Coarse-to-Fine Person Re-Identification with Auxiliary-Domain Classification and Second-Order Information Bottleneck

Anguo Zhang[1,2] Yueming Gao[1,2] Yuzhen Niu*,[3] Wenxi Liu[3] Yongcheng Zhou[1]
[1] *College of Physics and Information Engineering, Fuzhou University*
[2] *Key Laboratory of Medical Instrumentation and Pharmaceutical Technology of Fujian Province*
[3] *College of Mathematics and Computer Science, Fuzhou University*

Abstract

Person re-identification (Re-ID) is to retrieve a particular person captured by different cameras, which is of great significance for security surveillance and pedestrian behavior analysis. However, due to the large intra-class variation of a person across cameras, e.g., occlusions, illuminations, viewpoints, and poses, Re-ID is still a challenging task in the field of computer vision. In this paper, to attack the issues concerning with intra-class variation, we propose a coarse-to-fine Re-ID framework with the incorporation of auxiliary-domain classification (ADC) and second-order information bottleneck (2O-IB). In particular, as an auxiliary task, ADC is introduced to extract the coarse-grained essential features to distinguish a person from miscellaneous backgrounds, which leads to the effective coarse- and fine-grained feature representations for Re-ID. On the other hand, to cope with the redundancy, irrelevance, and noise contained in the Re-ID features caused by intra-class variations, we integrate 2O-IB into the network to compress and optimize the features, without increasing additional computation overhead during inference. Experimental results demonstrate that our proposed method significantly reduces the neural network output variance of intra-class person images and achieves the superior performance to state-of-the-art methods.

1. Introduction

Image-based person re-identification (Re-ID) aims to retrieve a particular person from a high volume of person images captured by different cameras. Considering its crucial applications in public security surveillance, pedestrian behavior analysis, etc., Re-ID has received increasing attention in recent years.

* Yuzhen Niu is the corresponding author (e-mail: yuzhenniu@gmail.com).

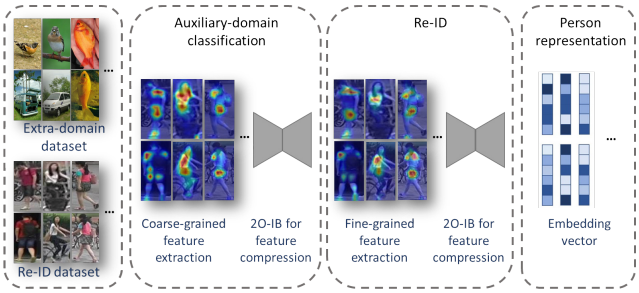

Figure 1. We present a novel coarse-to-fine Re-ID framework. Assisted by the task of *auxiliary-domain classification* (ADC), our Re-ID framework is enabling to learn from the coarse-grained to the fine-grained features for distinguishing persons. Specifically, an extra-domain dataset and a Re-ID dataset are jointly utilized in ADC. *Second-order information bottleneck* (2O-IB) module compresses the redundant information and noise in the features to obtain a more concise and core representation. Last, the computed embedding vector for person representation is generated for Re-ID.

Owing to the ability of extracting excellent representative features as well as the outstanding invariance embedding capability, convolutional neural networks (CNNs) have become the predominant choices for Re-ID. CNNs-based Re-ID methods can be roughly categorized into 1) the representation learning methods that use classification losses as the proxy targets to learn the person embeddings [12, 43, 15, 26]; 2) the metric learning methods that aim to learn the similarity between two person images so that images of the same person have greater similarity than those of different persons [25, 23, 22]; 3) the attribute learning methods that use the local attributes, such as clothing color, hair length/color, hat, and backpack, to identify a person [18, 19, 40, 31]; 4) the local feature learning methods that use image dividing, skeleton key point positioning, and posture correction to learn the features of each part of the person image, and then obtains the global features of the person through feature fusion [17, 37, 32, 33]; and 5) the generative adversarial network (GAN) based methods that expand and augment person images in the training dataset to obtain

more generalized CNN models [8, 41].

Despite plenty of research progress, Re-ID is still a challenging task, because there exist significant intra-class variations in the person images captured by different cameras, due to the changes in the background, illumination, viewpoint, and human poses. Learning robust Re-ID representations against intra-class variations has been an attractive research topic. To address these concerns, the attention-based methods [24, 15, 13, 2] tend to highlight the informative features and suppress the noise, while [14, 8, 41] apply GANs to synthesize training data to cover the changes in different scenes. However, these methods have not thoroughly resolve the concerns of intra-class variations, since the person images may be severely contaminated by miscellaneous backgrounds. Besides, the intra-class variations tend to make the Re-ID features vulnerable towards the low-level data uncertainty. Moreover, the complexity of existing models usually increase the computational overhead in the inference stage.

To cope with the concerns of intra-class variations, in this paper, we present an end-to-end hierarchical coarse-to-fine Re-ID framework for person Re-ID, as shown in Figure 1. In the coarse stage, the framework focuses on extracting features to answer the question, "how to describe a person in an image and what are the characteristics of a person different from other objects?" In the fine stage, it focuses on the question, "how to distinguish different persons in different images?" These two questions are formulated as a cascade of auxiliary-domain classification (ADC) task and the Re-ID task. Effective coarse- and fine-grained feature representations for Re-ID can be learned through multi-task learning of these two tasks. As shown in Figure 1, the proposed hierarchical framework is composed of the coarse-grained feature extraction (CGFE) module, the fine-grained feature extraction (FGFE) module, as well as our proposed second-order information bottleneck (2O-IB) layers.

In order to reduce the impact of miscellaneous background objects, we propose a simple yet effective solution. Since all the images in the Re-ID dataset are person images, we leverage an extra-domain dataset, composed of images with non-person objects, to jointly train the CGFE and the first 2O-IB layer in ADC. In this way, the ADC task can reduce the influence of intra-class variations caused by the distractors, such as background, occlusion, and illumination in the images.

