Learning with Twin Noisy Labels for Visible-Infrared Person Re-Identification

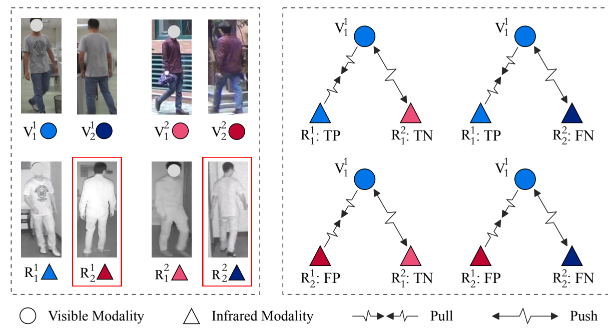
Mouxing Yang[1], Zhenyu Huang[1], Peng Hu[1], Taihao Li[2], Jiancheng Lv[1], Xi Peng[1*]
[1] College of Computer Science, Sichuan University.
[2] Zhejiang Lab, China.
{yangmouxing, zyhuang.gm, penghu.ml, pengx.gm}@gmail.com;
lith@zhejianglab.com; lvjiancheng@scu.edu.cn

Abstract

In this paper, we study an untouched problem in visible-infrared person re-identification (VI-ReID), namely, Twin Noise Labels (TNL) which refers to as noisy annotation and correspondence. In brief, on the one hand, it is inevitable to annotate some persons with the wrong identity due to the complexity in data collection and annotation, e.g., the poor recognizability in the infrared modality. On the other hand, the wrongly annotated data in a single modality will eventually contaminate the cross-modal correspondence, thus leading to noisy correspondence. To solve the TNL problem, we propose a novel method for robust VI-ReID, termed DuAlly Robust Training (DART). In brief, DART first computes the clean confidence of annotations by resorting to the memorization effect of deep neural networks. Then, the proposed method rectifies the noisy correspondence with the estimated confidence and further divides the data into four groups for further utilizations. Finally, DART employs a novel dually robust loss consisting of a soft identification loss and an adaptive quadruplet loss to achieve robustness on the noisy annotation and noisy correspondence. Extensive experiments on SYSU-MM01 and RegDB datasets verify the effectiveness of our method against the twin noisy labels compared with five state-of-the-art methods. The code could be accessed from `https://github.com/XLearning-SCU/2022-CVPR-DART`.

1. Introduction

Person re-identification (ReID) aims to match a specified person from the gallery set. However, most existing person Re-ID methods [3, 4, 28, 32–34] only focus on searching RGB images captured by visible cameras, which might fail to achieve encouraging results under poor illumination environments (*e.g.*, at night). To solve this problem, some visible-infrared person re-identification (VI-ReID) methods [17, 18, 23, 26, 29] have been proposed to find the corresponding identities across two modalities. More specifically, these methods usually leverage the identity annotations to establish the cross-modal correspondence so that the identity-aware discrimination is enlarged and the cross-modal discrepancy is eliminated.



(a) Noisy Annotations (b) Noisy Correspondence

Figure 1. The twin noisy labels in VI-ReID. In the figure, V_i^j/R_i^j denotes sample i with the annotated identity j from the visual/infrared modality, and the color indicates the latent correct identity. (a) Noisy annotations: due to the poor recognizability in the infrared modality, samples 2 of identity 1 and 2 will be mixed up thus being wrongly annotated with identity 2 and 1, respectively, *i.e.*, R_2^1 and R_2^2 are noisy annotations. (b) Noisy correspondence: as the cross-modal pairs are constructed resorting to the annotations, both the positive and negative pairs might be false due to noisy annotations, leading to the mismatching phenomenons. With such noisy correspondences, the false-positive and -negative would be wrongly pulled and pushed during training, respectively.

Although VI-ReID has achieved promising performance, its success heavily relies on high-quality annotated data. In practice, however, it is daunting and even impossible to precisely annotate all samples due to the poor recognizability, especially the color information is lost in the infrared modality as shown in Fig. 1(a). As a result, it is inevitable to result in noisy annotations (NA) problem, thus degrading the performance of ReID models. Although some studies [30, 31] have devoted to mitigating the perfor-

*Corresponding author

mance degradation caused by the noisy annotations, all of them only focus on the noisy annotation problem in visible modality ReID, while ignoring multi-modality cases such as VI-ReID. Furthermore, once multi-modality Re-ID is considered, another special noisy label will be encountered, *i.e.*, noisy correspondence (NC). More formally, we define the noisy correspondence as the mismatched cross-modal pairs whose correspondence is established by using the noisy annotations from their respective modalities. As illustrated in Fig. 1(b), such a cross-modal pair construction approach will inevitably lead to noisy correspondence for VI-ReID, *i.e.*, False Positive (FP) pairs, and False Negative (FN) pairs.

Based on the above observation, in this paper, we reveal a new problem for VI-ReID, termed Twin Noisy Labels (TNL). Different from the traditional noisy label studies [6, 6, 10, 12, 16] which only consider the NA challenge, TNL simultaneously consider NA in category and NC in cross-modal pairs. It should be pointed out that, it is intractable to adopt existing NA-oriented methods to rectify the noisy annotations in VI-ReID so that the TNL problem is solved due to the following reasons. First, the success of most existing noisy label methods is mainly limited to the case of small category numbers, whereas the category (person) number in ReID is in hundreds at least. Second, despite the issue of category number, it is impossible to fully rectify all noisy annotations and accordingly avoid the noisy correspondence problem. In other words, TNL is unavoidable in practice. Third, it is nontrivial and hard to achieve a promising result by employing the existing noisy annotation methods such as [10] to correct wrong annotations for VI-ReID due to the difficulty in sampling and joint optimization. For a comprehensive study, we present experiments to verify the above claim in the supplementary.

