Optimal Transport for Label-Efficient Visible-Infrared Person Re-Identification

Jiangming Wang[1], Zhizhong Zhang[1(✉)], Mingang Chen[2], Yi Zhang[3], Cong Wang[4], Bin Sheng[5], Yanyun Qu[6], and Yuan Xie[1(✉)]

[1] East China Normal University, Shanghai, China
[2] Shanghai Development Center of Computer Software Technology, Shanghai, China
[3] ZheJiang Lab, Hangzhou, China
[4] Huawei Technologies, Hangzhou, China
[5] Shanghai Jiao Tong University, Shanghai, China
[6] Xiamen University, Fujian, China
{51215901073}@stu.ecnu.edu.cn, {zzzhang,yxie}@cs.ecnu.edu.cn, {cmg}@sscenter.sh.cn, {zhangyi620}@zhejianglab.com, {wangcong64}@huawei.com, {shengbin}@sjtu.edu.cn, {yyqu}@xmu.edu.cnn

Abstract. Visible-infrared person re-identification (VI-ReID) has been a key enabler for night intelligent monitoring system. However, the extensive laboring efforts significantly limit its applications. In this paper, we raise a new label-efficient training pipeline for VI-ReID. Our observation is: RGB ReID datasets have rich annotation information and annotating infrared images is expensive due to the lack of color information. In our approach, it includes two key steps: 1) We utilize the standard unsupervised domain adaptation technique to generate the pseudo labels for visible subset with the help of well-annotated RGB datasets; 2) We propose an optimal-transport strategy trying to assign pseudo labels from visible to infrared modality. In our framework, each infrared sample owns a label assignment choice, and each pseudo label requires unallocated images. By introducing uniform sample-wise and label-wise prior, we achieve a desirable assignment plan that allows us to find matched visible and infrared samples, and thereby facilitates cross-modality learning. Besides, a prediction alignment loss is designed to eliminate the negative effects brought by the incorrect pseudo labels. Extensive experimental results on benchmarks demonstrate the effectiveness of our approach. Code will be released at https://github.com/wjm-wjm/OTLA-ReID.

Keywords: VI-ReID, Optimal-Transport, Label-efficient Learning

1 Introduction

Visible-infrared person re-identification (VI-ReID) [31,3,38,29,19,27] has been a key enabler for night intelligent monitoring system. It aims to properly find the target visible/infrared images when given a query image from another modality. Due to the significant difference in sensing processes, visible-infrared heterogeneous images have large appearance variations. Therefore, it's very different from conventional visible ReID problem [39,42,43].

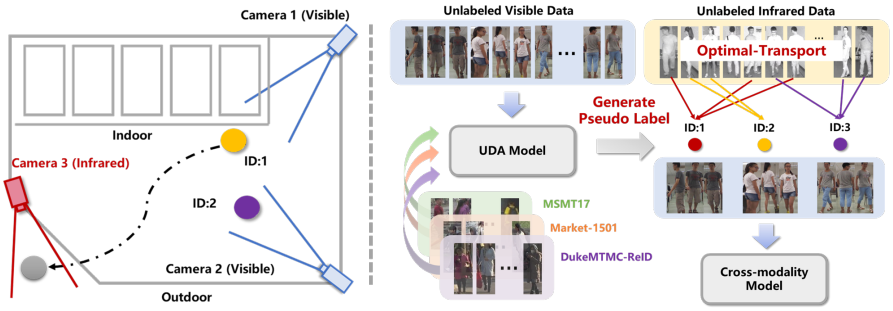

Fig. 1. Left: the real-life ReID application scenario. Right: Our proposed label efficient training pipeline.

Recently, impressive progress [38,39,24,35,37] in VI-ReID has been made to reduce the cross-modality variations. A common practice is to align the visible and infrared images on both image and feature level [37,35]. However, we have noticed that one of the important ingredient to their success is the availability of well-annotated training sets. These training sets need extensive labelling efforts, especially for infrared subsets due to lack of color information. Hence, a critical question comes up: Can we learn a cross-modality model only with one modal supervision or even without supervision?

To this end, we raise a new training pipeline towards label-efficient learning for VI-ReID. One key assumption of our approach is that only visible labels are accessible, or can be produced by self-training strategy from other visible datasets. This idea is inspired by the observation that: 1) The scale of existing cross-modality ReID dataset is relatively small, while visible ReID dataset has rich annotation information; 2) The cost of annotating infrared images is much more expensive, as it is difficult for annotators to recognize the identities without color information. In fact, this setting is also quite common in real-life scenario. For example, in a supermarket as shown in Fig. 1 left, there exits indoor and outdoor cameras which capture visible and infrared images, respectively. Hence, images from two modalities probably contain the same identity. But a ReID system is often deployed in the indoor scene (visible camera) but unprepared for the outdoor scene (infrared camera). It requires us to train a cross-modality model without infrared annotations, or directly avoid labour-extensive annotations by taking advantage of other well-annotated RGB datasets.

Driven by this analysis, our approach includes two key steps as shown in Fig. 1 right. Firstly, we utilize the standard UDA-ReID approach [12] to generate the pseudo labels for visible data by taking knowledge from the rich annotated dataset *e.g.,* Market-1501 [45], DukeMTMC-ReID [25], MSMT17 [32]. Secondly, to establish an explicit connection between cross-modality data, we propose an optimal-transport strategy for assigning the infrared images to the generated visible pseudo classes. In this module, each sample owns a label assignment choice viewed as supplier, and each label requires unallocated images viewed

as demander. By introducing the uniform sample-wise and label-wise prior, we can achieve a desirable assignment plan that allows us to find truly matched visible and infrared samples. To eliminate the negative effects brought by the inaccurate supervised signals, we also propose a prediction alignment learning module, which in practice is a batch-level prediction mix-up and further facilitate the learning modality-invariant representations.

We conduct extensive experiments against state-of-the-arts of three categories (*i.e.*, fully-supervised, unsupervised learning and unsupervised domain adaption methods) on the widely adopted benchmarks for VI-ReID. We empirically find that 1) Taking knowledge from the rich annotated dataset is necessary for label-efficient VI-ReID; 2) Our approach achieves promising results, *i.e.*, 48.2% in term of Rank-1 accuracy on SYSU-MM01 with mere visible ground-truth labels and 29.9% without ground-truth labels. Our contributions can be summarized as follows:

• We propose a new label-efficient learning pipeline which roots from real-world scenario. By taking advantage of rich annotated visible dataset, we produce reliable pseudo labels for RGB images and these labels in turn allow us to train a cross-modality model.

