] Nianchang Huang, Kunlong Liu, Yang Liu, Qiang Zhang, and Jungong Han. Cross-modality person re-identification via multi-task learning. *Pattern Recognition*, 128:108653, 2022. 3, 6, 7

[13] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018. 3

[14] Kajal Kansal, A. V. Subramanyam, Zheng Wang, and Shin'Ichi Satoh. Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3422–3432, 2020. 1, 2

[15] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4610–4617, Apr. 2020. 1, 2

[16] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 3

[17] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 6

[18] Baijiong Lin, Feiyang YE, Yu Zhang, and Ivor Tsang. Reasonable effectiveness of random weighting: A litmus test for multi-task learning. *Transactions on Machine Learning Research*, 2022. 6

[19] Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu, and David Zhang. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20973–20982, 2022. 6, 7

[20] Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *International Conference on Learning Representations*, 2021. 6

[21] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 6

[22] 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), 2017. 6

[23] 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 (ICCV)*, pages 12046–12055, October 2021. 7

[24] Nan Pu, Wei Chen, Yu Liu, Erwin M. Bakker, and Michael S. Lew. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2149–2158, New York, NY, USA, 2020. Association for Computing Machinery. 2

[25] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[26] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1531, June 2021. 5, 7

[27] 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. 1

[28] 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 (ICCV)*, pages 225–234, October 2021. 2

[29] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. Rgb-ir person re-identification by cross-modality similarity preservation. *International journal of computer vision*, 128(6):1765–1785, 2020. 2

[30] 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. 1, 6

[31] 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 (CVPR)*, pages 4330–4339, June 2021. 1, 2, 6, 7

[32] 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. 2, 6, 7

[33] 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 (ECCV)*, 2020. 7

[34] 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. *arXiv preprint arXiv:2001.04193*, 2020. 6

[35] 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 (CVPR)*, pages 7349–7358, June 2022. 2

[36] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 788–796, New York, NY, USA, 2021. Association for Computing Machinery. 6, 7

[37] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 462–479. Springer, 2022. 6, 7

[38] Zhiwei Zhao, Bin Liu, Qi Chu, Yan Lu, and Nenghai Yu. Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. *Proceedings of the AAAI Confer-ence on Artificial Intelligence*, 35(4):3520–3528, May 2021. 2

[39] Aihua Zheng, Peng Pan, Hongchao Li, Chenglong Li, Bin Luo, Chang Tan, and Ruoran Jia. Progressive attribute embedding for accurate cross-modality person re-id. MM '22, page 4309–4317, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2, 6, 7

[40] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020. 3