To further reduce the negative impact of data stochasticity, redundancy, and noise within data representation, we present a second-order information bottleneck, which is inspired by the information bottleneck (IB) proposed by Tishby *et al.* [21]. It has been widely applied to compress the stochasticity of the representation in the input signal, and improve the generalization ability of deep learning models. The proposed 2O-IB layers are introduced before the classification layer that follows the CGFE module and before the embedding output layer that follows the FGFE module to compress the stochasticity of features and strengthen the representativeness for the embedding vectors of the same person.

It is worth mentioning that the ADC task, 2O-IB layers, and the extra-domain dataset are only used in the training stage. On the inference stage, the ADC task and the extra-domain dataset are not necessary for Re-ID and the IB layers can be equivalently replaced by fully-connected layers, thereby avoiding increasing the inference cost.

In summary, the main contributions of this paper are four-fold:

1) We propose an end-to-end hierarchical coarse-to-fine framework for person Re-ID. The hierarchical model can realize person feature extraction and information compression from coarse-grained to fine-grained, and address the challenges of intra-class variations.

2) We formulate an auxiliary-domain classification task for the coarse-grained feature extraction. The ADC task can help the model learn how to extract the most essential features of "person" that are different from other objects by introducing an extra image classification dataset, thereby improving the Re-ID model's ability to overcome the intra-class variations.

3) We propose a 2O-IB method that can reduce the impact of data stochasticity, redundancy, noise, and intra-class variation on data representation. The 2O-IB layers are only used during training and can be equivalently replaced by fully-connected layers during inference, which saves the extra inference cost.

4) Experimental results show that the proposed hierarchical framework for Re-ID achieves superior performance to state-of-the-art Re-ID methods.

2. Related Works and Preliminaries

2.1. Related Works

Effective feature representation for high-accuracy discrimination of people images is still challenging. Recently, attention mechanism [24, 15, 13, 2] receives increasingly focus since it can highlight the informative features (e.g., spatial locations and human pose) and suppress the noisy parts (e.g., background and illumination).

Zhang *et al.* [35] proposed a two-stream network to solve the problem of changes in illumination which may affect the people representation of Re-ID methods significantly. The proposed network separates Re-ID features from lighting features to enhance the illumination robustness.

Liu *et al.* [14] considered the challenge of Re-ID robustness caused by human pose variations, and proposed a pose-transferrable framework which utilizes pose-

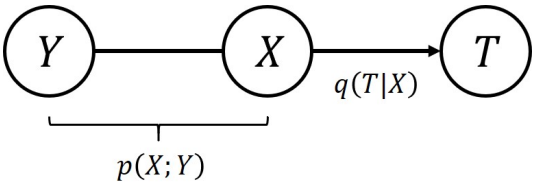

Figure 2. Probabilistic relation in the information bottleneck (IB) framework, where p denotes a fixed distribution probability, q denotes parametric distribution probability to be optimized. Source X, target Y and compression T form a Markov chain $Y \leftrightarrow X \leftrightarrow T$ with the joint distribution $p(X;Y)q(T|X)$. IB tries to find a compression representation T of X, i.e., minimize the mutual information $I(X;T)$, while preserving maximal relevance about Y, i.e, maximizing $I(T;Y)$ with a trade-off balance $\beta > 0$.

transferred sample augmentations to enhance the pose robustness. They trained a generator-guider-discriminator network to generate more novel samples with rich pose variations, and then added these samples into the target dataset to facilitate robust training. Gao *et al.* [4] proposed a pose-guided visible part matching method to learn the discriminative features with pose-guided attention and self-mines the part visibility. The proposed method exploits discriminative local features and estimates whether a part suffers the occlusion or not.

He *et al.* [5] considered another challenge, i.e., person occlusion caused by various obstacles, and proposed an alignment-free model for this scenario. They first leveraged the fully convolutional network and pyramid pooling to extract spatial pyramid features, and then developed an alignment-free matching approach to compute matching scores between occluded persons. They designed an occlusion-sensitive foreground probability generator to focus on clean human body parts. Following the GAN-based methods used for Re-ID tasks, Huang *et al.* [8] proposed an adversarially occluded samples method to augment the variation of training data for improving the generalization capability of Re-ID models.

In addition to the above problems, noisy training data may also affect the performance of deep learning based Re-ID models. Yu *et al.* [27] proposed a DistributionNet to solve two types of noise, i.e., label noise caused by human annotator errors and data outliers caused by person detector errors or occlusion. The DistributionNet models the person image as a Gaussian distribution with its variance representing the stochasticity of the extracted features.

2.2. Information Bottleneck

For a provided dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N) | x_i \in X, y_i \in Y, i = 1, \ldots, N\}$, where X, Y are two random variables that take values x, y from two finite sets with distribution $p(x), p(y)$, respectively. N is the number of i.i.d. training sample pairs. We use p to denote the distributions that are given to be fixed and use q to denote the distributions that are updated throughout the optimization process. Variable X and output Y follow the marginal distribution $p(X)$ and $p(Y)$, their joint distribution is denoted by $p(x, y)$. Let T be the intermediately compressed representation of X in the way of Markov chain $Y \leftrightarrow X \leftrightarrow T$. $q(t|x)$ denotes the conditional probability for a given encoding map from X to the variable T.

The IB principle [21] (as shown in Figure 2) aims to extract a compression T of X that is relevant for determining Y. T is found to maximize the relevance with Y, formulated in terms of the mutual information $I(Y;T)$, while under the constraint of input compression which formulated in terms of $I(X;T)$. This constrained optimization problem can be written by

$$\arg_{T \in \Delta} \max I(T;Y) \text{ s.t. } \mathrm{I}(X;T) \leq r, \tag{1}$$

where Δ is the variable space of T.

