QuadNet: Quadruplet Loss for Multi-view Learning in Baggage Re-identification

Hao Yang[a,*], Xiuxiu Chu[a,*], Li Zhang[a,b], Yunda Sun[a], Dong Li[a,b], Stephen J. Maybank[c]

[a] *R&D Center of Artificial Intelligence, NUCTECH Company Limited, Beijing, China*
[b] *Department of Engineering Physics, Tsinghua University, Beijing, China*
[c] *Department of Computer Science and Information Systems, Birkbeck College, London, United Kingdom*

Abstract

Recently, baggage re-identification (ReID) has become an attractive topic in computer vision because it plays an important role in intelligent surveillance. However, the wide variations in different views of baggage items degrade baggage ReID performance. In this paper, a novel QuadNet is proposed to solve the multi-view problem in baggage ReID at three levels. At the sample level, we propose a multi-view sampling strategy which samples hard examples from multiple identities in multiple views. The sampled baggage items are used to construct quadruplets. At the feature level, view-aware attentional local features are extracted from discriminative regions in each view. These local features are fused with global features to obtain better representations of the quadruplets. At the loss level, a multi-view quadruplet loss operating on the representations of quadruplets is proposed to reduce the intra-class distances caused by view variations and increase the inter-class distances of baggage images captured in the same view. A random local blur data augmentation is proposed to handle the motion blur which is often found in baggage images. The multi-task learning of materials is introduced to obtain discriminative features based on the

[*] Equal Contribution

Email addresses: `yanghao1@nuctech.com` (Hao Yang), `chuxiuxiu@nuctech.com` (Xiuxiu Chu), `zli@mail.tsinghua.edu.cn` (Li Zhang), `sunyunda@nuctech.com` (Yunda Sun), `li.dong@nuctech.com` (Dong Li), `steve.maybank@bbk.ac.uk` (Stephen J. Maybank)

materials of baggage surfaces.

Extensive experiments on three ReID datasets, MVB, Market-1501 and VeRi-776, indicate the remarkable effectiveness and good generalization of the QuadNet model. It has achieved the state-of-the-art performance on the three datasets.

Keywords: Baggage re-identification, Multi-view learning, Quadruplet loss, View-aware features.

1. Introduction

In recent years, people have demanded much higher requirements for travel safety. Baggage re-identification (baggage ReID) is a core component of current intelligent baggage check devices which are used in airports, customs and other places with a high demand for safety. The aim of baggage ReID is to spot a baggage item of interest in images taken by different cameras in a surveillance system. In traditional airport security checks, Radio Frequency IDentification (RFID) tags are usually attached to baggage items so that they can be traced. However, this method has several limitations as follows: 1) In many cases the RFID tags fall off in transit or are deliberately torn off to avoid inspection; 2) Attaching the RFID tags to baggage items is labour intensive and time consuming; 3) Any metal in the baggage may interfere with the detection of RFID tags. Therefore the re-identification of baggage based on visual appearance has attracted more attention recently [1] following the fast development of new technologies in intelligent surveillance.

In ReID tasks, person ReID [2, 3, 4] and vehicle ReID [5, 6, 7] have been well studied in recent decades. However, there is a lack of research on baggage ReID because more challenges such as occlusion, shape change, motion blur and view variation, are encountered, compared with the person ReID and vehicle ReID tasks. In particular, view variation is a major problem. On one hand, different baggage items might have very similar appearances in the same view, as shown in Fig. 1(a). On the other hand, a baggage item may have large appearance

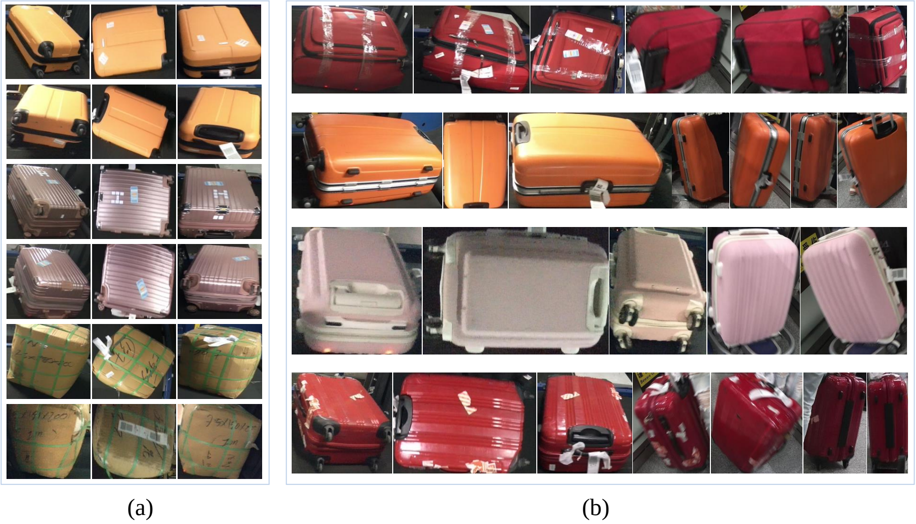

Figure 1: Several baggage examples from the MVB dataset. Each row corresponds to a single baggage item. (a) shows appearance similarities for different baggage items in the same view. (b) shows large appearance variations of the same baggage item in different views.

variations in different views, as shown in Fig. 1(b). Therefore, it is a significant challenge to obtain a baggage ReID model which is reliable given the inter-class similarities in a single view and the large intra-class variations in different views of the same baggage item.