• Two critical modules: Optimal-Transport Label Assignment module (OTLA) and Prediction Alignment Learning (PAL) are proposed. OTLA enables us to assign the infrared images to the generated visible pseudo classes, and thereby establish an explicit connection between visible and infrared data. PAL can reduce the negative effects brought by the inaccurate pseudo labels.

• We provide comprehensive evaluations on this challenge problem. Empirical results show that our approach achieves highly comparable results with fully supervised methods and outperforms recent UDA-ReID and USL-ReID methods.

2 Related Work

Visible-Infrared Person Re-Identification. Visible-infrared person ReID (VI-ReID) aims to match the person images between two modalities. Recently, some works [38,29,19,35,24] try to enhance the feature discrimination by using novel network structures (*e.g.*, graph convolution network and non-local module) or discovering nuanced but discriminative representation. While, another lines of novel works [3,31,15,30] attempt to excavate modality-invariant information by image generation. Besides, [5,40,18] have optimized metric learning items (*e.g.*, Triplet Loss) adapting to cross-modality learning.

Unsupervised Domain Adaptation Person Re-Identification. Unsupervised domain adaptation [23] aims to learn the knowledge of unlabeled target data with help of labeled source data. The recent application of UDA in ReID (UDA-ReID) can be regarded as an open set task, where label spaces between two domains are inconsistent. It can be roughly classified into three categories. The first category [48,49,21] attempts to reduce domain gap by digging up positive or negative pairs from labeled source data or unlabeled target data or both of them. The second category [9,11,12,44] has adopted unsupervised clustering

methods. The last category [7,33,47] wants to learn domain invariant information by mutually generating images from source and target domain.

Unsupervised Learning Person Re-Identification. Unsupervised learning person ReID (USL-ReID) aims to train a model with only unlabeled data. But previous works often restrict the problem to a single modality ReID task. In this setting, most methods [12,28,8,36,17,46] are mainly based on pseudo labels, which establish a bridge with supervised manner. For example, two representative works [12,36] try to obtain pseudo labels with traditional clustering method, DBSCAN or K-means. Besides, some hierarchical clustering ways [17,46] are designed to obtain high-quality pseudo labels.

Optimal Transport. Optimal transport (OT) [4] theory has obtained an increasing attention in the field of machine learning, which is often used to find correspondences with learnable features or measure the distribution distance. M. Asano *et al.* [1] has extended OT to self-supervised learning. In this framework, they alternate between the following two steps: 1) Making use of Sinkhorn-Knopp algorithm to produce pseudo labels for unlabeled data. 2) Doing classification with current pseudo labels. In fact, [1] is a clustering based self-supervised approach, which aims to find a good pretraining model. By contrast, we apply OT for a global data-label assignment problem.

Summary. By reviewing recent studies, VI-ReID often requires extensive labelling efforts. However, collecting a well annotated dataset is time-consuming and laborious. Besides, most UDA-ReID and USL-ReID methods restrict their studies to single modality problem, which can't meet the challenges posed by VI-ReID. Towards the label-efficient learning, Liang *et al.* [16] firstly designed an unsupervised framework by taking advantage of the clustering process. But it is sub-optimal since the rich annotated visible data is not utilized and the heterogeneous pseudo labels is also not well aligned. Considering the above problems, we divide label-efficient learning of VI-ReID into two parts: 1) Producing accurate pseudo visible labels by using recent well-established UDA-ReID or USL-ReID methods. 2) Formulating an optimal-transport task inspired by [1] so as to assign infrared data to visible pseudo classes.

3 Methodology

3.1 Problem Formulation and Overview

Suppose we are given a collection $\mathcal{X} = \{\mathcal{V}, \mathcal{R}\}$ consisting of cross-modality pedestrian images. $\mathcal{V} = \{\boldsymbol{x}_i^v\}_{i=1}^{N_v}$ and $\mathcal{R} = \{\boldsymbol{x}_i^r\}_{i=1}^{N_r}$ denote the visible and infrared images with N_v and N_r samples, respectively. To learn a cross-modality model only with one modal supervision or even without supervision, a natural idea is to utilize the supervision from the well annotated visible ReID dataset. Intuitively, these labeled data allow us to take advantage of UDA-ReID [9,41,21,12,11,44], and hence enable us to produce reliable pseudo labels $\mathcal{Y} = \{\boldsymbol{y}_i^v\}$ for visible subset $\mathcal{V}$. In our implementation, we adopt the SOTA clustering based method SpCL [12] to generate $\mathcal{Y}$, by taking RGB dataset *e.g.,* Market-1501 [45], DukeMTMC-ReID [25], MSMT17 [32] as the source domain, and $\mathcal{V}$ in visible-infrared dataset

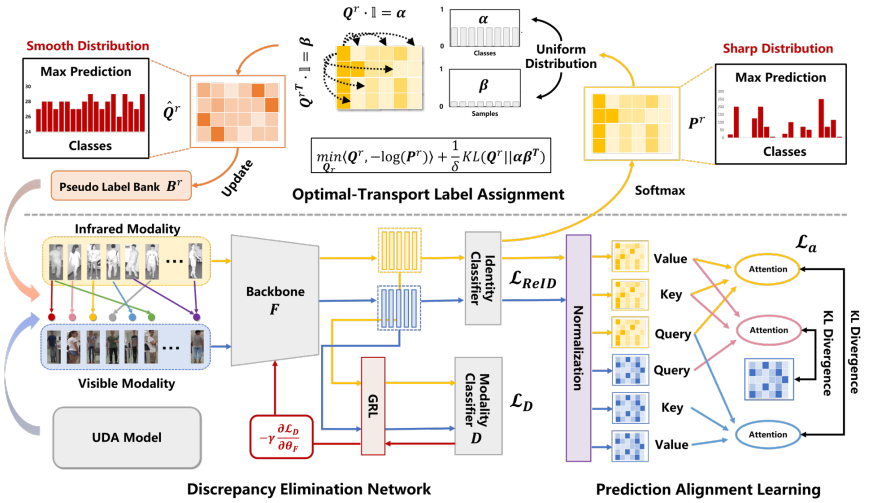

Fig. 2. The pipeline of our framework. Left Bottom: we first use an UDA model to generate the pseudo labels for visible images. Then we take both visible and infrared data into Discrepancy Elimination Network. Upper: The identity prediction of infrared images are sent into Optimal-Transport Label Assignment to assign labels. Right Bottom: The identity predictions are also forwarded into Prediction Alignment Learning to align the mixed predictions, so as to reduce the effects from incorrect pseudo labels.

as the target domain. Since the data in both domains are homogeneous RGB images, falling in the scope of standard UDA-ReID problem, we can obtain relatively reliable pseudo labels.