Rather than solving the constrained optimization problem in (1), the general IB method tries to minimize the so-called IB Lagrangian cost function:

$$\min_{q(t|x)} \mathcal{L}_{\mathrm{IB}}(T) = I(T;Y) - \beta I(X;T), \tag{2}$$

where β is the Lagrange multiplier, a non-negative free parameter that controls the trade-off between the input compression and output determination. Considering that the mutual information between two probability functions is non-negative, thus, minimizing $\mathcal{L}_{\mathrm{IB}}(T)$ is to minimize $I(X;T)$ while maximizing $I(T;Y)$ at the same time.

3. Proposed Method

In order to alleviate the influence of intra-class variations, we proposed an end-to-end hierarchical coarse-to-fine deep learning framework for Re-ID. In the coarse stage, an object-level feature representation is obtained with the help of an extra-domain dataset by using the ADC task. Specifically, the classification task, trained on both the Re-ID dataset and the extra-domain dataset, can improve the model's ability to extract more accurately person features. On this basis, the fine-grained feature representation is obtained in the fine stage, where only the Re-ID dataset is used for training with triple loss. In addition, to further address the intra-class variation, the second-order IB method is proposed and applied before the classification layer and embedding output layer. The 2O-IBs are calculated separately in these two stages, which can keep only the information most relevant to the target output and can obtain intra-class variation tolerant feature representation.

As shown in Figure 3, the proposed hierarchical framework is composed of a CGFE module, an FGFE module, two proposed 2O-IB layers, some convolutional layers, and pooling layers. Both original Re-ID datasets (e.g., Market1501 [39], DukeMTMC-reID [42], and CUHK03 [11]) and an extra-domain dataset of other object classes (e.g., cat, fish, bird, and car images in ImageNet) are prepared

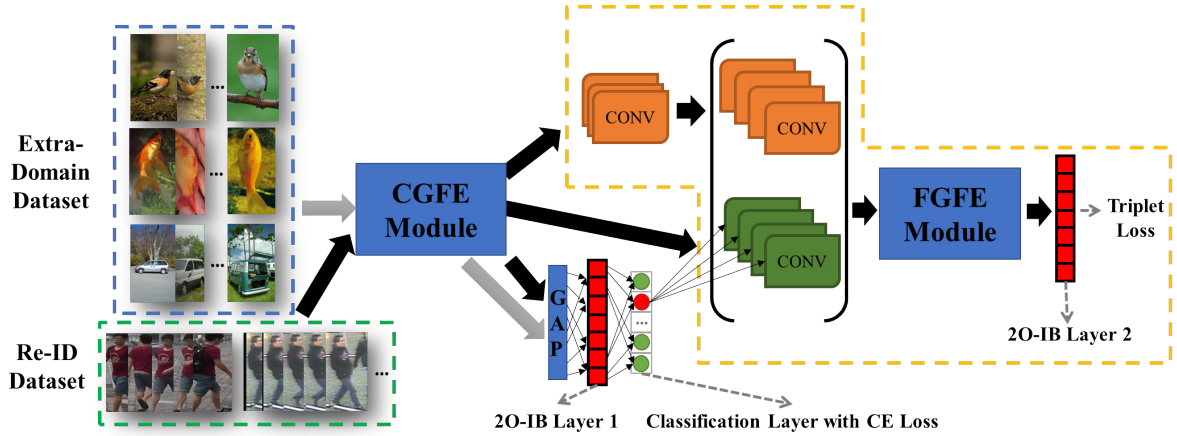

Figure 3. The proposed end-to-end hierarchical coarse-to-fine framework for person Re-ID. Both the original Re-ID dataset and the extra-domain dataset are used for training. *Coarse-grained feature extraction* (CGFE) module aims to extract features for image classification, as well as provide primary feature maps for *fine-grained feature extraction* (FGFE) module that extracts features for calculating the triplet loss. *GAP* and *CONV* denote Global Average Pooling and convolution operations, respectively. $(\cdot)$ of the *CONV* blocks denotes the concatenation operation. Only when the input samples are person images for Re-ID (meeting the input requirements of the triplet loss), the components in the yellow dashed box will be trained. The red dot in the classification layer represents the person class, while the three green dots represent the other object classes. The thin black arrows "$\rightarrow$" between *GAP*, *2O-IB Layer 1* and *Classification Layer with Cross-Entropy (CE) Loss* denote the neuronal connections. The thick black arrows indicate the data flow for images come from the Re-ID dataset, while the thick gray arrows indicate the data flow for images come from the extra-domain dataset. And the thin arrows with the gray dotted lines "$\dashrightarrow$" are used to point to the corresponding description texts.

for training. For the model training process, we propose an ADC task and a 2O-IB method. The ADC enables the CGFE module in the framework to learn from the joint datasets how to recognize people and what features can effectively describe a general person. Therefore, this coarse step helps the model filter out background noise and redundant information from the original images.

On this basis, the FGFE module learns how to distinguish different persons for the Re-ID task. The 2O-IB method we proposed can further improve the stability of network representation, overcome the uncertainty caused by non-essential variations of images, and make the model more robust and generalizable. The proposed hierarchical framework is trained end-to-end using multi-task learning of the ADC task with a cross-entropy loss and the Re-ID task with a triplet loss.

3.1. Auxiliary-Domain Classification

Conventional deep learning based person Re-ID models use initialized or pre-trained weights to start training on specialized Re-ID datasets. However, all current Re-ID datasets often have thousands of person categories, while each category has only a few dozens of images. For example, the training set of Market1501 has 751 categories, and each category has only 16 people images on average; MSMT17 contains 4101 person categories, but only 30 images per category on average. Therefore, the generalization ability of the model trained based on these original datasets is usually poor.