In the literature, to tackle the intra-class view variation, some ReID works [5, 6, 7] fuse the multi-view information that is learned with extra annotations, such as key points [5] or views [6, 7]. However, the cost of these annotations is exorbitant. Several loss functions have been proposed to tackle inter-class similarities in ReID. Examples are the contrastive loss [8], triplet loss [9, 3], quadruplet loss [10] and classification loss [2, 11, 12]. However, these loss functions do not consider the information in a range of views, and so are not sufficient to solve the multi-view problem in ReID.

To solve the multi-view problem in baggage ReID, we propose the QuadNet model, which includes a multi-view hard example sampling strategy, view-aware attentional local features and a novel quadruplet loss. The QuadNet model deals with the multi-view problem at three levels. Firstly, at the sample level, the multi-view hard example sampling strategy selects four examples from multiple identities and multiple views to construct a quadruplet. Secondly, at the feature level, we extract view-aware attentional local features from the discriminative regions in different views. The baggage items may have different discriminative features in different views, for example, the logo in the top view and the handle or wheels in the front view, as shown in Fig. 1. Typically, the features extracted from the different spatial regions in different views are treated equally. This introduces disturbance information into the ReID models. Then, these view-aware attentional features are fused with global features to obtain powerful representations of the quadruplets. Finally, at the loss level, a quadruplet loss is constructed on the representations of the quadruplets. Compared with other losses in ReID, the quadruplet loss optimizes the model in two aspects: 1) It reduces the distances between representations of the same baggage item in different views, in order to ensure that these distances are smaller than the distances between representations of different baggage items with the same view; 2) It increases the distances between representations of different baggage items in different views to ensure that they are larger than the distances between representations of the same baggage item in different views.

Baggage ReID is badly affected by motion blur which is often found in baggage images. In order to improve the robustness of the QuadNet model against motion blur, we propose a data augmentation method, namely random local blur, for pre-processing. The multi-task learning of materials is proposed to distinguish baggage items with similar appearances but made from different materials. This method strengthens discrimination of the QuadNet model in real applications.

The contributions of this paper can be summarized as follows:

- A novel quadruplet loss is proposed to solve the multi-view problem in baggage ReID. The quadruplet loss equipped with a multi-view sampling strategy effectively reduces the intra-class distances and increases the inter-class distances. To our knowledge, this is the first work that proposes the quadruplet loss for multi-view learning.

- The view-aware attentional local features are learned from the discriminative regions in different views. The local features are fused with global features to enhance the representations of baggage images.
- Random local blur is proposed to handle motion blur which is usually found in baggage images. The multi-task learning of materials is used to improve the discrimination of the QuadNet model.
- The proposed QuadNet model is evaluated extensively on the MVB dataset to demonstrate the effectiveness of QuadNet for baggage ReID. To be compared more fairly, the generalization of QuadNet is evaluated on the Market-1501 and VeRi-776 datasets. The QuadNet model achieves the state-of-the-art performance on all three datasets.

2. Related works

In recent decades, ReID has been widely studied in computer vision, *i.e.*, person ReID and vehicle ReID. Traditional ReID methods mainly concentrate on two aspects: discriminative feature extraction [13, 14, 15] and similarity measurement [16, 17]. The features and the metrics for similarities are studied independently. In recent years, deep learning has become the mainstream method in many computer vision tasks, such as, object recognition [18, 19], object detection [20, 21], object re-identification [1, 4, 22], *etc.* These methods have achieved better generalization, compared with the handcrafted features based methods [23, 24, 25]. Deep models for ReID are optimized end-to-end by error back-propagation. The features and metrics are learned jointly. For example, the early works [2, 12, 26] treat the ReID problem as a general classification task. A cross-entropy loss is adopted to train deep networks to enlarge the distances between representations with different identities. However, these methods may not be enough to distinguish the identities that are not in the training set. In addition, the number of parameters increases as the number of identities increases. Other works, *e.g.* [8], introduce the contrastive loss in person ReID to enlarge the margin between positive and negative pairs. Additionally, the works [9, 3, 10] treat the ReID problem as a ranking task and exploit a triplet loss to train their networks. TriNet [3] proposes a variant of triplet loss for the online mining of hard negative examples without additional cost. ImpTrpLoss [9] introduces an additional constraint to ensure that the distances between positive pairs are less than a predefined value. In [10], a combination of two triplet losses is constructed on the absolute distances between the positive and negative samples. These loss functions still suffer from weak generalization because the view information is not considered in sufficient detail.

In order to deal with the variations in appearance of the same item in different camera views, several methods [5, 6, 7] employ deep neural networks to extract view-invariant features. The Viewpoint-aware Attentive Multi-view Inference model (VAMI) [6] transforms single-view features into global multi-view features via an adversarial training architecture. The method [7] exploits the Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to learn transformations across different views of vehicles by inferring all views information from the only one input view. The orientation invariant feature embedding [5] is proposed by utilizing the annotated key points of vehicles. In addition, some methods [27, 28, 29] make use of the information from human pose. SpindleNet [27] facilitates feature learning using human structure information. The method [28] proposes the pose invariant embedding (PIE) as the representations of pedestrians. The above methods require large scale image annotations with key points or views. The annotations are labour intensive and costly in applications. In contrast, our method overcomes the intra-class variations in baggage ReID without extra annotations.