In the second stage, infrared images are required to be assigned to the generated pseudo labels for cross-modality training. To this end, three key components: Discrepancy Elimination Network (DEN), Optimal-Transport Label Assignment module (OTLA), and Prediction Alignment Learning module (PAL) are proposed. DEN is implemented with a backbone network (*e.g.,* ResNet-50) and a modality classifier (bottom left in Fig. 2), which is served as a feature extractor to reduce the modality gap. OTLA (upper in Fig. 2) is proposed to transport the infrared images to the generated visible pseudo classes. By introducing class-wised uniform distribution $\boldsymbol{\alpha}$ and sample-wised uniform distribution $\boldsymbol{\beta}$, OTLA can effectively produce matched visible and infrared data. PAL (bottom right in Fig. 2) minimizes the KL-divergence between the original predictions and the mixed predictions using a batch-level self-attention technique. We will elaborate each module and illustrate how they cooperate with each other.

3.2 Discrepancy Elimination Network (DEN)

Discrepancy elimination network enables us to reduce the modality gap, served as a strong baseline for learning modality-invariant features. Specifically, given a batch of visible and infrared images, we forward them into a ResNet-50 backbone

$\boldsymbol{F}$ for feature extraction, *i.e.*, $\boldsymbol{f}_i^v = \boldsymbol{F}(\boldsymbol{x}_i^v)$, $\boldsymbol{f}_i^r = \boldsymbol{F}(\boldsymbol{x}_i^r)$. To make both $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^r$ modality-invariant, we deploy a modality classifier $\boldsymbol{D}$ to determine which modality the feature comes from. The learning objective is thus formulated as:

$$\mathcal{L}_D = \max_{\boldsymbol{F}} \min_{\boldsymbol{D}} \mathbb{E}_{\boldsymbol{f}_i^v}[\log(1 - \boldsymbol{D}(\boldsymbol{f}_i^v)] + \mathbb{E}_{\boldsymbol{f}_i^r}[\log(\boldsymbol{D}(\boldsymbol{f}_i^r)]. \tag{1}$$

Note that Eq. (1), in fact, is an adversarial loss widely used in the field of domain adaptation. We achieve Eq. (1) by using a gradient reversal layer (GRL) [10]. During the forward propagation, GRL acts as an identity transform. In the back propagation, GRL filps the gradient of modality classifier (*i.e.* multiply gradient by $-\gamma$) and passes it to the preceding layer. To design DEN, we experimentally find that GRL would degrade the performance in the fully-supervised VI-ReID, but is effective in semi-supervised/unsupervised case. DEN appears to be able to reduce modality discrepancy, especially in the absence of accurate guidance.

3.3 Optimal-Transport Label Assignment (OTLA)

To train a discriminative model for VI-ReID, only using generated pseudo labels $\boldsymbol{y}_i^v$ for visible data is insufficient. For example, Triplet Loss [26] is widely used in VI-ReID community, and its common step is to choose the positive and negative samples to construct the triplet. However, without the matched infrared and visible data, triplet loss can hardly promote the cross-modality matching performance. A intuitive solution is to use the self-training technique to assign labels for infrared images. However, if we send the features $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^r$ into an identity classifier, and use the pseudo visible label $\boldsymbol{y}_i^v$ with standard cross-entropy and triplet loss to optimize it:

$$\mathcal{L}_{\text{V-ReID}} = \mathcal{L}_{\text{Tri}}^v + \mathcal{L}_{\text{CE}}^v, \tag{2}$$

it will lead a so-called degeneration of classifier problem. In this case, most infrared samples are assembled in a few classes. This phenomenon may not happen in single modality training, because Eq.(2) encourages the classifier pay more attention on the visible data while neglecting the discrimination in infrared images. So, if we follow the self-training methods [9], using the maximum value of classifier output $\boldsymbol{p}_i^r$ or clustering result as the infrared labels, the cross-modality learning would be significantly biased to visible data.

To solve such issue, we propose an Optimal-Transport Label Assignment (OTLA) module to find infrared samples associated with visible data. Inspired by [1], we formulate the label assignment task as an optimal transport problem. In our framework, the infrared samples are viewed as suppliers, while the pseudo labels are considered as demands. The goal is to transport samples in suppliers to demands at the lowest cost via an optimal plan $\boldsymbol{Q}^r$. To prevent the degeneration of classifier, we start from two intuitions: first, each infrared image owns an assignment choice that corresponds to a generated pseudo label; second, each generated pseudo label owns approximately the same number of infrared images.

To this end, we define a supplier vector $\boldsymbol{\alpha} \in \mathbb{R}^{N_r}$ indicating that each sample owns a label assignment choice, and a demander vector $\boldsymbol{\beta} \in \mathbb{R}^{N_p}$ indicating the

desired assigned results. Besides, let $\boldsymbol{P}^r \in \mathbb{R}^{N_r \times N_p}$ denote the softmax output of classifier for the infrared images, where N_p stands for the total number of identities. We use $\boldsymbol{P}^r$ to act as as a kind of cost measuring the difficulty of each image assigned to the identity. Thus we can define a label assignment objective:

$$\min_{\boldsymbol{Q}^r} \langle \boldsymbol{Q}^r, -\log(\boldsymbol{P}^r) \rangle + \frac{1}{\delta} KL(\boldsymbol{Q}^r || \boldsymbol{\alpha}\boldsymbol{\beta}^T).$$
$$\text{s.t.} \begin{cases} \boldsymbol{Q}^r \mathbb{1} = \boldsymbol{\alpha}, & \boldsymbol{\alpha} = \mathbb{1} \cdot \frac{1}{N_r}, \\ {\boldsymbol{Q}^r}^T \mathbb{1} = \boldsymbol{\beta}, & \boldsymbol{\beta} = \mathbb{1} \cdot \frac{1}{N_p}, \end{cases} \tag{3}$$

where $\boldsymbol{Q}^r \in \mathbb{R}^{N_r \times N_p}$ represents the plan used for pseudo label assignment, $\langle \cdot \rangle$ denotes the Frobenius dot-product, δ is a hyper-parameter, and KL denotes the KL-divergence. In essence, Eq. (3) is a transport problem, and also a trade-off between prediction and smooth assignment. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represent the class-wise prior uniform distribution vector and the sample-wise prior uniform distribution vector, respectively. Through them, the infrared samples can be forced to be assigned to equally-sized subsets, avoiding samples are grouped together.