To solve the TNL problem in VI-ReID, we propose a novel method for learning with both noisy annotations and correspondence. The proposed DuAlly Robust Training (DART) consists of co-modeling and pair-division modules with a novel objective function. In detail, the co-modeling module first computes the clean confidence for each sample resorting to the memorization effect of deep neural networks. Then the pair-division module rectifies the noisy correspondence with the confidence and further divides the noisy pairs into four subsets, *i.e.*, true positive pairs (TP), true negative pairs (TN), false positive pairs (FP), and false negative pairs (FN). Finally, to achieve robust VI-ReID, we propose a novel dually robust objective function which consists of a soft identification loss and adaptive quadruplet loss. In short, the soft identification loss is employed to penalize samples of noisy annotation while learning the identity-aware representation. The adaptive quadruplet loss leverages the above four kinds of pairs to alleviate the modality discrepancy.

The contributions and novelties of this work could be summarized as follows:

- We reveal a new problem for VI-ReID, termed twin noisy labels, which could be a new paradigm for noisy labels. Different from the existing noisy label studies which only consider the NA problem, TNL refers to both the NA in the category and accompanying NC between cross-modal pairs. Notably, as far as we know, there is no study on VI-ReID with noisy annotation, not to mention the more practical and challenging TNL problem.
- To achieve robust VI-ReID, we propose a novel method for learning with TNL, termed dually robust training. To the best of our knowledge, the proposed method could be the first successful solution towards TNL.
- Extensive experiments on SYSU-MM01 and RegDB datasets verify the effectiveness of our method against twin noisy labels compared with five state-of-the-art methods.

2. Related Works

In this section, we will briefly introduce two related topics with this study, namely, VI-ReID and learning with noisy labels.

2.1. Visible-infrared Person Re-identification

To alleviate the cross-modality discrepancy, a number of VI-ReID approaches [2, 11, 14, 17, 22, 23] have been proposed during past years. According to the choice in alleviating the discrepancy, the existing methods could be divided into the following three groups: i) the network-design based methods [2, 11, 22, 23, 27] which aim at learning the discriminative representation shared across modalities; ii) metric-design based methods [28, 29] which aim at designing different metrics or losses to alleviate the modality discrepancy; iii) modality-transform based methods [7, 18, 20, 21, 26] which aim at finding transformation or augmentation strategies to bridge the gap of modalities.

Almost all existing VI-ReID methods assume that the annotations are faultless. However, the noisy annotations are inevitably introduced in the data collection, which would simultaneously result in noisy correspondence as elaborated above. Noticed again, as far as we know, there is no effort has been devoted to VI-ReID with noisy annotation so far, not mention to the twin noisy labels revealed in this paper.

2.2. Learning with Noisy Labels

Learning with noisy labels is a long-standing problem in the machine learning community. Most of the existing methods [5, 6, 10, 12, 16, 19] aim at combating the noisy annotation in the classification task by designing a robust loss, noise filter, or robust architecture. Recently, [30, 31] focuses on handling the noisy annotation in the visible person re-identification (ReID) task by proposing the instance-reweighing strategy and the feature uncertainty loss, respectively. In terms of cross-modal retrieve, [8] proposes to combat with noisy annotation by resorting to the negative learning strategy. Besides the noisy annotation, in very recent, [24, 25] finds that the correspondence of negative pairs in contrastive learning might be false, *i.e.*, false negatives, and designs a noise-robust contrastive loss to handle with that. [9] formally releases that the correspondence of cross-modal pair may be false and proposes to handle the false positives to achieve robust cross-modal matching.

Among the aforementioned studies, [8, 9, 24, 25, 30, 31] might be the most relevant, while they are remarkably different in the following aspects. First, [30, 31] only considers the noisy annotation problem in single-modality ReID, whereas our study reveals the noisy correspondence problem accompanied with the noisy annotation in VI-ReID. Second, [8] directly uses the off-the-shelf cross-modal pairs at the instance-level which is unavailable in the VI-ReID task. Besides, [24, 25] and [9] show the existence of FN and FP respectively and the proposed methods could be only either robust against FN or FP. In contrast, DART takes all the possible noisy correspondence cases into consideration and the proposed loss is robust to TP, TN, FP, and FN.

3. Method

In this section, we elaborate on the proposed DART which could be one of the first attempts to solve the twin noisy label problem in VI-ReID. In brief, Section 3.1 will present a formal definition of the TNL problem in VI-ReID. Then, Section 3.2 introduces the co-modeling module which aims to compute the correctly annotated confidence for each sample by resorting to the memory effect of deep neural networks (DNNs). Based on the confidences, Section 3.3 elaborates on how to divide pairs into different groups and rectify their correspondences. Finally, based on the confidences and pair partitions, Section 3.4 details the proposed robust objective function which consists of the soft identification and adaptive quadruplet losses.

3.1. Problem Formulation

For clarity, we use $\mathcal{V} = \{\boldsymbol{x}_i^v, \boldsymbol{y}_i^v\}_{i=1}^{N_v}$ and $\mathcal{R} = \{\boldsymbol{x}_i^r, \boldsymbol{y}_i^r\}_{i=1}^{N_r}$ to denote the visible images $\boldsymbol{x}_i^v$ and infrared images $\boldsymbol{x}_i^r$ with the corresponding annotation $\boldsymbol{y}_i^t$, where N_t is the number of images and $t \in \{v, r\}$. Given a visible/infrared query, VI-ReID aims to match the images of the same identity from the infrared/visible gallery set. To this end, most existing methods first construct cross-modal positive and negative pairs, *i.e.*, $(\boldsymbol{x}_i^v, \boldsymbol{x}_j^r)$, where the correspondence $y_{ij}^p = 1$ *i.f.f.* $\boldsymbol{y}_i^v = \boldsymbol{y}_j^r$, otherwise $y_{ij}^p = 0$. After that, the triplet-based loss and identification loss are employed to alleviate the modality discrepancy while guaranteeing the identity-aware discrimination. However, the noisy annotations would probably result in some noisy correspondences. More specifically, the correspondence between positive pairs ($y_{ij}^p = 1$) or negative pairs ($y_{ij}^p = 0$) may be wrongly established as $y_{ij}^p = 0$ or $y_{ij}^p = 1$ because it is intractable to known the sample i with clean annotation $\hat{\boldsymbol{y}}_i^t$. To solve such a twin noisy label problem, we propose DART which consists of co-modeling and pair division modules, and a joint robust objective function as shown in Fig. 2.