Attention is an important mechanism of human vision. It focuses selectively on salient regions in complex scenes. Attention is a popular topic in recent computer vision research, such as image classification [18], object detection [21] and person ReID [30, 31, 32]. In person ReID, HydraPlus-Net [31] captures

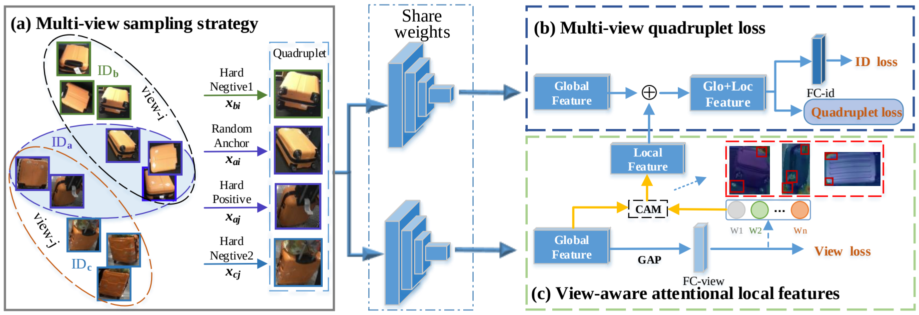

Figure 2: The proposed QuadNet model for multi-view learning in baggage ReID, (a) multi-view sampling strategy which samples hard examples from multiple identities and multiple views to construct quadruplets, (b) view-aware attentional local features are extracted from discriminative regions in different views, (c) multi-view quadruplet loss which minimizes intra-class variations in multiple views and maximizes inter-class similarities in the same view.

multi-scale selectiveness of attentive features from low-level to semantic-level to enrich the representations for a pedestrian image. The Multi-Scale Context Aware Network (MSCAN) [30] is designed to locate latent regions and the combination of full-body and body-part features is used for person ReID. However, these methods lack an explicit mechanism to ensure attention consistency. Class Activation Mapping (CAM) [21] learns a class-aware importance for each spatial position of the image by back-projecting the weights of the output layer to the convolutional feature maps. Meanwhile, CAM is trained with weak supervision and locates objects in a single forward pass. In this paper, the CAM is extended to extract view-aware local features from different discriminative regions in each view.

3. The proposed model

In this section, we introduce the QuadNet model for multi-view learning in baggage ReID. The QuadNet model consists of three parts: the multi-view hard example sampling strategy, the view-aware attentional local features, and the multi-view quadruplet loss. The overall architecture of the model is illustrated in Fig. 2.

3.1. The multi-view hard example sampling and quadruplet loss

To make the logic clearer, we first introduce the multi-view hard example sampling strategy and the multi-view quadruplet loss.

3.1.1. The multi-view hard example sampling

In ReID tasks, the most widely studied triplet loss based methods [9, 3, 4] usually randomly sample P identities and K images from each identity to form a training batch. However, these methods ignore the view information in images. In this paper, we propose a multi-view hard example sampling strategy to support multi-view learning in baggage ReID at the sample level. On the one hand, this sampling strategy samples baggage images from multiple views. On the other hand, this sampling strategy selects the hard samples from different identities or different views, which accelerates training of the QuadNet model.

Specifically, the QuadNet model samples $P \times K$ quadruplets to form a training batch denoted as B. Each quadruplet in B consists of four images $\{x_{ai}, x^*_{aj}, x^*_{bi}, x^*_{cj}\}$, where x_{ai} is sampled randomly from the training batch with the subscript a denoting the baggage identity and i denoting the label of view. We select the hardest positive sample x^*_{aj} having the same identity with x_{ai} but in a different view, *i.e.* $j \neq i$. The distance D_{ap} between x_{ai} and x^*_{aj} is:

$$D_{ap} = D(f(x_{ai}), f(x^*_{aj})), \tag{1}$$

$$x^*_{aj} = arg \max_{x_{aj} \in B_{a*}} [D(f(x_{ai}), f(x_{aj}))], \tag{2}$$

where $j \neq i$ and $j = 1, 2, ..., N$. N is the number of views. $B_{a*} \in B$ is a sample set having the same identity with x_{ai} in the training batch B. $f(x_{ai})$ and $f(x^*_{aj})$ denote the deep features of the sample x_{ai} and x^*_{aj}. These features are obtained from the last fully connected layer of the backbone network. Then, we select the hardest negative sample x^*_{bi} from view i and another hardest negative sample x^*_{cj} from view j:

$$x^*_{bi} = arg \min_{x_{bi} \in B_{*i}} [D(f(x_{ai}), f(x_{bi}))], \tag{3}$$

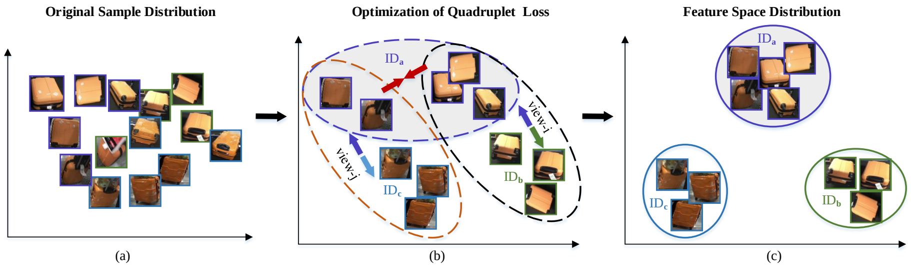
Original Sample Distribution
Optimization of Quadruplet Loss
Feature Space Distribution
IDa
view-j
view-i
IDc
IDb
IDa
IDc
IDb
(a)
(b)
(c)

Figure 3: The optimizing process of the proposed quadruplet loss, different colors of border denoting different identities of baggage images, (a) the distribution of the original sample space, (b) the optimization of quadruplet loss, (c) the distribution in feature space after optimizing with the quadruplet loss.