However, traditional approaches are not applicable to solve this transport objective due to the large amount of data points and identities. Instead, such constraint leads us to adopt the Sinkhorn-Knopp algorithm [4]. As a result, the optimal solution $\hat{\boldsymbol{Q}}^r$ can be achieved through the iteratively conducted Sinkhorn-Knopp algorithm with a simple matrix scaling operation:

$$\forall i : \boldsymbol{\alpha}_i \leftarrow [(\boldsymbol{P}^r)^\delta \boldsymbol{\beta}]_i^{-1} \quad \forall j : \boldsymbol{\beta}_j \leftarrow [\boldsymbol{\alpha}^T (\boldsymbol{P}^r)^\delta]_j^{-1}, \tag{4}$$

where initialize $\boldsymbol{\alpha}$ with $\frac{1}{N_r} \cdot \mathbb{1}$ and $\boldsymbol{\beta}$ with $\frac{1}{N_p} \cdot \mathbb{1}$. When the iteration meets the termination conditions or exceeds the maximum number, the auxiliary vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed. One primary advantage of this approach is it can equivalently convert Eq. (3) as:

$$\boldsymbol{Q}^r = \text{diag}(\boldsymbol{\alpha})(\boldsymbol{P}^\text{r})^\delta \text{diag}(\boldsymbol{\beta}), \tag{5}$$

where $\text{diag}(\cdot)$ denotes the square diagonal matrix with the elements of vector on the main diagonal.

On this basis, we reassign each sample with the class-wise and sample-wise smooth prior, and hold a pseudo label bank $\boldsymbol{B}^r$ according to the maximum value of optimal plan $\hat{\boldsymbol{Q}}^r$ to store the assigned results. This bank is updated epoch by epoch, and then assigns a reliable label for each infrared image. Finally, we use the pseudo labels of both visible and infrared data for training a standard VI-ReID model with cross-entropy and triplet loss. Our experimental results show that the identities of $\{\boldsymbol{x}_i^r\}_{i=1}^m$ and $\{\boldsymbol{x}_i^v\}_{i=1}^m$ can gradually coincide with each other, and the computational cost of assignment is extremely low.

3.4 Prediction Alignment Learning (PAL)

With the help of pseudo labels, we can sample the cross-modality images belong to the same identity to construct triplet, enabling us to complete the cross

Algorithm 1 Label-efficient VI-ReID

Require: Unlabeled visible-infrared data $\{\mathcal{V}, \mathcal{R}\}$, other labeled visible data $\{\mathcal{V}^{'}\}$.

Using $\mathcal{V}^{'}$ and $\mathcal{V}$ to generate reliable pseudo labels $\{\boldsymbol{y}_i^v\}$ by SpCL [12].
Initialize pseudo label bank $\boldsymbol{B}^r$ for all infrared data.
for *epoch* $= 1 : M$ **do**
  for *batch* $= 1 : N$ **do**
    According to $\{\boldsymbol{y}_i^v\}$ and $\boldsymbol{B}^r$ sample a batch of visible and infrared data.
    Calculate $\mathcal{L}_{\text{ReID}}$, $\mathcal{L}_D$, $\mathcal{L}_a$ with $\{\boldsymbol{y}_i^v\}$ and $\boldsymbol{B}^r$ and update the parameters.
  end for
  Extracting prediction $\boldsymbol{P}^r$ for all infrared data.
  Initialize $\boldsymbol{\alpha}^0$ with $\frac{1}{N_r} \cdot \mathbb{1}$ and $\boldsymbol{\beta}^0$ with $\frac{1}{N_p} \cdot \mathbb{1}$.
  while $||\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^k||_1 < \epsilon$ **do**
    $\forall i : \boldsymbol{\alpha}_i^k \leftarrow [(\boldsymbol{P}^r)^\delta \boldsymbol{\beta}^{k-1}]_i^{-1} \quad \forall j : \boldsymbol{\beta}_j^k \leftarrow [\boldsymbol{\alpha}^{k-1^T} (\boldsymbol{P}^r)^\delta]_j^{-1}$.
  end while
  $\boldsymbol{Q}^r = \text{diag}(\boldsymbol{\alpha})(\boldsymbol{P}^{\text{r}})^\delta \text{diag}(\boldsymbol{\beta})$.
  Using $\boldsymbol{Q}^r$ to update $\boldsymbol{B}^r$.
end for

modality training. However, there are still incorrect labels harming the training process. To eliminate the negative effects brought by incorrect labels, we propose a batch-level prediction mix-up, which aligns the prediction distributions between modalities from a batch perspective.

Specifically, for a batch of samples, we first normalize the prediction of classifier to obtain $\boldsymbol{S}^v \in \mathbb{R}^{B \times N_p}$ and $\boldsymbol{S}^r \in \mathbb{R}^{B \times N_p}$, where the superscript v/r represents visible/infrared modality and B is the number of visible/infrared images. Note that in order to adopt the triplet loss, we typically sample equal-sized cross-modality images according to the pseudo labels and hence the size of $\boldsymbol{S}^r$ and $\boldsymbol{S}^v$ are the same. To encourage the classifier to make consistent predictions on these sampled images, we conduct self-attention by taking $\boldsymbol{S}^v$ as query, and $\boldsymbol{S}^r$ as key and value. Formally, we have:

$$\boldsymbol{S}^{vr} = \text{softmax}\left(\boldsymbol{S}^v (\boldsymbol{S}^r)^T\right) \boldsymbol{S}^r. \tag{6}$$

After that, we compute the KL divergence between the source prediction $\boldsymbol{S}^{vr}$ and the target prediction $\boldsymbol{S}^v$ to get the alignment loss:

$$\mathcal{L}_a^{vr} = KL(\boldsymbol{S}^v || \boldsymbol{S}^{vr}). \tag{7}$$

Intuitively, Eq. (7) forces the visible prediction fused with infrared images to be consistent with $\boldsymbol{S}^v$. Even though there unfortunately exists an incorrect label, self-attention would eliminate its negative effect by emphasising the truly-related samples while neglecting the incorrect ones, by promoting the instance-level alignment to batch-level alignment. This strategy is like the mix-up technique by fusing samples from two modalities in a batch. The mixed prediction is hence encouraged to filter the outliers and reduce the prediction gap between two modalities.