During the training process of the proposed ADC, the samples of each batch are randomly selected from the Re-ID dataset or the extra-domain dataset. In particular, if the samples of a batch come from the Re-ID dataset, the selection should meet the sample requirements for the triple loss training. All samples first go through the CGFE module to extract the coarse-grained feature maps, then go through the global average pooling (GAP) layer and the second-order information bottleneck layer (which will be introduced in Subsection 3.2), and finally go through the classification layer to calculate the classification loss (implemented by the cross-entropy loss function). If these samples come from the extra-domain dataset, then the samples of this batch have been trained.

Otherwise, if the samples come from the Re-ID dataset, the module within the yellow dashed box in Figure 3 continues to be trained. Specifically, the concatenation of two groups of feature maps are used as the input for the FGFE module. Finally, a 2O-IB layer and a fully connected layer are applied to the output of the FGFE module to map an image into an embedding vector. In this paper, the triplet loss with hard positive/negative mining [6] is used to train the network to distinguish people similarity. Triplet loss renders a cubical possible triplet number of training images. Assume a batch is formed by randomly sampling P people classes and randomly sampling K images for each class. For each sample a in the batch, we first select the hardest positive and the hardest negative samples within the batch to form a triplet. Then, the triplet loss with hard posi-

tive/negative mining is formulated as

$$\mathcal{L}_{triplet}(\theta, \Psi_{ReID}) = \sum_{i=1}^{P}\sum_{a=1}^{K}\left[m + \max_{p=1,\ldots,K} D\big(f_\theta(\psi_a^i), f_\theta(\psi_p^i)\big) - \min_{\substack{j=1,\ldots,P \\ n=1,\ldots,K \\ j\neq i}} D\big(f_\theta(\psi_a^i), f_\theta(\psi_n^j)\big)\right]_+, \quad (3)$$

where P denotes the classes (person identities) that randomly sampled in a batch, K denotes the image number of each class (person). Ψ_{ReID} is the batch of PK Re-ID data samples, ψ_j^i corresponds to the j-th image of the i-th person in the batch, $f_\theta(\psi)$ denotes the output embedding vector by model hyperparameter set θ for sample ψ, D denotes the distance measurement between two vectors.

It is worth mentioning that ADC task and the two 2O-IB layers are only applied during the training stage, and therefore do not cause additional computational burden to the model's inference. Specifically manifested in two aspects: 1) Among the neuron nodes of the ADC task, only the node representing "person" is retained in the inference process, and the rest nodes and connection weights representing other domain category targets are discarded; 2) The two 2O-IB layers introduce additional calculations and storage during the training, but in the inference process, the corresponding connections and nodes can be completely replaced by common fully-connected layers, which greatly simplifies the calculation overhead.

3.2. Second-Order Information Bottleneck

We have demonstrated that the CGFE module aims to extract the feature of *"how to tell whether it is a person or not"*, while the FGFE module is to extract the feature of *"how to distinguish different persons"*. However, the significant changes in background, illumination, viewpoint, and human pose make the feature extraction modules difficult to overcome the intra-class variations of persons. Section 2.2 stated the advantages of IB, i.e., compressing the input observation as much as possible while keeping only the information most relevant to the target output, which helps reduce the stochasticity of network output caused by input changes.

In this paper, we propose a 2O-IB for training the deep learning framework:

$$\mathcal{L}_{IB}(T) = H^2(Y|T) + \beta I^2(X;T), \quad (4)$$

where $\beta > 0$ is a parameter to balance the trade-off between compression and prediction. X, Y, T denote the input, target output and bottleneck variable, respectively.

The conventional IB problem of minimizing $I(X;T)$ is a convex function of the encoding mapping $q(t|x)$ for fixed $p(x)$. Maximizing $H(Y|T) = H(Y) - I(T;Y)$ is a concave function of decoding mapping $q(y|t)$ for a fixed $p(x, y)$. The entropy $H(Y)$ only depends on the distribution of the variable (output) Y, and it can be regarded as a constant for the whole dataset $\mathcal{D}$, or with a relatively small fluctuation for a batch of training data. Thus, the 2O-IB optimization function $\mathcal{L}_{IB}$ is a concave function to determine the global or local minimum.

We define the optimal parameter set of the network to achieve the best performance (i.e., the minimal value of the loss function) to be ω. We let the predicted (decoded) representation by the hyperparameter set θ be $\hat{Y}$, and $\hat{y}$ is a random instance of $\hat{Y}$. Thus, we have the following:

$$\begin{aligned} H(q_\theta(Y|T)) &\leq H(q_\theta(Y|T)) + D_{KL}(q_\omega(Y|T) \parallel q_\theta(Y|T)) \\ &= -\mathbb{E}_{q_{\omega(Y,T)}}[\log q_\theta(Y|T)] \\ &\overset{(a)}{=} -\mathbb{E}_{q_{\omega(Y,\hat{Y})}}\left(\mathbb{E}_{q_{\theta(\hat{Y},T)}}\left[\log q_\theta(Y|\hat{Y})\right]\right) \\ &= -\mathbb{E}_{q_{\omega(Y,\hat{Y})}}\left[\log q_\theta(Y|\hat{Y})\right], \end{aligned} \quad (5)$$

where $\mathbb{E}[\cdot]$ denotes the expectation value, D_{KL} is the Kullback-Leibler (KL) divergence, and equation (a) in (5) is due to

$$q_{Y|T}(y|t) = q_{Y|q(\hat{y}|t)}(y|q(\hat{y}|t)) = q_{Y|\hat{Y}}(y|\hat{y}). \quad (6)$$

Remark 1: For auxiliary-domain classification problem (*2O-IB Layer 1* in Figure 3), (5) can be rewritten by

$$\begin{aligned} H(q_\theta(Y|T)) &\leq \sum_q \left[q_{Y|\hat{Y}}(y|\hat{y}) \log q_\theta(y|\hat{y})\right] \\ &= \mathcal{C}\left(q_\theta(Y|\hat{Y})\right), \end{aligned} \quad (7)$$

where $\mathcal{C}(\cdot)$ is the well-known cross entropy loss function and y denotes the target class label for an input sample.