3.2. Co-modeling

To begin, DART will project the visible and infrared modalities into a shared latent space to compute features and predict identity of $\{\boldsymbol{x}_i^v\}_{i=1}^{N_v}$ and $\{\boldsymbol{x}_i^r\}_{i=1}^{N_r}$ via two modal-specific networks of $\{F^v, C^v\}$ and $\{F^r, C^r\}$, where F^v and F^r are two feature extractors with some shared layers [26, 28], and C^v and C^r are two classifiers. With features $F^t(\boldsymbol{x}_i^t)$ and annotations $\boldsymbol{y}_i^t$, the pairwise correspondence matrix $\mathbf{Y}^p$ and distance matrix $\mathbf{D}$ are obtained respectively, where y_{ij}^p is the noisy correspondence and d_{ij} denotes the distance between $(\boldsymbol{x}_i^v, \boldsymbol{x}_j^r)$ in the latent space. Namely,

$$d_{ij} = \|F^v(\boldsymbol{x}_i^v) - F^r(\boldsymbol{x}_j^r)\|_2. \tag{1}$$

As discussed in 3.1, both the annotation $\boldsymbol{y}_i^t$ and correspondence y_{ij}^p might be noisy. To handle the noisy annotation problem, we adopt the empirical finds [**?**] in the memory effect of DNNs. More specifically, DNNs are apt to fit the simple patterns, thus leading to a relatively small loss for the clean (*i.e.*, simple) samples in the initial training phase. Based on this observation, one could compute the probability of samples being correctly annotated by modeling the loss distribution. Specifically, given the modal-specific network $\{F^t, C^t\}$ with parameter θ_t ($t \in \{v, r\}$), we compute the per-sample identification (cross-entropy) loss via

$$\ell^{id}(\theta_t) = \left\{\ell_i^{id}\right\}_{i=1}^{N_t} = \left\{\mathcal{L}^{id}(\boldsymbol{x}_i^t, \boldsymbol{y}_i^t)\right\}_{i=1}^{N_t}, \tag{2}$$

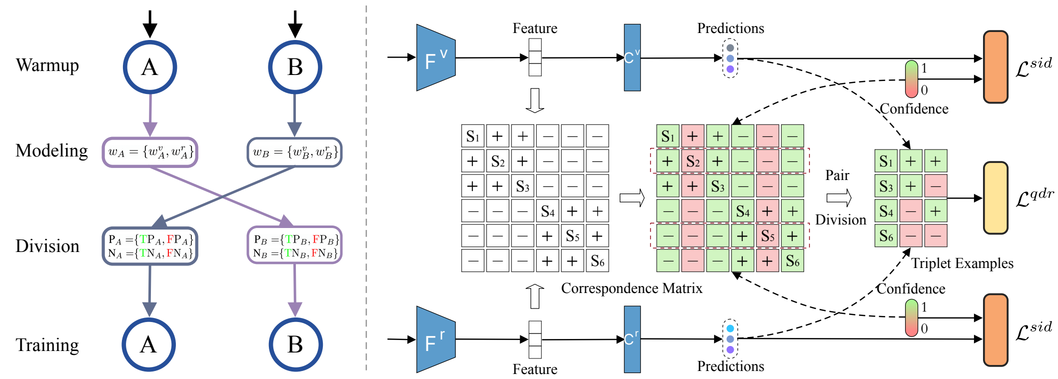
where $\mathcal{L}^{id}$ is the vanilla identification loss defined by

$$\mathcal{L}^{id}(\boldsymbol{x}_i^t, \boldsymbol{y}_i^t) = -\log P\left(\boldsymbol{y}_i^t \mid C^t\left(F^t\left(\boldsymbol{x}_i^t\right)\right)\right). \tag{3}$$

Following [10], we fit the per-sample loss distribution of all training data by modeling a two-component Gaussian Mixture Model as below:

$$p(\ell^{id} \mid \theta_t) = \sum_{k=1}^{K} \gamma_k \phi(\ell^{id} \mid k), \tag{4}$$

where γ_k and $\phi(\ell^{id} \mid k)$ are the mixture coefficient and probability density of the k-th component, respectively. According to the memory effect of DNNs, we could compute



(a) Training pipeline

(b) Dually robust training framework

Figure 2. Overview of the proposed method. (a): Training pipeline of DART. In brief, DART consists of two individual networks (A, B) which work in a co-teaching manner. More specifically, DART first warmups both A and B by using Eq. 2 for initialization. After that, at each epoch, the following procedure is performed. First, the network A/B models the per-sample identification loss distribution to estimate the correctly annotated confidence w for each sample and then feed w into B/A for further training. The next step will divide the data pairs into four subsets, *i.e.*, TP, FP, TN, and FN, and rectify their correspondence. Finally, the estimated confidence and rectified pairs are used to train the networks. (b): Dually robust training framework for A and B. In the figure, "S", "+", and "-" denote the anchor, positive, and negative samples, respectively. The sample whose confidence is above a specific threshold would be in green, otherwise in red. As shown, the backbone will first extract the features for visible and infrared modalities, respectively. Then, the features are fed to classifiers to get predictions and used to construct the correspondence matrix. After that, the correspondence matrix is established with the estimated confidence and the anchor in red would be discarded due to over-low confidence. With the help of the pair division module, the pairs would be categorized into four groups which are then combined as triplets (see (b) for some combination examples) for optimization. Finally, the predictions, triplets, and confidences are used to achieve dually robust training by minimizing our losses.

the correctly annotated confidence w_i for each sample i via poster probability over the small mean value component κ, *i.e.*,

$$w_i = p(\kappa \mid \ell_i^{id}). \tag{5}$$

However, as pointed in [6], it may introduce error accumulation if the neural network is simply trained with the self-modeling confidence. To avoid the bias, we adopt a co-modeling approach. To be specific, we individually train two sets of network which are with the same architecture while different initializations, *i.e.*, $A = \{F_A^v, C_A^v, F_A^r, C_A^r\}$ and $B = \{F_B^v, C_B^v, F_B^r, C_B^r\}$. At each epoch, the networks A or B will model a GMM to fit the loss distribution for computing the confidences, respectively. Then, the confidences are fed into the other network for further training. Notably, following [1, 6, 10, 25], a warm-up strategy is adopted for each network by using the vanilla cross-entropy loss (Eq. 3) for initialization.