Due to the more noise in $\boldsymbol{S}^r$, it is not wise to design a symmetric loss $\mathcal{L}_a^{rv}$, which may bring permutation to the training process. Instead, we define two mixed prediction $\boldsymbol{S}^{rr}$ and $\boldsymbol{S}^{rv}$:

$$\begin{aligned} \boldsymbol{S}^{rr} &= \text{softmax}\left(\boldsymbol{S}^r(\boldsymbol{S}^r)^T\right)\boldsymbol{S}^r, \\ \boldsymbol{S}^{rv} &= \text{softmax}\left(\boldsymbol{S}^r(\boldsymbol{S}^v)^T\right)\boldsymbol{S}^v. \end{aligned} \tag{8}$$

With them, we can finally obtain another alignment loss:

$$\mathcal{L}_a^{rv} = KL(\boldsymbol{S}^{rv}||\boldsymbol{S}^{rr}). \tag{9}$$

The reason for this design is that two mixed predictions are less effected by the incorrect labels. Based on the above analysis, the proposed prediction alignment loss $\mathcal{L}_a$ is formulated as:

$$\mathcal{L}_a = \lambda_a^{vr}\mathcal{L}_a^{vr} + \lambda_a^{rv}\mathcal{L}_a^{rv}. \tag{10}$$

where λ_a^{vr} and λ_a^{rv} are the coefficients (set to 0.1 and 0.5 in our experiments).

3.5 Optimization

The training process is summarized in Algorithm 1. The total training loss $\mathcal{L}$ can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ReID}} + \lambda_1\mathcal{L}_D + \lambda_2\mathcal{L}_a. \tag{11}$$

where $\mathcal{L}_{\text{ReID}}$ denotes the standard cross-entropy and triplet loss of both modalities, λ_1 and λ_2 are trade-off hyperparameters (empirically set them to 1.0).

4 Experiments

In this section, we conduct extensive experiments to provide a basic yet comprehensive evaluation on this new challenge problem. We report the results under two experimental settings *i.e.,* unsupervised VI-ReID (USVI-ReID, visible labels are generated by SpCL [12]), semi-supervised VI-ReID (SSVI-ReID, with ground-truth visible labels). For USVI-ReID and SSVI-ReID, any ground-truth label of infrared images is **inaccessible** during the training process.

4.1 Experimental Settings

Datasets. The proposed methods are evaluated on two widely adopted benchmarks **SYSU-MM01** [34] and **RegDB** [22]. Specifically, SYSU-MM01 is a large-scale dataset which is collected by four RGB and two infrared cameras from both indoor and outdoor environments. It composed of 287,628 visible images and 15,792 infrared images for 491 different identities. RegDB is collected by two aligned cameras (one visible and one infrared), and it includes 412 identities, where each identity has 10 infrared images and 10 visible images.

Evaluation Metrics. On both datasets, we follow the popular protocols [38] for evaluation, in which cumulative match characteristic (CMC) and mean average precision (mAP) are adopted. SYSU-MM01 contains two different testing settings, *i.e., all-search* and *indoor-search* mode. For all-search mode, the gallery consists of all visible images (captured by CAM1, CAM2, CAM3, CAM4) and the query is composed of all infrared samples (captured by CAM5, CAM6). For indoor-search mode, images captured only from indoor scene are adopted, excluding CAM4 and CAM5. On both search mode, the proposed method is evaluated under single-shot setting. For RegDB [22], we report the average result by randomly splitting of training and testing set 10 times.

4.2 Implementation Details

Training. We implement our model using MindSpore and PyTorch on one NVIDIA TITAN RTX. The batch size is fixed to 64 for all experiments. With the pseudo labels, in a batch we sample 4 different identities, and each identity includes 8 visible images and 8 infrared images. The model is optimized by Adam optimizer with an initial learning rate of 3.5×10^{-3}. The learning rate is incorporated with a warm-up strategy [20] and decays 10 times at the 20-th and the 50-th epoch. The total of training epochs is set to 80. All the pedestrian images are resized to 288×144. The margin ρ of triplet loss is set to 0.3. The δ of OTLA is fixed to 25. In the training stage, the input images are randomly flipped and erased with 50% probability, while visible images are extra randomly grayscale with 50% probability.
Critical Architectures. We adopt ResNet-50 [13] pretrained on ImageNet [6] as backbone, where last stride size of is set to 1. The modality classifier in DEN is implemented with three FC layers and a BN layer [14] is added before the output. The GRL [10] is a non-parametric module and $\gamma = 2/(1+\exp(-\tau\frac{iter}{maxiter})) - 1$ controls the scale of the reversed gradient. τ is fixed to 10 and $maxiter$ is set to 10000. The $iter$ linearly increases as the training goes on.

4.3 Main Results

We compare our approach with four related ReID settings to demonstrate its effectiveness, *i.e.,* fully-supervised VI-ReID (SVI-ReID), unsupervised VI-ReID (USVI-ReID), unsupervised domain adaptation ReID (UDA-ReID) and unsupervised learning ReID (USL-ReID). For UDA-ReID methods, we use ground-truth labeled visible data as source domain and unlabeled infrared data as target domain. For USL-ReID methods, we use both unlabeled visible and infrared data to train the model. The main results are shown in Tab. 1.
Comparison with Unsupervised Methods. H2H [16] is a representative unsupervised VI-ReID method most relevant with our approach. However, it ignores the rich annotated visible data and the heterogeneous pseudo labels are not well aligned, leading to a inferior performance. Other unsupervised methods are designed for single-modality ReID task, so it is somewhat unfair to directly

Table 1. Comparisons with SOTA methods on SYSU-MM01 (single-shot) and RegDB, including unsupervised domain adaptation ReID (UDA-ReID), unsupervised ReID (USL-ReID), fully-supervised VI-ReID (SVI-ReID) and unsupervised VI-ReID (USVI-ReID). All methods are measured by CMC(%) and mAP(%). † indicates we re-implement the result with official code. ‡ indicates the results are copied from [16].