The inequality (5) becomes an equality only if $q_\theta(y|t)$ equals to the "optimal" mapping $q_\omega(y|t)$. Further, it holds that $H(q_\theta(Y|T)) \to \mathcal{C}\left(q_\theta(Y|\hat{Y})\right)$ if the KL divergence $D_{KL}(q_\omega(Y|T) \parallel q_\theta(Y|T)) \to 0$, which implies that if the distance between hyperparameter set θ and the "optimal" set ω is trained to be smaller, the term $H(q_\theta(Y|T))$ is closer to its upper bound.

The term $I(X;T)$ represents the information compressed from input signal X to the intermediate activation T:

$$\begin{aligned} I(X;T) &= \sum_{x,t} q(x,t) \log\left(\frac{q(x,t)}{p(x)\,q(t)}\right) \\ &= \sum_{x,t} q(x,t) \log\left(\frac{q(t|x)}{q(t)}\right) \\ &= \sum_{x,t} q(x,t) \log q(t|x) - \sum_t q(t) \log q(t). \end{aligned} \quad (8)$$

However, computing the marginal distribution of T, $q(t) = \sum_x q(t|x)p(x)$ might be difficult. Inspired by the variational IB [1], we also use the variational distribution $r(t)$ to approximate $q(t)$. Because the KL divergence is defined to be nonnegative, thus we have the following:

$$D_{KL}\left(q(T) \parallel r(T)\right) = \sum_{t} q(t)\log q(t) - \sum_{t} q(t)\log r(t) \geq 0. \quad (9)$$

According to (8) and (9), we have

$$\begin{aligned} I(X;T) &\leq \sum_{x,t} q(x,t)\log q(t|x) - \sum_{t} q(t)\log r(t) \\ &= \sum_{x,t} p(x)q(t|x)\log q(t|x) - \sum_{x,t} p(x)q(t|x)\log r(t) \\ &= \frac{1}{N}\sum_{x_i\in\Psi_{all}} q(t|x_i)\log\frac{q(t|x_i)}{r(t)} \\ &= \frac{1}{N}\sum_{x_i\in\Psi_{all}} D_{KL}\Big[q\left(T|x_i\right) \parallel r\left(T\right)\Big], \end{aligned} \quad (10)$$

where Ψ_{all} is a batch of N data samples.

Combining (5) and (10) with the fact that $H(Y|T) \geq 0$ and $I(X;T) \geq 0$, we quantify the upper bound $\bar{\mathcal{L}}_{ADC}$ of the second-order IB for ADC task as

$$\begin{aligned} \mathcal{L}_{IB-ADC} \leq \bar{\mathcal{L}}_{IB-ADC} &= \left[\mathcal{C}\left(q_\theta(Y|\hat{Y})\right)\right]^2 \\ &+ \beta\left[\frac{1}{N}\sum_{x_i\in\Psi_{all}} D_{KL}\big[q\left(T|x_i\right) \parallel r\left(T\right)\big]\right]^2. \end{aligned} \quad (11)$$

We can convert the problem of minimizing the loss function $\mathcal{L}_{IB-ADC}$ into minimizing the upper bound $\bar{\mathcal{L}}_{IB-ADC}$, that is, reducing $\bar{\mathcal{L}}_{IB-ADC}$ to achieve the purpose of reducing $\mathcal{L}_{IB-ADC}$.

Remark 2: For the triplet loss with hard positive/negative mining (3) optimization problem, if we use Euclidean distance to measure the difference D between network output vectors, as noted in [9] (pp. 132-134) that the mean squared error can be interpreted as a cross-entropy term, thus similar to (11), we have the 2O-IB based (*2O-IB Layer 2* in Figure 3) triplet loss version as:

$$\begin{aligned} \mathcal{L}_{IB-triplet} \leq \bar{\mathcal{L}}_{IB-triplet} &= \mathcal{L}^2_{triplet} \\ &+ \beta\left[\frac{1}{PK}\sum_{x_i\in\Psi_{ReID}} D_{KL}\big[q\left(T|x_i\right) \parallel r\left(T\right)\big]\right]^2, \end{aligned} \quad (12)$$

where P, K, and Ψ_{ReID} have the same meaning as in (3). We can also minimizing the upper bound $\bar{\mathcal{L}}_{IB-triplet}$ to reduce the loss $\mathcal{L}_{IB-triplet}$.

Thus, the total loss $\mathcal{L}$ of our proposed framework is then defined by

$$\mathcal{L} = \sum_{\Psi_i\in\Phi} \mathcal{L}_{IB-ADC}(\Psi_i) + \lambda\cdot\sum_{\Psi_i\in\Phi_r} \mathcal{L}_{IB-triplet}(\Psi_i), \quad (13)$$

where Ψ_i is a batch of training images, Φ is the union of the extra-domain dataset and the Re-ID dataset, Φ_r is the Re-ID dataset, $\mathcal{L}_{IB-ADC}(\Psi_i)$ and $\mathcal{L}_{IB-triplet}(\Psi_i)$ are the cross entropy loss and triplet loss for the batch Ψ_i, respectively, and λ is a coefficient to adjust the contribution of the triplet loss.

4. Experiments

4.1. Datasets and Evaluation Metrics

Experiments are conducted on three commonly-used large-scale person Re-ID datasets, i.e., Market1501 [39], DukeMTMC-reID [42], and CUHK03 [11]. The extra-domain dataset is constructed by selecting images from 100 categories of the ImageNet dataset, such as cat, dog, plane, etc. Both the extra-domain dataset and the Re-ID dataset (person category) serve as the training dataset for the 101 categories classification task. Performance is evaluated by using the most popular metrics, including the cumulative match characteristic (CMC), which is reported via the rank-1, rank-5, and rank-10 accuracy values, and the mean average precision (mAP).