3.3. Pair Division

Thanks to the co-modeling module, the clean confidence of the annotation could be estimated, which will be used to partition the data pairs into clean and noise portions. After that, we will further divide these portions into four subsets, *i.e.*, true positive (TP), true negative (TN), false positive (FP), and false negative (FN) pairs.

Following [28], we adopts the modality mix strategy to construct pairs from both within and across modalities. For given pairs $(\boldsymbol{x}_i^{t_1}, \boldsymbol{x}_j^{t_2})$ with the noisy correspondence y_{ij}^p where $t_k \in \{v, r\}$, $k \in \{1, 2\}$, we set a threshold η on the estimated annotation confidence to divide them into clean portion $\mathcal{S}^c = \{(\boldsymbol{x}_i^{t_1}, \boldsymbol{x}_j^{t_2}), y_{ij}^p \mid w_i > \eta, w_j > \eta\}$ and noisy portion $\mathcal{S}^n = \{(\boldsymbol{x}_i^{t_1}, \boldsymbol{x}_j^{t_2}), y_{ij}^p \mid w_i > \eta, w_j \leq \eta\}$, where i and j index the anchor and positive ($y_{ij}^p = 1$) or negative ($y_{ij}^p = 0$) sample, respectively. Notably, we discard pairs whose anchors' confidences are less than η as they cannot be correctly divided. After rectifying the correspondence of $\{\mathcal{S}^c, \mathcal{S}^n\}$, the newly obtained correspondence is denoted by $\{\hat{\mathcal{S}}^c, \hat{\mathcal{S}}^n\}$. The rectifying operation is as below:

$$\hat{y}_{ij}^p = \mathbb{I}(y_{ij}^p \in \mathcal{S}^c) \odot y_{ij}^p, \tag{6}$$

where $\mathbb{I}(y_{ij}^p \in \mathcal{S}^c)$ denotes whether the pair belongs to the clean portion or not, and $\odot$ is the xnor operation. Eq 6 means that if the positive pair ($y_{ij}^p = 1$) comes from $\mathcal{S}^c$, then it is TP; otherwise FP. Similarly, the negatives from $\mathcal{S}^c$ and $\mathcal{S}^n$ are treated as TN and FN, respectively.

For FN pairs, we will further refine it to improve the accuracy. To be specific, for a negative sample $\boldsymbol{x}_j^{t_2} \in \mathcal{S}^n$, it should be TN but wrongly treated as FN, if its confidence w_j is not larger than η (*i.e.*, $w_j \leq \eta$) and meanwhile its identity is different from the anchor sample (*i.e.*, $\boldsymbol{y}_i^{t_1} \neq \boldsymbol{y}_j^{t_2}$). To

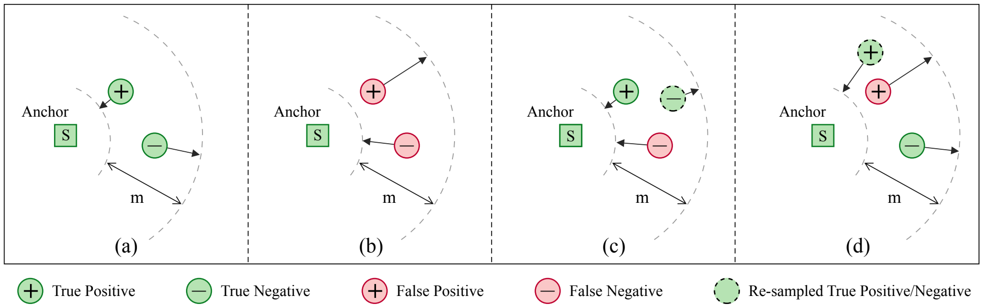


Figure 3. Four kinds of triplets due to twin noisy labels. For each triplet, the proposed quadruplet loss (Eq. 10) would adaptively turn into different variants to achieve robust learning.

recall such TN pairs, we revised their correspondence by

$$\hat{y}^p_{ij} = \mathbb{I}(C^t(F^t(\boldsymbol{x}^{t_1}_i)) = C^t(F^t(\boldsymbol{x}^{t_2}_j))), \forall(\boldsymbol{x}^{t_1}_i, \boldsymbol{x}^{t_2}_j) \in \mathcal{S}^n \tag{7}$$

where $C^t(F^t(\boldsymbol{x}^t_i))$ is the annotation prediction. With Eq. 6–7, all training pairs will be divided into one of the TP, FP, TN and FN subsets.

3.4. Dually Robust Objective Function

Thanks to the co-modeling and divide modules, DART could obtain the annotation confidence and rectify the correspondence of pairs. Then, we employ the following loss to achieve robust VI-ReID:

$$\mathcal{L} = \mathcal{L}^{sid} + \mathcal{L}^{qdr}, \tag{8}$$

where $\mathcal{L}^{sid}$ and $\mathcal{L}^{qdl}$ are soft identification loss and quadruplet loss which are proposed to combat noisy annotations and noisy correspondence, respectively. In the following, we elaborate on each loss one by one.

Robust on Noisy Annotations: Instead of simply discarding the samples with noisy annotations [6], for either the network A or B, we utilize the confidence w_i to penalize the noise during optimization. To this end, the following soft identification loss is proposed:

$$\mathcal{L}^{sid} = -w_i \log P\left(\boldsymbol{y}^t_i \mid C^t\left(F^t\left(\boldsymbol{x}^t_i\right)\right)\right), \tag{9}$$

where $t \in \{v, r\}$ denotes the visible or infrared modality.