Settings			SYSU-MM01				RegDB			
			All Search		Indoor Search		Visible2Thermal		Thermal2Visible	
Type	Method	Venue	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
UDA-ReID	SSG‡[9]	ICCV'19	2.3	12.7	-	-	2.2	2.9	-	-
	ECN‡[48]	CVPR'19	8.1	5.0	-	-	1.9	3.2	-	-
	D-MMD†[21]	ECCV'20	12.5	10.4	19.0	15.4	2.2	3.7	2.0	3.6
	MMT†[11]	ICLR'20	13.9	8.4	21.0	15.3	5.3	7.1	11.0	12.1
	SpCL(UDA)†[12]	NIPS'20	15.1	6.5	19.5	12.1	3.3	4.3	8.4	9.5
	GLT†[44]	CVPR'21	7.7	9.5	12.1	18.0	2.9	4.5	6.3	7.6
USL-ReID	BUC†[17]	AAAI'19	8.2	3.2	12.5	6.0	4.7	4.5	8.8	6.0
	SpCL(USL)†[12]	NIPS'20	18.7	11.4	27.1	20.9	20.6	17.3	19.0	16.6
	MetaCam†[36]	CVPR'21	14.7	9.3	23.9	17.1	23.1	17.5	20.9	16.5
	HCD†[46]	ICCV'21	18.0	17.9	24.4	28.8	10.8	12.3	12.4	13.7
SVI-ReID	JSIA-ReID[29]	AAAI'20	38.1	36.9	43.8	52.9	48.5	49.3	48.1	48.9
	Hi-CMD[3]	CVPR'20	34.9	35.9	-	-	70.9	66.0	-	-
	AGW[39]	TPAMI'21	47.5	47.7	54.17	63.0	70.1	66.4	70.5	65.9
	NFS[2]	CVPR'21	56.9	55.5	62.8	69.8	80.5	72.1	78.0	69.8
	LbA[24]	ICCV'21	55.4	54.1	58.5	66.3	74.2	67.6	72.4	65.5
	CAJL[37]	ICCV'21	69.9	66.9	76.3	80.4	85.0	79.1	84.8	77.8
	MPANet[35]	CVPR'21	70.6	68.2	76.7	81.0	83.7	80.9	82.8	80.7
USVI-ReID	H2H[16]	TIP'21	25.5	25.2	-	-	14.1	12.3	13.9	12.7
	Ours	-	**29.9**	**27.1**	**29.8**	**38.8**	**32.9**	**29.7**	**32.1**	**28.6**
SSVI-ReID	Ours	-	**48.2**	**43.9**	**47.4**	**56.8**	**49.9**	**41.8**	**49.6**	**42.8**

compare them, since most of them don't consider the cross-modality discrepancy. We report here because very few methods have studied this problem before.

Comparison with Unsupervised Domain Adaptation Methods. It appears that recent state-of-the-art UDA-ReID methods cannot effectively deal with the huge modality discrepancy. Notice that the supervised signal used in UDA-ReID is even stronger than ours, but the highest accuracy is much lower than our method. That indicates our approach is able to learn robust multi-modality representation, significantly outperforming all UDA-ReID methods. On the other hand, USL-ReID methods appear to achieve better results than UDA-ReID approaches. We conjecture this is because most UDA-ReID methods [9,21,11,44] rely heavily on the labeled source domain which drives the model to overfit on visible data. However, USL-ReID methods tend to fuse two modal data so as to achieve better results than UDA-ReID methods.

Comparison with Fully-supervised Methods. Surprisingly, our approach only with ground-truth visible data outperforms several fully-supervised VI-ReID methods on SYSU-MM01 dataset, and achieves closed results on RegDB. Such phenomenon indicates label information of infrared images could be learned from optimal transport assignment. Besides, we should admit there is still a large gap between our method and SOTA fully-supervised results.

Table 2. Ablation study in terms of CMC(%) and mAP(%) on SYSU-MM01.

Order	Approach				All Search							
					USVI-ReID				SSVI-ReID			
	$\mathcal{L}_{\text{V-ReID}}$	$\mathcal{L}_D$	$\mathcal{L}_a$	$OTLA$	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
1	✓	-	-	-	12.62	41.91	57.27	12.73	12.25	46.49	62.24	14.66
2	✓	✓	-	-	16.62	49.91	64.53	15.94	23.69	59.56	73.10	24.71
3	✓	-	✓	-	12.65	42.39	57.03	12.81	13.86	46.67	61.87	14.72
4	✓	✓	-	✓	20.90	59.53	73.86	19.83	33.89	73.89	85.49	32.44
5	✓	-	✓	✓	19.64	61.16	77.31	19.74	36.31	77.31	86.93	34.66
6	✓	✓	✓	✓	29.98	71.79	83.85	27.13	48.15	85.30	92.64	43.86

4.4 Ablation Study

In this subsection, we conduct ablation study to show the effectiveness of each component in our approach. We firstly clarify various settings. $OTLA$ indicate whether the OTLA mechanism is used. $\mathcal{L}_{\text{V-ReID}}$ is visible basic ReID loss functions defined in Sec. 3. As shown in Tab. 2, the main observations are:
(1) The modality classifier in semi-supervised setting works well, which brings improvement of 11.44%@Rank-1 and 10.05%@mAP (see 1^{st} row and 2^{nd} row). When combined with $OTLA$ and $\mathcal{L}_a$, it also boosts huge performance of unsupervised setting (see 5^{th} and 6^{th} row).
(2) Though noisy in the first few epochs, the pseudo labels for infrared images can be gradually rectified through the proposed OTLA (see Fig. 4 Left Upper). It also lays a crucial and solid foundation, where removing this technique leads to a dramatic performance drop (see 2^{nd} row and 4^{th} row, 3^{rd} row and 5^{th} row).
(3) Prediction alignment learning loss significantly boosts the performance when combined with $OTLA$ and $\mathcal{L}_D$ (see Fig. 5, 4^{th} row and 6^{th} row). That indicates a further promotion would be expected when aligning the predictions between two modalities. It seems that the batch-level mix-up can eliminate the negative effects brought by incorrect pseudo labels.