Market-1501 dataset was collected at Tsinghua University and made public in 2015. The training set contains 12,936 images from 751 persons, and the test set contains 19,732 images from 750 other persons, where each pedestrian is captured by at least two cameras.

DukeMTMC-reID dataset was collected at Duke University. It is a multi-target multi-camera person tracking dataset. The dataset was captured by eight cameras, where the training set contains 16,522 images from 792 persons, and the test set contains 19,889 images from 702 other persons and 408 distractors.

CUHK03 dataset was collected by five cameras at the Chinese University of Hong Kong. It contains 14,096 images from 1,467 persons. In this paper, the training set includes the images of 767 identities, and the test set includes the images of the other 700 identities. CUHK03 provides two types of annotations, including person bounding boxes labeled manually and detected automatically by Deformable-Part-Model (DPM) [3] detector, named CUHK03 labeled and CUHK03 detected, respectively.

4.2. Implementation Details

We build our Re-ID model based on ResNet50, as shown in Table 1, where "conv1, conv2_x, conv3_x and conv4_x, conv5_x, avg_pool" are the original ResNet50 components. The CGFE module is composed of "conv1, conv2_x, conv3_x, and conv4_x". The FGFE module is composed of "conv5_x and avg_pool". Between the CGFE and FGFE modules, there are a GAP layer, a 2O-IB layer, a classification layer for the ADC task, and some convolutional layers. After the FGFE module, there are a 2O-IB layer and a fully connected layer for the Re-ID task.

We select 100 categories of images from the ImageNet dataset as the extra-domain dataset. In order to ensure that the aspect ratios of the images in the extra-domain dataset are close to those in the Re-ID dataset, we select images with an aspect ratio between 2:3 and 2:5 from these categories. All the images are resized to 256×128, and

Table 1. Detailed model architecture of the proposed neural network framework for Re-ID.

Module	Layer name	Output size	Operation
	input	256×128	
Coarse grained feature extraction (CGFE)	conv1	128×64	7×7, 64, stride 2
	conv2_x	64×32	3×3, maxpooling, stride 2
			$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$
	conv3_x	32×16	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 4$
	conv4_x	16×8	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 6$
			global average pooling (GAP) → CONV: 3×3,512
			2O-IB layer 1 → CONV: 3×3,512
			classification layer
	concat		feature maps concatenation
Fine grained feature extraction (FGFE)	conv5_x	8×4	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 3$
	avg_pool	256×1	average pooling, 256-d fc
	IB2	256×1	2O-IB layer 2
	output embedding	512×1	512-d fc

Table 2. Comparison with state-of-the-art methods on **Market1501** dataset.

Method	Rank-1	Rank-5	Rank-10	mAP
HCT [28]	80.0	91.6	**95.2**	56.4
MGCAM [15]	83.6	-	-	74.3
MGCAM-Siamese[15]	83.8	-	-	74.3
VCFL [29]	89.3	95.6	-	74.5
FGSAM [44]	91.5	**96.8**	**97.3**	85.4
VCFL+ [29]	91.9	**96.7**	-	**90.8**
IANet [7]	94.4	-	-	83.1
SNR [10]	94.4	-	-	84.7
DSA [33]	**95.7**	-	-	87.6
RGA-SC [34]	**96.1**	-	-	**88.4**
PCB+RPP($\mathcal{G}$)[20]	93.8	-	-	81.6
Ours	**94.8**	**97.2**	**98.0**	**87.7**

augmented by random flipping, hue shifting, cropping, and small-angle rotation. For the triplet loss, we sample $P = 16$ identities and $K = 4$ images for each identity as a batch and the margin parameter m in (3) is set to 0.3.

4.3. Performance Evaluation

We compare our proposed framework for person Re-ID with recent state-of-the-art methods on Market1501, DukeMTMC-Re-ID, CUHK03 labeled, and CUHK03 detected in Tables 2, 3, 4, and 5, respectively. In each Table, a symbol "-" indicates that the corresponding value is not provided in the corresponding paper. The top three performance values in each column are highlighted in red, blue, and green colors, respectively.

As shown in Tables 2, 3, 4, and 5, our proposed method

Table 3. Comparison with state-of-the-art methods on **DukeMTMC-reID** dataset.

Method	Rank-1	Rank-5	Rank-10	mAP
HCT [28]	69.6	83.4	**87.4**	50.7
VCFL [29]	80.3	**90.2**	-	63.4
VCFL+ [29]	82.7	**91.3**	-	65.7
PCB+RPP($\mathcal{G}$)[20]	83.3	-	-	69.2
SNR [10]	84.4	-	-	72.9
FGSAM [44]	85.9	**90.2**	**93.4**	**74.1**
DSA [33]	**86.2**	-	-	**74.3**
IANet [7]	**87.1**	-	-	73.4
Ours	**87.4**	**92.1**	**95.5**	**74.9**

Table 4. Comparison with state-of-the-art methods on **CUHK03 labeled** dataset.

Method	Rank-1	Rank-5	Rank-10	mAP
NFST [30]	54.7	84.8	94.8	-
PDC [16]	78.3	**94.8**	**97.2**	-
DSA [33]	78.9	-	-	**75.2**
RGA-SC [34]	**81.1**	-	-	**77.4**
PAR [38]	**81.6**	**97.3**	**98.4**	-
Ours	**80.6**	**94.1**	**98.2**	**79.3**

Table 5. Comparison with state-of-the-art methods on **CUHK03 detected** dataset.