$$x^*_{cj} = arg \min_{x_{cj} \in B_{*j}} [D(f(x_{ai}), f(x_{cj}))], \tag{4}$$

where B_{*i} and B_{*j} are the sample sets of view i and j in the batch B. The distances of the two negative pairs are given respectively as:

$$D_{an_1} = D(f(x_{ai}), f(x^*_{bi})), \tag{5}$$

$$D_{an_2} = D(f(x_{ai}), f(x^*_{cj})). \tag{6}$$

3.1.2. The quadruplet loss for multi-view learning

To overcome the weakness of traditional methods in multi-view learning, we propose a multi-view quadruplet loss to deal with the inter-class similarities and the intra-class variations in baggage ReID. Based on a sampled quadruplet $\{x_{ai}, x^*_{aj}, x^*_{bi}, x^*_{cj}\}$ and the distances D_{ap}, D_{an_1}, and D_{an_2}, the quadruplet loss is given by:

$$L_{quad} = \alpha[D_{ap} - D_{an_1} + m_1]_+ + (1-\alpha)[D_{ap} - D_{an_2} + m_2]_+, \tag{7}$$

where $[u]_+ = max(u, 0)$. The first term in Eqn. 7 is optimized to pull the distance D_{ap} between x_{ai} and x^*_{aj} much closer than the distance D_{an_1} between x_{ai} and x^*_{bi} with a margin m_1. The second term in Eqn. 7 is utilized to push the distance D_{an_2} between x_{ai} and x^*_{cj} much farther than the distance D_{ap} with a margin m_2. The parameter α is used to balance the two optimization items. The margin m_2 is much larger than m_1 to ensure that the distance between a negative pair from different views is larger than the distance between a negative pair from the same view. In the experiments, we set $\alpha = 0.5$, $m_1 = 0.3$ and $m_2 = 1.2$ empirically.

The optimizing process of the quadruplet loss is shown in Fig. 3. In the original distribution, some distances between different baggage items with the same view are less than some distances of same baggage item, as shown in Fig. 3(a). After optimizing with the quadruplet loss, the distribution of baggage items in the feature space is shown in Fig. 3(c). The samples with the same identity are clustered compactly and the samples with different identities are pushed away. It indicates that the quadruplet loss optimizes the inter-class similarities and the intra-class variations effectively.

3.2. View-aware attentional local features

Learning discriminative features for different identities always plays the most important role in ReID. In baggage ReID, baggage images captured from different views have different discriminative spatial regions. We propose a view-aware attention mechanism to extract the discriminative local features from the discriminative spatial regions in each view. As shown in Fig. 2(b), the view-aware attentional local features are extracted by feeding the global features in the last convolutional layer into a Global Average Pooling (GAP) layer which is followed by a view classifier. The GAP layer retains the remarkable localization ability of the convolutional units until the final layer [21]. The cross entropy loss for view classification enables the network to focus on the most discriminative spatial regions in each view.

Formally, given a baggage image I, the convolutional features in the last convolutional layer of the backbone network are denoted as $F_{glob}(I) = f(I; W_b) \in R^{C \times H \times H}$, where C and H denote the number of channels and the spatial size of the feature maps respectively, and W_b denotes the weights of the backbone

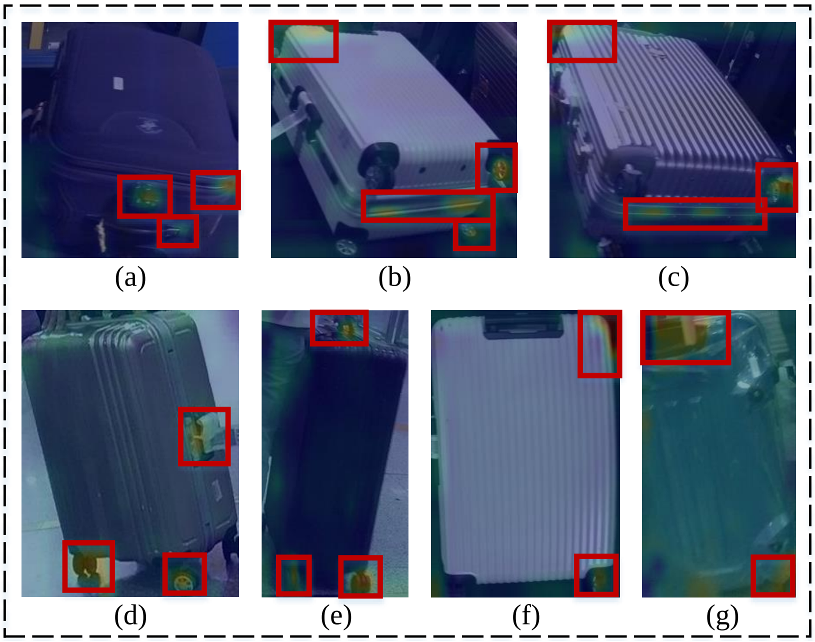

Figure 4: Seven visualization examples of the view-aware local features in the MVB dataset. Best viewed in color.

network. The global features $F_{glob}(I)$ are considered as stacked feature vectors from $H \times H$ spatial locations. The feature vector at a spatial location (x, y) is denoted as $F(x, y) \in R^C$. The features $F_{glob}(I)$ are fed into a GAP layer and a fully connected layer for view classification. For a view class v, the class score S_v is formulated as:

$$S_v = \sum_{c=1}^{C} \omega_v^c \sum_{x=1}^{H} \sum_{y=1}^{H} F(x, y) = \sum_{x=1}^{H} \sum_{y=1}^{H} \sum_{c=1}^{C} \omega_v^c F(x, y), \tag{8}$$

where ω_v^c denotes the weight of unit c corresponding to view class v, $v \in 1, 2, \ldots, V$ and V is the number of views. We define M_v as the view-aware local features for the view v. The activation of M_v at the spatial location (x, y) is given by:

$$M_v(x, y) = \sum_{c=1}^{C} \omega_v^c F(x, y), v = 1, 2, \ldots, V. \tag{9}$$

where $x, y = 1, 2, \ldots, H$ and $M_v(x, y)$ denotes the discrimination of the spatial location (x, y) in supporting the classification of view v. The final features used for baggage classification are constructed by concatenating the global features $F_{glob}(I) \in R^{C\times H\times H}$ and the view-aware attentional local features $\{M_v \in R^{H\times H}\}_{v=1}^{V}$ at the channel dimension. The final features are denoted as $F_{final}(I) \in R^{(C+V)\times H\times H}$.