4.5 Discussion

Effects of RGB Source Domain. We analysis the effects of various source RGB domains in SpCL (*e.g.,* Market-1501 [45], DukeMTMC-ReID [25] and MSMT17 [32]). The results are shown in Fig. 3 Right. X-axis 'SYSU' indicates we use the USL mode of SpCL to generate the pseudo labels and hence only the SYSU-MM01 data is involved for training. It seems that using annotated visible data achieves better results and Market-1501 is the most effective domain. The reason and more discussion can be seen in supplementary materials.
Performance on SVI-ReID Setting. Since modality classifier and prediction alignment loss can also be deployed under fully-supervised setting, we conduct additional experiments to study their effects. As shown in Fig. 3 Left, we observe that modality classifier seems to be helpless under fully-supervised setting, while PAL loss consistently gains obvious promotion. It appears that some tricks in semi-supervised or unsupervised setting may fail in supervised setting, which motivates us to highlight their differences.

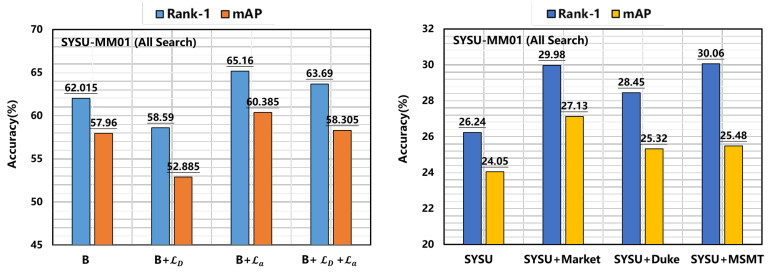

Fig. 3. Left: The supervised ablation results of discriminative loss $\mathcal{L}_D$ and prediction alignment loss $\mathcal{L}_a$ (**B** means baseline model). Right: The effects of source RGB datasets.

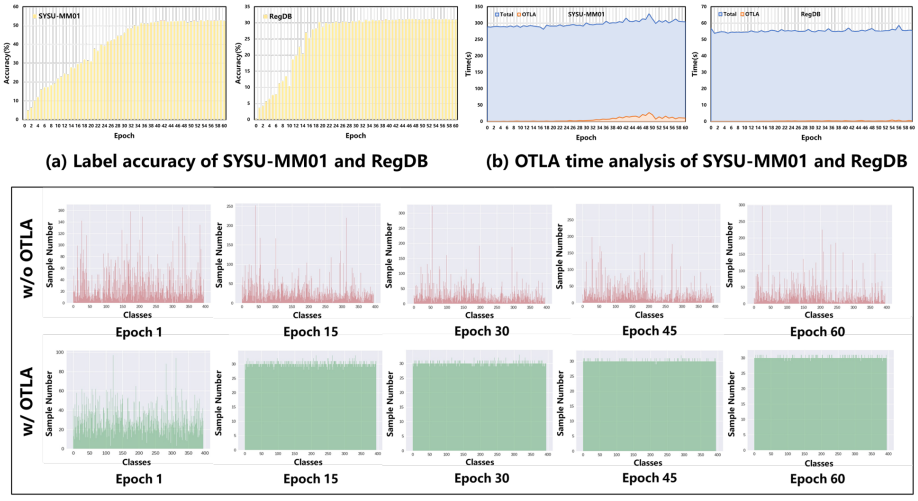

Fig. 4. Left upper: infrared pseudo label assignment accuracy of semi-supervised setting produced by OTLA epoch by epoch. Right upper: the elapsed time of OTLA and total training process in each epoch on SYSU-MM01 and RegDB. Bottom: the assigned infrared label distribution as the training goes on (green area means pseudo label distribution w/ OTLA, red area means pseudo label distribution w/o OTLA).

OTLA Time Analysis. As illustrated in the Fig. 4 (b), we summarize the total training time and OTLA running time for each epoch. For SYSU-MM01 and RegDB, the average elapsed time of OTLA is 6.832s and 0.353s, which has merely occupied 2.293% and 0.639% of the total training time per epoch. Therefore, the computational cost of OTLA seems to be negligible.

Label Distribution of Infrared Images. As shown in Fig. 4(a), the pseudo label accuracy of semi-supervised setting is iteratively improved as the training continues. It can achieve about 50% accuracy on SYSU-MM01 and 30% accuracy

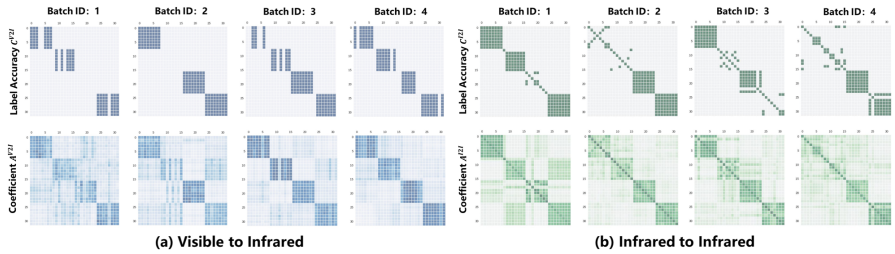

Fig. 5. Visualization of self-attention matrix in prediction alignment learning with random batches. (a) Visible to Infrared. (b) Infrared to Infrared. The upper figures are ground-truth binary matrices and the bottom are our learned attention map in PAL.

on RegDB. From Fig. 4(c), it seems that OTLA can alleviate the degeneration of the classifier as training goes on. Without OTLA the degradation of classifier is significant, *i.e.,* the pseudo label distribution is sharp during the training, indicating that the classifier can not distinguish the identities of infrared images.

Visualization of Prediction Alignment Learning. We visualize the attention map to help understand the influence of prediction alignment learning. As shown in Fig. 5, we draw the attention coefficients $\boldsymbol{A}^{V2I} = \text{softmax}(\boldsymbol{S}^v(\boldsymbol{S}^r)^T)$ and $\boldsymbol{A}^{I2I} = \text{softmax}(\boldsymbol{S}^r(\boldsymbol{S}^r)^T)$ in PAL. To show its effectiveness, we also visualize the binary matrices $\boldsymbol{C}^{V2I}$ and $\boldsymbol{C}^{I2I}$ using ground-truth labels in the upper of Fig. 5, where 1 (fill color) indicates two samples share same identity, and 0 (not fill color) otherwise. From this figure, we can find that the learned attention matrices are consistent with the binary ground-truth label accuracy matrices, which indicates our prediction alignment mechanism can more or less filter the incorrect labels and align the cross-modality predictions.