Robust on Noisy Correspondences: With the four pair subsets (TP, FP, TN, and FN), DART needs to alleviate the modality discrepancy with their help. As the vanilla triplet losses can only handle the combination of TP and TN, it is necessary to develop a novel method that could handle all possible combinations (in a form of triplets) of the four subsets.

To this end, $\mathcal{L}^{qdr}$ is designed, which could be adaptive to different combinations. Formally, given a triplet $(\boldsymbol{x}^{t_1}_i, \boldsymbol{x}^{t_2}_j, \boldsymbol{x}^{t_3}_k)$, where $t_k \in \{v, r\}$, $k \in \{1, 2, 3\}$, its pairwise distances are denoted as d_{ij} and d_{ik}, where i, j, k denote the index of anchor, positive ($y^p_{ij} = 1$) and negative ($y^p_{ik} = 0$) samples, respectively. Thanks to our pair division module, these triplets could be grouped into one combination of TP-TN ($\hat{y}^p_{ij} = 1$, $\hat{y}^p_{ik} = 0$), FP-FN ($\hat{y}^p_{ij} = 0$, $\hat{y}^p_{ik} = 1$), TP-FN ($\hat{y}^p_{ij} = 1$, $\hat{y}^p_{ik} = 1$), FP-TN ($\hat{y}^p_{ij} = 0$, $\hat{y}^p_{ik} = 0$) with their rectified correspondence. Notably, the last three combinations are with noisy correspondence focused in this paper.

To combat such a noisy correspondence problem, we propose the following adaptive quadruplet loss:

$$\mathcal{L}^{qdr} = \mathcal{L}^{tri} + \mathcal{L}^{qdt}, \tag{10}$$

where $\mathcal{L}^{tri}$ is defined as:

$$\mathcal{L}^{tri} = m + \frac{(-1)^{(\hat{y}^p_{ij} \otimes \hat{y}^p_{ik})(1-\hat{y}^p_{ij})} d_{ij} + (-1)^{(\hat{y}^p_{ij} \otimes \hat{y}^p_{ik})(1-\hat{y}^p_{ik})} d_{ik}}{(-1)^{(1-\hat{y}^p_{ij})(1-\hat{y}^p_{ik})} 2^{\hat{y}^p_{ij} \odot \hat{y}^p_{ik}}}, \tag{11}$$

where m is a margin fixed as a constant in our experiment, $\otimes$ and $\odot$ denote the xor and xnor operations, respectively. As both TP-FN and FP-TN consist of homogeneous pairs, the existing triplet losses cannot handle such a case. Hence, we propose additionally sampling a pair $(\boldsymbol{x}^{t_1}_i, \boldsymbol{x}^{t_4}_s)$ of confidence $w_s > \eta$ for use of the following quadruplet term:

$$\mathcal{L}^{qdt} = (-1)^{\hat{y}^p_{ij} \hat{y}^p_{ik}} (\hat{y}^p_{ij} \odot \hat{y}^p_{ij}) d_{is}. \tag{12}$$

Eq. 12 will take effect when the pairs $(\boldsymbol{x}^{t_1}_i, \boldsymbol{x}^{t_2}_j)$ and $(\boldsymbol{x}^{t_1}_i, \boldsymbol{x}^{t_3}_k)$ are of the same correspondence, *i.e.,* they are TP-FN or FP-TN triplet. In the following, we will elaborate on the robustness enjoyed by $\mathcal{L}^{qdr}$ in different situations:

- TP-TN (Fig. 3(a)): For the pair divided into TP ($\hat{y}^p_{ij} = 1$) or TN ($\hat{y}^p_{ik} = 0$), the goal is to decrease the pairwise distance of TP while increasing that of TN. In the case, $\mathcal{L}^{qdr}$ will adaptively turn into the vanilla triplet loss as below:

$$\mathcal{L}^{qdr} = [d_{ij} - d_{ik} + m]_+. \tag{13}$$

- FP-FN (Fig. 3(b)): For the pair divided into FP ($\hat{y}^p_{ij} = 0$) or FN ($\hat{y}^p_{ik} = 1$), the goal is to increase the pairwise distance of FP while decreasing that of FN. Then, $\mathcal{L}^{qdr}$ becomes:

$$\mathcal{L}^{qdr} = [-d_{ij} + d_{ik} + m]_+, \tag{14}$$

- TP-FN (Fig. 3(c)): For the pair divided into TP ($\hat{y}_{ij}^p = 1$) or FN ($\hat{y}_{ik}^p = 1$), the goal is to increase the pair-wise distance of both TP and FN pairs. Then, $\mathcal{L}^{qdr}$ becomes:

$$\mathcal{L}^{qdr} = [-d_{is} + \frac{d_{ij} + d_{ik}}{2} + m]_+. \tag{15}$$

- FP-TN (Fig. 3(d)):For the pair divided into FP ($\hat{y}_{ij}^p = 0$) or TN ($\hat{y}_{ik}^p = 0$), the goal is to decrease the pairwise distance of both FP and TN. Then, $\mathcal{L}^{qdr}$ becomes:

$$\mathcal{L}^{qdr} = [d_{is} - \frac{d_{ij} + d_{ik}}{2} + m]_+. \tag{16}$$

4. Experiment

In this section, we carry out experiments on the SYSU-MM01 [22] and RegDB [13] datasets to verify the robustness of DART against twin noisy labels. Due to the space limitation, we present more experiments in the supplementary materials.

4.1. Experiment Settings

DART is a general framework which could endow almost all existing VI-ReID methods with robustness against twin noisy labels. Hence, ADP [26] is used to verify the effectiveness of DART, which is a very recently-proposed VI-ReID method. In detail, we retain the backbone and pipeline of ADP except for the loss function. To endow the robustness on ADP, we adopt the co-modeling and pair division modules with the dually robust objective function instead.

In the experiment, DART is implemented in PyTorch 1.7.0 [15] and all the evaluations are carried out on GeForce RTX 3090 GPUs on Ubuntu 20.04 OS. The margin in Eq. 11, the threshold for confidence estimation, and the warm-up epoch are fixed as 0.3, 0.5, and 1 for all experiments, respectively. In the testing phase, similar to most multi-view learning methods [**?**, **?**], we simple use the mean outputs of model A and B as the final representation for inference.