Method	Rank-1	Rank-5	Rank-10	mAP
NFST [30]	69.1	86.9	**91.8**	-
VCFL+ [29]	69.9	81.4	-	70.5
VCFL [29]	70.4	81.1	-	70.4
PAR [38]	75.0	**93.5**	**97.1**	-
DSA [33]	78.2	-	-	**73.1**
RGA-SC [34]	**79.6**	-	-	**74.5**
Spindle [36]	**79.9**	**94.4**	**97.1**	-
Ours	**81.3**	**94.1**	**97.2**	**84.1**

achieves superior or comparable performance on all these datasets. Specifically, the proposed method achieves 10 out of 16 best performance values across all four datasets, including 2 out of 4 best rank-1 values, 2 out of 4 best rank-5 values, 3 out of 4 best rank-10 values, and 3 out of 4 best mAP values. Especially on the DukeMTMC-reID dataset, as shown in Table 3, the proposed method achieves all best performance values in terms of rank-1, rank-5, rank-10, and mAP.

4.4. Ablation Study

We further perform ablation study to validate the effectiveness of the proposed ADC task and the 2O-IB method. In the experiments, "2O-IB L2 only" means that extra-domain dataset based ADC and 2O-IB Layer 1 are removed, and we only use 2O-IB Layer 2 for triplet loss in Figure 3. "ADC only" means that we do not use 2O-IB layers for feature compression, but replaced the 2O-IB layers with the fully-connected layers for both training and inference.

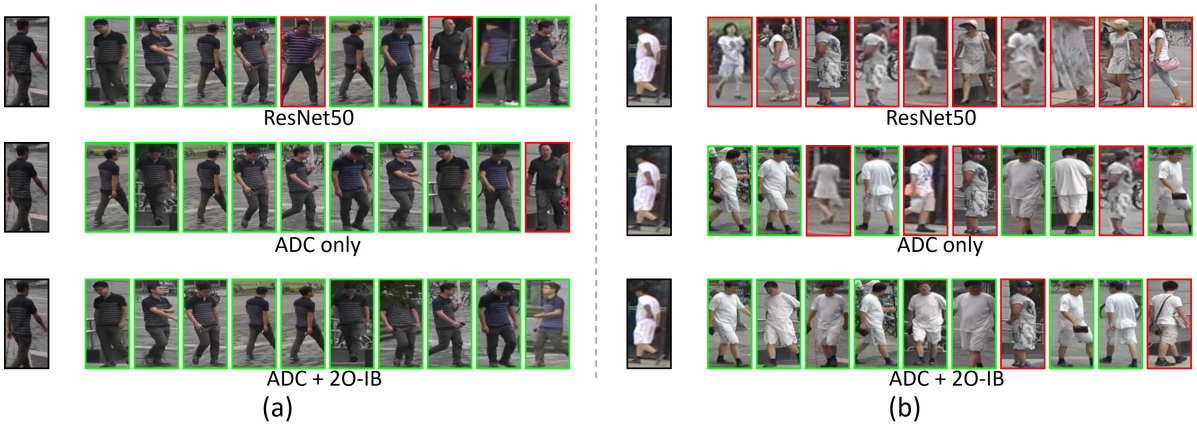

Figure 4. Visual results for comparing the person images retrieved by original ResNet50, ADC only and Ours (ADC+2O-IB), where ADC only means the proposed 2O-IB is not used to compress feature information. In each subfigure, the left image in a single column shows the query image, and the right images in ten columns are the retrieved results in retrieved order.

Table 6. Effectiveness of the proposed ADC and second-order IB on the **Market1501** dataset.

Method	Rank-1	Rank-5	Rank-10	mAP
Baseline	88.6	92.9	95.5	74.3
2O-IB L2 only	90.1	93.0	94.9	82.0
ADC only	91.3	94.8	96.0	84.2
ADC+2O-IB L1	**92.6**	**95.6**	**97.2**	**85.5**
ADC+2O-IB L2	**94.1**	**96.0**	**96.5**	**87.1**
ADC+2O-IB	**94.8**	**97.2**	**98.0**	**87.7**

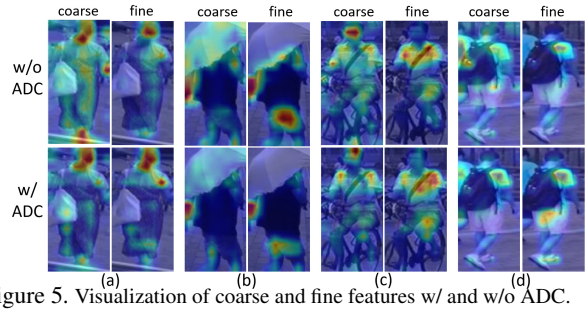

Figure 5. Visualization of coarse and fine features w/ and w/o ADC.

"ADC+2O-IB L1" and "ADC+2O-IB L2" means we use ADC and only one of the 2O-IB layers.

In Table 6, we observe that using the ADC only or using each 2O-IB layer only also significantly outperforms the Baseline, but is inferior to the proposed method that uses both ADC and 2O-IB. It indicates that the main improvements come from both the proposed ADC and 2O-IB.

In Figure 4, we show two examples to illustrate the effectiveness of the proposed ADC and 2O-IB on the person recognition ability. Figure 4(a) shows an easy re-identification example, while Figure 4(b) shows a relatively difficult re-identification example. It can be seen that the number of incorrect images retrieved by the original ResNet50 baseline is larger than that by the proposed method ("ADC+2O-IB"). Furthermore, among the retrieved images of "ADC only" and "ADC+2O-IB" we proposed, the accuracy of the first few images (especially for rank-5) retrieved by the proposed method ("ADC+2O-IB") is higher than that by "ADC only". It implies that, compared with the original ResNet50, both the ADC and 2O-IB we proposed can improve the model representation ability, thereby retrieving more correct person images. Especially when ADC and 2O-IB are used together, the computational accuracy of the model is greatly improved. We also show some visualization examples of the coarse and fine features w/ and w/o ADC in Figure 5.