5 Conclusion

In this paper, we raise a novel label-efficient learning pipeline for VI-ReID, where the visible labels can be produced by UDA-ReID approach with the help of rich annotated RGB datasets. In this setting, we propose a Discrepancy Elimination Network to reduce modality gap. An Optimal-Transport Label Assignment mechanism is designed to uniformly assign labels for infrared images and thereby connect two kinds of modal data. We also propose a Prediction Alignment Learning to eliminate the negative effect brought by incorrect assignment. Extensive experimental results highlight the state-of-the-art performance of our approach. Finally, we hope our study can help researchers to understand the VI-ReID problem from a new perspective.

Acknowledgements: This work is supported by grants from the National Key Research and Development Program of China (2021ZD0111000), National Natural Science Foundation of China No.62106075, 62176092, Shanghai Science and Technology Commission No.21511100700, Natural Science Foundation of Shanghai (20ZR1417700), CAAI-Huawei MindSpore Open Fund.

References

1. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. ICLR (2020) 4, 6
2. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for rgb-infrared person re-identification. In: CVPR. pp. 587–597 (2021) 11
3. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: CVPR. pp. 10257–10266 (2020) 1, 3, 11
4. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. NIPS **26**, 2292–2300 (2013) 4, 7
5. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: IJCAI. vol. 1, p. 6 (2018) 3
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE (2009) 10
7. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR. pp. 994–1003 (2018) 4
8. Ding, Y., Fan, H., Xu, M., Yang, Y.: Adaptive exploration for unsupervised person re-identification. TOMM **16**(1), 1–19 (2020) 4
9. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: CVPR. pp. 6112–6121 (2019) 3, 4, 6, 11
10. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. pp. 1180–1189. PMLR (2015) 6, 10
11. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. ICLR (2020) 3, 4, 11
12. Ge, Y., Zhu, F., Chen, D., Zhao, R., Li, H.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. NIPS (2020) 2, 3, 4, 8, 9, 11
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 10
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456. PMLR (2015) 10
15. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: AAAI. vol. 34, pp. 4610–4617 (2020) 3
16. Liang, W., Wang, G., Lai, J., Xie, X.: Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. TIP **30**, 6392–6407 (2021) 4, 10, 11
17. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: AAAI. vol. 33, pp. 8738–8745 (2019) 4, 11
18. Liu, H., Tan, X., Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. TMM **23**, 4414–4425 (2020) 3
19. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: CVPR. pp. 13379–13389 (2020) 1, 3
20. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. TMM **22**(10), 2597–2609 (2019) 10

21. Mekhazni, D., Bhuiyan, A., Ekladious, G., Granger, E.: Unsupervised domain adaptation in the dissimilarity space for person re-identification. In: ECCV. pp. 159–174. Springer (2020) 3, 4, 11
22. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017) 9, 10
23. Pan, S.J., Yang, Q.: A survey on transfer learning. TKDE **22**(10), 1345–1359 (2009) 3
24. Park, H., Lee, S., Lee, J., Ham, B.: Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In: CVPR. pp. 12046–12055 (2021) 2, 3, 11
25. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV. pp. 17–35. Springer (2016) 2, 4, 12
26. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015) 6
27. Tian, X., Zhang, Z., Lin, S., Qu, Y., Xie, Y., Ma, L.: Farewell to mutual information: Variational distillation for cross-modal person re-identification. In: CVPR. pp. 1522–1531 (2021) 1
28. Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: CVPR. pp. 10981–10990 (2020) 4
29. Wang, G.A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z.G.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: AAAI. vol. 34, pp. 12144–12151 (2020) 1, 3, 11
30. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: CVPR. pp. 3623–3632 (2019) 3
31. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: CVPR. pp. 618–626 (2019) 1, 3
32. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR. pp. 79–88 (2018) 2, 4, 12
33. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR. pp. 79–88 (2018) 4
34. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: ICCV (2017) 9
35. Wu, Q., Dai, P., Chen, J., Lin, C.W., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover cross-modality nuances for visible-infrared person re-identification. In: CVPR. pp. 4330–4339 (2021) 2, 3, 11
36. Yang, F., Zhong, Z., Luo, Z., Cai, Y., Lin, Y., Li, S., Sebe, N.: Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In: CVPR. pp. 4855–4864 (2021) 4, 11
37. Ye, M., Ruan, W., Du, B., Shou, M.Z.: Channel augmented joint learning for visible-infrared recognition. In: ICCV. pp. 13567–13576 (2021) 2, 11
38. Ye, M., Shen, J., J. Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: ECCV. pp. 229–247. Springer (2020) 1, 2, 3, 10
39. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. TPAMI (2021) 1, 2, 11
40. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI. vol. 1, p. 2 (2018) 3

41. Zhai, Y., Ye, Q., Lu, S., Jia, M., Ji, R., Tian, Y.: Multiple expert brainstorming for domain adaptive person re-identification. In: ECCV. pp. 594–611. Springer (2020) 4
42. Zhang, Z., Xie, Y., Li, D., Zhang, W., Tian, Q.: Learning to align via wasserstein for person re-identification. TIP **29**, 7104–7116 (2020) 1
43. Zhang, Z., Xie, Y., Zhang, W., Tang, Y., Tian, Q.: Tensor multi-task learning for person re-identification. TIP **29**, 2463–2477 (2019) 1
44. Zheng, K., Liu, W., He, L., Mei, T., Luo, J., Zha, Z.J.: Group-aware label transfer for domain adaptive person re-identification. In: CVPR. pp. 5310–5319 (2021) 3, 4, 11
45. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV. pp. 1116–1124 (2015) 2, 4, 12
46. Zheng, Y., Tang, S., Teng, G., Ge, Y., Liu, K., Qin, J., Qi, D., Chen, D.: Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In: CVPR. pp. 8371–8381 (2021) 4, 11
47. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: ECCV. pp. 172–188 (2018) 4
48. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: CVPR. pp. 598–607 (2019) 3, 11
49. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Learning to adapt invariance in memory for person re-identification. TPAMI **43**(8), 2723–2738 (2020) 3