For evaluation, we use all two publicly available datasets. To be specific,

- SYSU-MM01 [22]: It is a large-scale VI-ReID dataset collected in the SYSU campus using four visible cameras and two near-infrared ones under both indoor and outdoor environments. The training set consists of 22,258 visible images and 11,909 infrared images distributed over 395 identities, while the query and gallery set is composed of 3,803 infrared images and 301 randomly sampled visible images from 96 identities for single-shot evaluation, respectively.
- RegDB [13]: The dataset contains 8,240 images of 412 identities collected by a dual-camera (one visible and one infrared) system. For each of the identities, there are 10 visible and 10 infrared images.

To verify the robustness of DART against noisy labels, we refer to the setting used in [31]. In detail, we randomly select a specific percentage of training images from each modality and randomly assign wrong identities to them.

For fair comparisons, we follow the common testing settings used in the most existing VI-ReID methods. In brief, the SYSU-MM01 dataset consists of two testing modes, namely, the *all-search* and *indoor-search* modes. The RegDB dataset contains two test settings, namely, *visible-to-infrared* and *infrared-to-visible*. Following [17, 23, 26], for the SYSU-MM01 dataset, we evaluate the performance under both testing modes with 10 random gallery sets chosen. For the RegDB dataset, we perform 10 trials with different training/testing splits under both testing settings. In the evaluation, we report the average results on Cumulative Matching Characteristic (CMC), mean Average Precision (mAP), and mINP [26, 28].

4.2. Comparison with State of the Arts

In this section, we compare DART with five state-of-the-art VI-ReID methods, namely, AGW [28], DDAG [27], LbA [14], MAPNet [23], and ADP [26] on the SYSU-MM01 and RegDB dataset. For extensive evaluations, the noise ratio varies from 0%, 20%, to 50%. In addition, we also report the results of ADP on the clean SYSU-MM01 and RegDB datasets by discarding samples with noisy annotation, denoted by ADP-C. Clearly, ADP-C is a quite strong baseline since the used data does not contain any noisy labels.

When the noise ratio is 0%, we refer to the results reported in the corresponding papers. For other noise ratios, we train the baselines with the recommended settings and report the corresponding results in Table 1 and 2. From the results, one could observe that DART is competitive to ADP under the noise-free setting even though DART is specially designed for combating twin noisy labels. When the data is contaminated by the noisy annotations, DART remarkably outperforms all the baselines by a large margin. Besides, even comparing with ADP-C which is trained on the clean data, DART improves mAP by 4.16%, 2.87%, 3.89%, and 3.94% on SYSU-MM01 and 5.94%, 4.02%, 0.43% and 2.11% on RegDB in the four valuations under noise ratio of 20% and 50%.

4.3. Ablation Study

In this section, we perform ablation studies on SYSU-MM01 to verify the importance of each component in DART. As DART endows ADP the robustness with three

Table 1. Comparisons with state-of-the-art methods on the SYSU-MM01 dataset under the noise ratio of $0\%, 20\%$ and 50%, respectively. The best and second best results are highlight in **bold** and underline.

Noise	Methods	All-Search					Indoor-Search				
		Rank-1	Rank-10	Rank-20	mAP	mINP	Rank-1	Rank-10	Rank-20	mAP	mINP
0%	AGW (TPAMI2021)	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
	DDAG (ECCV2020)	54.75	90.39	95.81	53.02	39.62	61.02	94.06	98.41	67.98	62.61
	LbA (ICCV2021)	55.41	–	–	54.14	–	58.46	–	–	66.33	–
	MPANet (CVPR2021)	**70.58**	96.21	98.8	**68.24**	–	**76.74**	**98.21**	**99.57**	**80.95**	–
	ADP (ICCV2021)	69.88	95.71	98.46	66.89	**53.61**	76.26	97.88	99.49	80.37	**76.79**
	DART (Ours)	68.72	**96.39**	**98.96**	66.29	53.26	72.52	97.84	99.46	78.17	74.94
20%	AGW (TPAMI2021)	17.68	56.80	72.45	18.15	8.55	20.83	65.01	82.43	29.80	25.31
	DDAG (ECCV2020)	14.55	46.58	61.81	13.99	5.56	15.13	50.68	69.33	22.37	18.34
	LbA (ICCV2021)	9.86	39.47	55.85	10.23	3.84	10.10	44.06	64.45	17.39	13.97
	MPANet (CVPR2021)	21.59	63.58	78.71	21.21	–	23.80	70.18	86.44	33.17	–
	ADP (ICCV2021)	25.44	67.55	80.88	23.71	11.05	26.61	70.68	85.19	34.97	29.61
	ADP-C (ICCV2021)	63.67	94.13	97.78	61.57	48.02	68.52	96.13	98.73	73.82	69.66
	DART (Ours)	**66.31**	**95.31**	**98.38**	**64.13**	**50.69**	**70.52**	**97.08**	**99.03**	**75.94**	**72.30**
50%	AGW (TPAMI2021)	7.93	37.56	55.78	9.75	4.38	9.61	47.87	70.47	18.14	15.22
	DDAG (ECCV2020)	6.68	28.95	43.77	7.52	2.93	8.39	37.87	57.86	15.12	12.33
	LbA (ICCV2021)	2.67	17.78	30.27	4.15	1.85	4.87	29.39	48.97	10.96	8.63
	MPANet (CVPR2021)	6.98	32.75	49.16	8.20	–	8.47	40.71	61.37	15.85	–
	ADP (ICCV2021)	8.00	42.55	62.14	10.83	5.21	11.49	52.99	76.77	20.81	17.53
	ADP-C (ICCV2021)	59.17	92.52	97.28	56.49	41.80	62.99	94.84	98.08	69.05	64.29
	DART (Ours)	**60.27**	**93.41**	**97.47**	**58.69**	**45.33**	**65.74**	**95.04**	**98.23**	**71.77**	**68.14**