Through a softmax layer, the scores of all view classes can be expressed as $p_{view} = [S'_1, S'_2, \ldots, S'_V]$, then the view classification loss is computed as:

$$L_{view} = \sum_{v=1}^{V} -y_{view}^{v} \log(S'_v), \tag{10}$$

where y_{view} is the view label of the baggage image, if $y_{view} = v$, y_{view}^{v} is set as 1, otherwise it is set as 0.

In Fig. 4, seven examples of the view-aware local features learned by QuadNet are presented. We normalize the view-aware local features M_v to $[0, 1]$ and use them as weights to weight each channel of baggage images. In examples (d), (e) and (f), the wheels and handles are highlighted. In (b), (c) and (g), our model selectively focuses on the zipper and trolley. These local features above are discriminative for baggage ReID, especially for different baggage items with similar appearances. These examples indicate that the QuadNet model is able to extract the view-aware local features from the discriminative spatial regions in different views.

3.3. Optimization

In baggage ReID, images are often degraded by motion blur. We propose a random local blur method to deal with the motion blur in baggage images. The random local blur data augmentation method introduces a local mean blur operation to simulate the motion blur in images during training. This improves the robustness of the baggage ReID model against motion blur. Given an image I, the probability of it undergoing a random blur operation is set as p_r, and the probability of remaining unchanged is $1 - p_r$. This method randomly selects a rectangular area in the image, denoted as $I_b = [x_b, y_b, x_b + W_b, y_b + H_b]$, where $x_b + W_b \leq W$ and $y_b + H_b \leq H$, (x_b, y_b) is the coordinate of the top left point of the rectangle I_b. W_b and H_b are the width and height of the rectangle I_b.

W and H are the width and height of the image I. Then a local mean blur operation is performed on I_b. Each spatial element of the image I is given by

$$I(x,y) = \begin{cases} \frac{1}{m \times n} \sum_{(s,t) \in S_{xy}} I_b(s,t) & (x,y) \in I_b \\ I(x,y) & (x,y) \notin I_b, \end{cases} \tag{11}$$

where S_{xy} is a local area centered at the position (x,y) with the rectangular size of $m \times n$. In the experiments, we set $m = 15$ and $n = 1$, as the motion blur usually occurs in the horizontal direction.

By analyzing the baggage images in the MVB dataset, we find that some baggage items with different identities may have the same appearance. Fortunately, the surface materials of these baggage items are often different, as shown in Fig. 5. To make full use of the annotation of materials in the MVB dataset, the multi-task learning of materials is introduced as an auxiliary task, which can distinguish baggage items with similar appearances but different materials. This task is optimized by the cross-entropy loss L_{ma}.

$$L_{ma} = \sum_{q=1}^{Q} -y_{ma}^{q} \log(p_q) \quad \begin{cases} y_{ma}^{q} = 1, y_{ma} = q \\ y_{ma}^{q} = 0, y_{ma} \neq q \end{cases} \tag{12}$$

where y_{ma} is the label of baggage material and p_q is the prediction probability of class q.

In order to further optimize the model, we combine the cross-entropy loss of baggage identities as in the previous works [12, 4]. It is denoted as L_{ID}. The final optimization objective is formulated as:

$$L = L_{quad} + \lambda_1 L_{ID} + \lambda_2 L_{ma} + \lambda_3 L_{view}, \tag{13}$$

where λ_1, λ_2, λ_3 are the weights of different losses. In our experiments, $\lambda_1 = 1$, $\lambda_2 = 0.3$ and $\lambda_3 = 0.2$ achieve the best performance.

4. Experiments

4.1. Datasets and evaluation metrics

The **MVB** dataset [1] is a large scale benchmark for baggage ReID. It contains 22,660 images of 4,519 labelled baggage items. These baggage images are