5. Conclusion

In this paper, we proposed a hierarchical coarse-to-fine person Re-ID framework based on auxiliary-domain classification (ADC) and second-order information bottleneck (2O-IB), where ADC is used to enable the network to achieve step-by-step person representation. For a person image, the proposed framework first extracts the features that can answer "What is a person? " through the coarse-grained feature extraction (CGFE) module and the ADC task, and then extracts the features that can answer "How to distinguish between different persons? " through the fine-grained feature extraction (FGFE) module and the Re-ID task. The proposed 2O-IB is used to compress the redundancy and noise in the input person image, while ensuring the maximum correlation between the compressed information and the expected output as much as possible. The experimental results on the Market1501, DukeMTMC-Re-ID, and CUHK03 datasets show that our proposed method outperforms other state-of-the-art Re-ID methods.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under grant No. U1505251, No. 61672158, No. 61702104, No. 61972097, and No. 62072110, Chinese Ministry of Science and Technology under grant No. 2016YFE0122700, National Science Foundation of Fujian Province under grant No. 2019J02006, No. 2020J01494, and No. 2018J07005, and Fujian Provincial Department of Science and Technology under grant No. 2018I0011.

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. In *International Conference on Representation Learning*, pages 1–19, 2017. 5

[2] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *IEEE International Conference on Computer Vision*, pages 371–381, 2019. 2

[3] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 6

[4] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11741–11749, 2020. 3

[5] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *IEEE/CVF International Conference on Computer Vision*, pages 8449–8458, 2019. 3

[6] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *ArXiv*, abs/1703.07737, 2017. 4

[7] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-And-Aggregation Network for Person Re-Identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9309–9318, 2019. 7

[8] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018. 2, 3

[9] Goodfellow Ian, Bengio Yoshua, and Courville Aaron. *Deep Learning*. Cambridge, MA: MIT Press. 6

[10] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 7

[11] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014. 3, 6

[12] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference on Artificial Intelligence*, pages 2194–2200, 2017. 1

[13] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious Attention Network for Person Re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018. 2

[14] Jinxian Liu, Bingbing Ni, Yichao Yan, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018. 2

[15] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-Guided Contrastive Attention Model for Person Re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. 1, 2, 7

[16] Chi Su, Li Jianing, Shiliang Zhang, Xing Junliang, Gao Wen, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *IEEE International Conference on Computer Vision*, pages 3980–3989, 2017. 7

[17] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-Driven Deep Convolutional Model for Person Re-identification. In *IEEE International Conference on Computer Vision*, pages 3980–3989, 2017. 1

[18] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, 2016. 1

[19] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition*, page S0031320317302686, 2017. 1

[20] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European Conference on Computer Vision*, 2018. 7

[21] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. 2000. 2, 3

[22] Nicolai Wojke and Alex Bewley. Deep Cosine Metric Learning for Person Re-Identification. In *IEEE Winter Conference on Applications of Computer Vision*, pages 748–756, 2018. 1

[23] Qiqi Xiao, Hao Luo, and Chi Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *ArXiv*, abs/1710.00478, 2017. 1

[24] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-Aware Compositional Network for Person Re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018. 2

[25] Hong Xing Yu, Ancong Wu, and Wei Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE International Conference on Computer Vision*, 2017. 1

[26] Hong Xing Yu, Wei Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian Huang Lai. Unsupervised person Re-Identification by Soft Multilabel Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 2143–2152, 2019. 1

[27] Tianyuan Yu, Da Li, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *IEEE/CVF International Conference on Computer Vision*, pages 552–561, 2019. 3

[28] Kaiwei Zeng, Munan Yang, Yaohua Wang, and Yang Guo Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 7

[29] Lei Zhang, Fangyi Liu, and David Zhang. Adversarial view confusion feature learning for person re-identification. *IEEE*

Transactions on Circuits and Systems for Video Technology, pages 1–1, 2020. 7

[30] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016. 7

[31] Shun Zhang, Yantao He, Jiang Wei, Shaohui Mei, and Ke Chen. Person re-identification with joint verification and identification of identity-attribute labels. *IEEE Access*, (99):1–1, 2019. 1

[32] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. *Arxiv*, abs/1711.08184, 2017. 1

[33] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019. 1, 7

[34] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 7

[35] Ziyue Zhang, Richard Yi Da Xu, Shuai Jiang, Yang Li, Congzhentao Huang, and Chen Deng. Illumination adaptive person reid based on teacher-student model and adversarial training. In *IEEE International Conference on Image Processing*, pages 2321–2325, 2020. 2

[36] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Yi Shuai, Wang Xiaogang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 907–915, 2017. 7

[37] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-Learned Part-Aligned Representations for Person Re-identification. In *IEEE International Conference on Computer Vision*, pages 3239–3248, 2017. 1

[38] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *IEEE International Conference on Computer Vision*, pages 3219–3228, 2017. 7

[39] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015. 3, 6

[40] Wang Zheng, Bai Xiang, Mang Ye, and Shin'Ichi Satoh. Incremental deep hidden attribute learning. In *ACM Multimedia Conference*, 2018. 1

[41] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint Discriminative and Generative Learning for Person Re-Identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2133–2142, 2019. 2

[42] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision*, 2017. 3, 6

[43] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person re-identification. *Acm Transactions on Multimedia Computing Communications & Applications*, 14(1), 2018. 1

[44] Qinqin Zhou, Bineng Zhong, Xiangyuan Lan, Gan Sun, Yulun Zhang, Baochang Zhang, and Rongrong Ji. Fine-Grained Spatial Alignment Model for Person Re-Identification With Focal Triplet Loss. *IEEE Transactions on Image Processing*, 29:7578–7589, 2020. 7