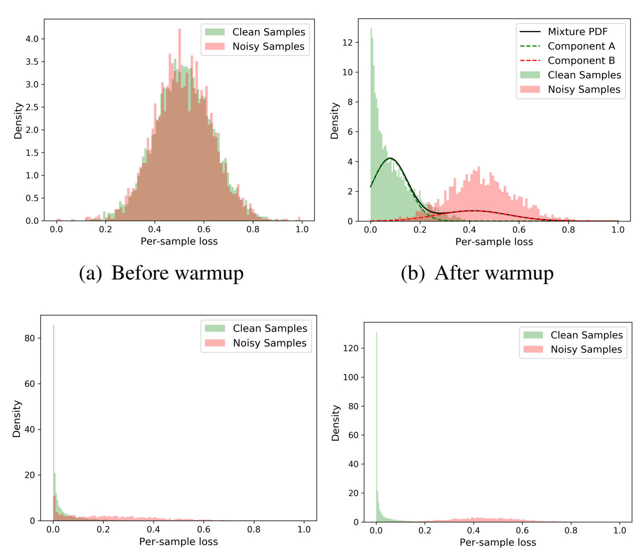
Table 2. Comparisons with state-of-the-art methods on the RegDB dataset under the noise ratio of $0\%, 20\%$ and 50%, respectively. The best and second best results are highlight in **bold** and underline.

Noise	Methods	Visible to Thermal			Thermal to Visible		
		Rank-1	mAP	mINP	Rank-1	mAP	mINP
0%	AGW (TPAMI2021)	70.05	66.37	50.19	70.49	65.9	51.24
	DDAG (ECCV2020)	69.34	63.46	49.24	68.06	61.80	48.62
	LbA (ICCV2021)	74.17	67.64	–	72.43	65.46	–
	MPANet (CVPR2021)	83.70	**80.90**	–	82.80	**80.70**	–
	ADP (ICCV2021)	**85.03**	79.14	**65.33**	**84.75**	77.82	**61.56**
	DART (Ours)	83.60	75.67	60.60	81.97	73.78	56.70
20%	AGW (TPAMI2021)	47.77	31.35	12.43	47.18	30.86	11.85
	LbA (ICCV2021)	35.99	23.48	7.49	36.18	22.75	6.74
	DDAG (ECCV2020)	39.27	25.74	10.03	37.69	25.07	9.61
	MPANet (CVPR2021)	33.83	23.50	–	32.62	22.06	–
	ADP (ICCV2021)	50.71	35.92	14.12	49.98	34.75	12.62
	ADP-C (ICCV2021)	78.39	70.02	51.80	75.81	68.95	51.19
	DART (Ours)	**82.04**	**74.18**	**57.89**	**79.48**	**71.72**	**54.47**
50%	AGW (TPAMI2021)	21.87	13.40	3.93	20.98	12.95	3.70
	DDAG (ECCV2020)	24.03	14.44	4.25	21.46	13.38	4.28
	LbA (ICCV2021)	11.65	6.68	1.53	10.24	6.34	1.46
	MPANet (CVPR2021)	9.51	6.13	–	11.41	6.67	–
	ADP (ICCV2021)	17.04	11.25	3.55	20.28	12.31	3.24
	ADP-C (ICCV2021)	77.43	66.75	47.25	74.89	63.05	41.83
	DART (Ours)	**78.23**	**67.04**	**48.36**	**75.04**	**64.38**	**43.62**

components, we conduct the study on the following variants. More specifically, the proposed co-modeling module and soft identification loss (Eq. 9) are added on ADP to verify the robustness of DART on noisy annotation, which is denoted as "B + $\mathcal{L}^{sid}$". Besides, the pair division module and special triplet loss (Eq. 11) are added to verify the robustness against the noisy correspondence. The third variant is adding the quadruplet term (Eq. 12) to verify the capacity of DART in handling the FP-TN and TP-FN triplets, and such a capacity could further boost the robustness on noisy correspondence. The results are summarized in Table 3 which illustrates that each component plays an inseparable role in combating the twin noisy labels.

Table 3. Ablation studies on SYSU-MM01 with noise ratio of 20% under the all-search mode.

Method	SYSU-MM01 under All-search Evaluation				
	Rank-1	Rank-10	Rank-20	mAP	mINP
B	25.44	67.55	80.88	23.71	11.05
B + $\mathcal{L}^{sid}$	49.24	89.14	95.66	46.78	31.32
B + $\mathcal{L}^{sid}$ + $\mathcal{L}^{tri}$	65.44	95.01	98.13	63.15	50.35
B + $\mathcal{L}^{sid}$ + $\mathcal{L}^{tri}$ + $\mathcal{L}^{qdt}$	**66.31**	**95.31**	**98.38**	**64.13**	**50.69**



(a) Before warmup (b) After warmup

(c) Without Eq. 9 (d) With Eq. 9

Figure 4. Per-sample loss distribution under different situations.

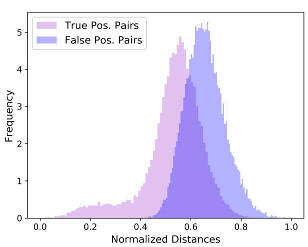
4.4. Visualization on the Robustness

In this section, we qualitatively analyze the robustness of DART against the noisy annotations and noisy correspondence on the SYSU-MM01 dataset with the noise ratio of 20%.

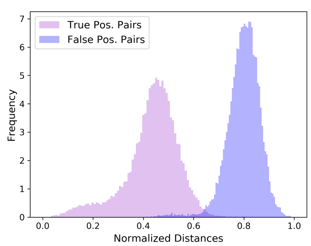
Robustness against Noisy Annotations: As discussed in Section 3.2 and 3.4, DART enjoys robustness against noisy annotations with the help of the soft identification loss (Eq. 9). To visually show the achieved robustness, we illustrate the per-sample identification loss distribution of all the training samples before and after warmup, as well as without and with the help of Eq. 9. The results are shown in Fig. 4, from which one could have the following observations. First, after the warmup stage, the losses of most clean sample are smaller than that of noisy sample, which verifies that the neural networks will fit the clean samples first. As there is still a non-negligible mixture of clean and noisy samples, DNNs would continually fit the noise as the optimization goes without Eq. 9. In other words, our loss will prevent the noisy annotations from dominating the network optimization.