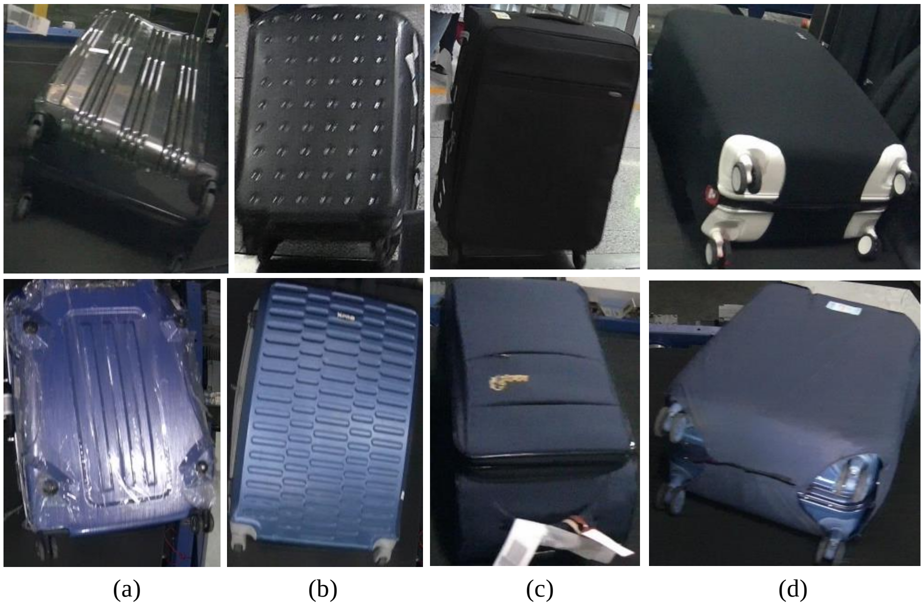

Figure 5: Baggage examples from the MVB dataset. Each column corresponds to one surface material. (a) aluminum; (b) plastic; (c) cloth; (d) other material.

captured from seven cameras in an airport. The pose of baggage items and cameras are relatively fixed. The labels of cameras are used as the annotates of views. In addition, each baggage item is annotated with one of four labels to specify the surface material. There are 4019 identities with 20,176 images for training and 500 identities with 2,484 images for testing. The test set is split into 1,052 probe images and 1,432 gallery images.

The **VeRI-776** dataset [33] consists of vehicle images captured in the real-word unconstrained traffic scenario. It contains about 50,000 images of 776 vehicles, in which each vehicle is captured by 2∼18 cameras with different view points, illuminations, resolutions and occlusions. The vehicles are labeled with bounding boxes, types, colors and brands. There are 37,778 images used for training. The remaining 11,579 images are used for testing.

The **Market-1501** dataset [34] contains bounding boxes from a person detector which have a large intersection over union overlap with manually annotated bounding boxes. It contains 32,668 images of 1,501 persons from 6 camera views. There are 751 identities with 12,936 images for training and 750 identities with 23,100 images for testing. The test set is split into 3,368 probe images and 19,732 gallery images.

The performance in the ReID tasks, *e.g.*, person ReID and vehicle ReID, is assessed using the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) [34]. So the CMC and the mAP are employed in this work to evaluate the performance of the QuadNet ReID models. We first conduct ablation studies on the MVB dataset. Then we compare the proposed QuadNet with the state-of-the-art methods on the three datasets.

4.2. Important details

The QuadNet is built using the Pytorch framework. ResNet50 [19] or a variant of ResNext101 [35] is used as the backbone of QuadNet. The backbone networks are pretrained on ImageNet [36]. The number of output channels of the fully connected layer is set as 4019 which is the number of baggage identities. In training, the input image is resized and cropped to 356 × 356. The Adam method [37] is adopted to optimize the model, where the batch size is set as 24 and the initial learning rate is set as 0.00015. The learning rate is divided by 10 at the $70th$ epoch and the $100th$ epoch. The training stops at the $120th$ epoch. The random horizontal flips with 0.5 probability are used to prevent overfitting and the random local blur with 0.5 probability is used to deal with the motion blur in baggage images. The weights of feature extractors for the view classifier and the identity classifier are shared.

In testing, the horizontal and vertical flipping are applied to the original image. We combine the features of the original image with the variants obtained by horizontal and vertical flipping for feature matching. The k-reciprocal re-ranking method [38] is used to improve the performance of ReID models.

4.3. Ablation study

In this section, two groups of experiments are designed to illustrate the effectiveness of each component in the proposed QuadNet model on the MVB

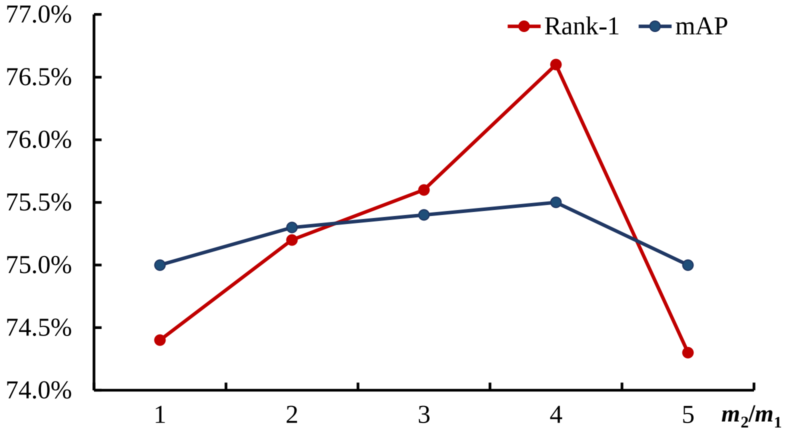

Figure 6: Influence of the margins m_1 and m_2 on the proposed quadruplet loss. m_1 is set as 0.3 empirically. The performance of the QuadNet model is evaluated on the MVB dataset when $m_2 = km_1$, where $k = 1, 2, 3, 4, 5$.

dataset. The first group of experiments compare different loss functions and backbone networks for model selection. The second group of experiments evaluate the effectiveness of the proposed view-aware attentional local features, the random local blur and the multi-task learning of materials.

4.3.1. Evaluating effectiveness of the quadruplet loss

We conduct experiments with different loss functions and different backbone networks, namely ResNet50 [19] and ResNext101 [35]. These experiments are divided into four parts. Firstly, we analyze the influence of the two margins m_1 and m_2 in Eqn. 7 on our QuadNet model. The two margins are used to control the distances between positive and negative pairs, which is important for feature learning. In this experiment, the value of m_1 is set as 0.3 empirically and the performance of our model is evaluated when $m_2 = km_1$, where $k = 1, 2, 3, 4, 5$. As shown in Fig. 6, our model with quadruplet loss has the best accuracy and mAP when $m_2 = 4m_1$. Then, we compare the proposed quadruplet loss with the well-known triplet loss [3] and the identification loss (IDE) [2] individually. Next, we combine the quadruplet loss with the IDE and center loss [39] for better performance in baggage ReID. Finally, we evaluate the use of different depth

Table 1: Comparison of different losses on the MVB dataset.