Table 4. Statistics of fours kinds of pair

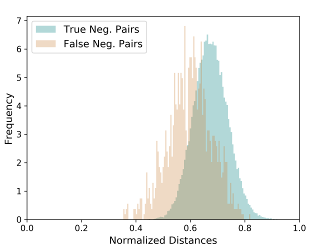
Type	TP	FP	TN	FN
Percentage(%)	60.62	39.38	95.99	4.01



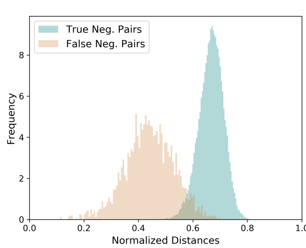
(a) TP-FP distributions of variant



(b) TP-FP distributions of DART



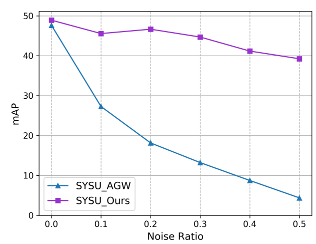
(c) TN-FN distributions of variant



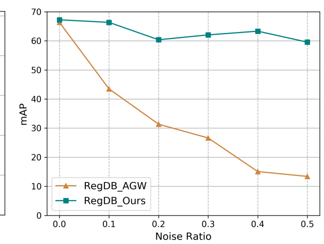
(d) TN-FN distributions of DART

Figure 5. Pairwise distance distributions of TP and FP pairs, TN and FN pairs computed through ADP [26] and DART, respectively.

Robustness against Noisy Correspondence: As discussed in Section 3.3 and 3.4, DART enjoys robustness against noisy correspondence by resorting to the adaptive quadruplet loss (Eq. 10) and pair division module. To show the achieved robustness, we visualize the pairwise distance distribution of TP and FP pairs, and TN and FN pairs on DART compared with the variant which only uses Eq. 9 and vanilla triplet loss. The statistics of the four kinds of pairs and their distribution are shown in Table 4 and Fig. 5. From Fig. 5, one could observe that the variant cannot handle the noisy correspondence. As a result, TP and FP pairs, as well as TN and FN pairs are mixed up. In contrast, DART could correctly distinguished these cases because it will prevent the noisy correspondence from dominating the network optimization. In other words, DART will enforce the distance of TN and FP larger than that of FN and TP during training, thus eliminating the influence of the noisy correspondence.



(a)



(b)

Figure 6. Performance comparisons between DART+AGW and AGW on SYSU-MM01 and RegDB with varying noise ratios.

4.5. Study on the Generalizability

In this section, we verify the generalizability of DART by endowing AGW [28] the robustness to twin noisy labels. As shown in Fig. 6, our method (DART + AGW) performs better than AGW by a considerable performance margin when the noise ratio varies from 0% to 50% with an interval of 10%. This demonstrates the generalizability and robustness of DART.

5. Conclusion

In this paper, we study a new problem in VI-ReID, *i.e.*, twin noisy labels (TNL), which refers to the noisy annotations and noisy correspondence. To solve this problem, we propose DART which estimates the clean confidences of annotation and then rectifies the noisy correspondence. By dividing data pair into four subsets, DART employs a novel dually robust loss for learning with twin noisy labels. We believe this work might remarkably enrich the learning paradigm with noisy labels by simultaneously considering the noisy annotations and accompanying noisy correspondence, especially, in the VI-ReID community. In the future, we plan to explore other scenarios of the twin noisy labels, such as category-level cross-modal retrieve, face recognization, and so on.

6. Acknowledgments

This work was supported in part by NSFC under Grant U21B2040, 62176171, 61836006, and U19A2078; in part by Open Research Projects of Zhejiang Lab under Grant 2021KH0AB02.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, pages 312–321. PMLR, 2019. 4

[2] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, pages 10257–10266, 2020. 2

[3] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *ICLR*, 2020. 1

[4] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *NeurIPS*, 2018. 1

[5] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. *arXiv:1805.08193*, 2018. 2

[6] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8527–8537, 2018. 2, 4, 5

[7] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *ICCV*, pages 16403–16412, 2021. 2

[8] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *CVPR*, pages 5403–5413, 2021. 3

[9] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *NeurIPS*, 2021. 3

[10] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv:2002.07394*, 2020. 2, 3, 4

[11] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, pages 13379–13389, 2020. 2

[12] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, pages 6543–6553, 2020. 2

[13] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 6

[14] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *ICCV*, pages 12046–12055, 2021. 2, 6

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv:1912.01703*, 2019. 6

[16] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv:2007.08199*, 2020. 2

[17] 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 *CVPR*, pages 1522–1531, 2021. 1, 2, 6

[18] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3623–3632, 2019. 1, 2

[19] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *CVPR*, pages 9358–9367, 2019. 2

[20] 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 *CVPR*, pages 618–626, 2019. 2

[21] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *ICCV*, pages 225–234, 2021. 2

[22] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017. 2, 6

[23] 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 *CVPR*, pages 4330–4339, 2021. 1, 2, 6

[24] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jian Cheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[25] Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. Partially view-aligned representation learning with noise-robust contrastive loss. In *CVPR*, June 2021. 3, 4

[26] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, 2021. 1, 2, 3, 6, 8

[27] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, 2020. 2, 6

[28] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 3, 4, 6, 8

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

[30] Mang Ye and Pong C Yuen. Purifynet: A robust person re-identification model with noisy labels. *IEEE Transactions on Information Forensics and Security*, 15:2655–2666, 2020. 1, 3

[31] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *ICCV*, pages 552–561, 2019. 1, 3, 6

[32] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2012. 1

[33] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, 2015. 1

[34] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, pages 4678–4686, 2015. 1