Models	Network	MVB	
		rank-1	mAP
Triplet [3]	ResNet50	72.6%	74.1%
IDE [2]	ResNet50	68.8%	64.6%
Quadruplet	ResNet50	**76.6%**	**75.5%**
Triplet+IDE	ResNet50	77.5%	74.6%
Quadruplet+IDE	ResNet50	**80.3%**	**76.0%**
Triplet+IDE+Center	ResNet50	77.8%	75.2%
Quadruplet+IDE+Center	ResNet50	**82.6%**	**80.1%**
Triplet+IDE	ResNext101	82.8%	81.1%
Triplet+IDE+Center	ResNext101	83.1%	81.4%
Quadruplet+IDE+Center (QuadNet-1)	ResNext101	**85.5%**	**84.8%**

convolutional networks in the backbone of the QuadNet model. The results for all above models are listed in Tab. 1.

As shown in Tab. 1, the rank-1 accuracy of our quadruplet loss outperforms the triplet loss by 4% and outperforms the IDE loss by nearly 8%. It indicates that the quadruplet loss can effectively solve the multi-view problem in baggage ReID. Then, we combine the quadruplet loss with the IDE loss and center loss, denoted as Quadruplet+IDE+Center. This combination achieves an improvement of about 5% over the original model for both rank-1 and mAP. It outperforms the counterpart model of triplet loss (*i.e.*, Triple+IDE +Center [4]) by over 6% for both rank-1 and mAP. These experimental results demonstrate the advantages of the multi-view quadruplet loss with the multi-view sampling strategy. In addition, we increase the depth of the backbone network from ResNet50 to ResNext101. The experimental results show that the Quadruplet+IDE+Center model with deeper backbone network, denoted as QuadNet-1, achieves a better performance for baggage ReID. Therefore, the subsequent experiments are based on the ResNext101 backbone.

Table 2: Evaluating the performance of components of QuadNet on the MVB dataset.

Models	rank-1	
	w/o re-rank	re-rank
QuadNet-1	85.5%	86.1%
+Random local blur	86.1%	86.8%
+View-aware local feature	86.6%	87.1%
+Material multi-task (QuadNet-2)	87.3%	87.5%
+Geometry transformation (QuadNet-3)	**87.5%**	**88.0%**

4.3.2. Evaluating effectiveness of the other components

In this section, we evaluate the effectiveness of the proposed view-aware attentional local features, the online random local blur and the multi-task learning of materials in baggage ReID. We introduce these three methods into the QuadNet-1 model one by one to evaluate their individual effectiveness in baggage ReID. The experimental results of rank-1 for these methods are listed in Tab. 2. First, fusing the view-aware local features with global features achieves a better performance over the original QuadNet-1. This improvement indicates that our model can extract discriminative local features from different views. Then, on adding these three training methods in QuadNet-1, the resulting network, QuadNet-2, achieves 87.3% of rank-1 on the MVB dataset. These results indicate that the random local blur improves robustness of the model against motion blur, and the multi-task learning of materials helps to distinguish baggage items with similar appearances but different materials. In the test phase, the geometric transformation and k-reciprocal re-ranking method [38] improve the performance of our model further. The final model, QuadNet-3, reaches 88.0% of rank-1 on the MVB dataset.

4.4. Comparison with the state-of-the-art

In order to highlight the significance of the proposed QuadNet for ReID tasks, we compare it with several recent state-of-the-art methods using the three datasets. First, we compare the QuadNet model with other methods in

Table 3: Comparison with the state-of-the-art methods on the MVB dataset.

Models	MVB	
	rank-1	mAP
MSN (PRCV 2019) [1]	50.2%	-
IDE-resnet50 (arXiv 2016) [2]	68.8%	64.6%
IDE-densenet121 (arXiv 2016) [2]	69.6%	64.9%
Verification-IDE (ACM TMM 2019) [12]	68.8%	63.8%
PCB (ECCV 2018) [11]	66.3%	63.6%
OSNet-x1 (ICCV 2019) [40]	74.8%	73.1%
OSNet-x0 (ICCV 2019) [40]	71.3%	68.0%
PCB+MHN4 (ICCV 2019) [41]	70.0%	67.8%
IDE+MHN6 (ICCV 2019) [41]	68.7%	65.6%
DGNet (CVPR 2019) [42]	67.1%	60.9%
Strong baseline (CVPRW 2019) [4]	77.8%	75.2%
QuadNet-3 (Ours)	**87.5%**	**86.1%**

the MVB dataset [1] to evaluate its effectiveness in baggage ReID. Then, in order to compare more fairly with well-known ReID methods and to evaluate the generalization of the QuadNet model in other ReID tasks, *i.e.*, person ReID and vehicle ReID, we compare the QuadNet model with other state-of-the-art methods on the VeRi-776 dataset [33] and the Market-1501 dataset [34].

The results for the MVB dataset are listed in Tab. 3. The Merged Siamese Network (MSN) [1] is a typical Baggage ReID method. It achieves 50.2% of rank-1 on the MVB dataset. The proposed QuadNet model outperforms it by 37.3% for rank-1. The Verification-IDE model uses dual losses combination (IDE loss and verification loss) to enhance discrimination of the features. The proposed QuadNet model outperforms the Verification-IDE model by 18.7% and 22.3% for rank-1 and mAP respectively. Our model outperforms a local features based PCB model [11] by 21.2% and 22.5% for rank-1 and mAP respectively. The OSNet [40] proposes the omni-scale features as the representations of images

Table 4: Comparison with the state-of-the-art methods on the VeRi-776 dataset.

Models	VeRi-776	
	rank-1	mAP
OIFE+ST (ICCV 2017) [5]	92.35%	51.42%
GS-TRE (IEEE TMM 2018) [43]	96.24%	59.47%
PNVR (CVPR 2019) [44]	94.3%	74.3%
RS+MT+K+S (ICCV 2019) [45]	92.86%	71.88%
VANet (ICCV 2019) [22]	89.78%	66.34%
AAVER (ICCV 2019) [46]	89.78%	66.34%
VAMI+STR (CVPR 2018) [6]	85.92%	61.32%
QuadNet-3 (Ours)	**96.6%**	**80.1%**

for person ReID. We re-implement it and test it on the MVB dataset. Our model outperforms OSNet by 12.7% and 13% for rank-1 and mAP respectively. The proposed QuadNet model also outperforms the recent state-of-the-art ReID methods, MHN [41] and DGNet [42], by 17.5% and 20.4% for rank-1 respectively.

In the work VAMI [6], the authors provide annotations of views for the VeRi-776. The results tested on the VeRi-776 dataset are listed in Tab. 4. VANet [22] is proposed to tackle the challenge of view variation in vehicle images. It achieves 89.78% for rank-1 and 66.34% for mAP on VeRi-776. Our QuadNet model outperforms it by 6.8% and 13.8% for rank-1 and mAP respectively. The GS-TRE method [43] learns a group-sensitive triplet embedding by deep metric learning. It achieves a state-of-the-art performance, 96.24% for rank-1 and 59.47% for mAP. Our QuadNet model outperforms it by 0.4% and 20.6% respectively. The QuadNet model also outperforms many other recently proposed methods on the VeRi-776 dataset, such as the PNVR [44], RS+MT+K+S [45] and AAVER [46] models.

The views of a person can be divided into front, right, left and back. It is time-consuming and laborious to manually annotate view labels, so we train a view classifier using the RAP dataset [52] for view classification. We report the

Table 5: Comparison with the state-of-the-art methods on the Market-1501 dataset.

Models	Market-1501	
	rank-1	mAP
SVDNet (ICCV 2017) [47]	82.3%	62.1%
PCB (ECCV 2018) [11]	93.8%	81.6%
MLFN (CVPR 2018) [48]	90.0%	74.3%
CACM (CVPR 2019) [49]	94.7%	84.5%
OSNet (ICCV 2019) [40]	94.8%	84.9%
DGNet (CVPR 2019) [42]	94.8%	86.0%
MHN-6 (ICCV 2019) [41]	95.1%	85.0%
ABD-Net (ICCV 2019) [50]	95.6%	88.3%
Deep-Person (PR 2020) [51]	92.3%	79.6%
End-to-end Ensemble (PR 2020) [26]	93.1%	82.2%
QuadNet-3 (Ours)	**95.4%**	**88.6%**

results on the Market-1501 dataset in Tab. 5. The QuadNet model outperforms the local features based PCB model [11] by 1.6% for rank-1 and 7% for mAP. Our model outperforms the omni-scale features based OSNet [40] by 0.6% for rank-1 and 3.7% for mAP. The DGNet model [42] proposes a joint learning framework that couples feature learning and data generation. QuadNet outperforms it by 0.6% for rank-1 and 2.6% for mAP. Deep-Person [51] proposes a novel three-branch framework to learn highly discriminative features for person ReID. The QuadNet model outperforms it by 3.1% and 9% for rank-1 and mAP respectively on the Market-1501 dataset. The End-to-end Ensemble method [26] is proposed to address the problem of overfitting for person ReID. Our QuadNet model outperforms it by 2.3% for rank-1 and 6.4% for mAP. Our QuadNet model also outperforms other recent methods on the Market-1501 dataset, such as CACM [49] and MHN-6 [41].

5. Conclusion

In this paper, we contribute the QuadNet model to solving the multi-view problem in baggage ReID at three levels. It consists of a multi-view hard example sampling strategy, view-aware attentional local features and a novel quadruplet loss. To our knowledge, it is the first work that proposes the quadruplet loss for multi-view learning, which explicitly concentrates on reducing the intra-class distances caused by view variations and increasing the inter-class distances of baggage images captured in the same view. The final framework QuadNet obtains the state-of-the-art performances on the MVB, VeRi-776 and Market-1501 datasets. In the future work, we will optimize the view-aware attentional module in the QuadNet model with transformer which has demonstrated great effectiveness in computer vision tasks. We will extend the multi-view quadruplet loss for face recognition to tackle the large pose variations of face.

References

[1] Z. Zhang, D. Li, J. Wu, Y. Sun, L. Zhang, MVB: A large-scale dataset for baggage re-identification and merged siamese networks, in: Springer Conference on Pattern Recognition and Computer Vision, 2019, pp. 84–96.

[2] L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: Past, present and future, arXiv preprint arXiv:1610.02984 (2016).

[3] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737 (2017).

[4] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2019, pp. 4321–4329.

[5] Z. Wang, L. Tang, X. Liu, Z. Yao, et al, Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification, in: IEEE International Conference on Computer Vision, 2017, pp. 379–387.

[6] Y. Zhou, L. Shao, Viewpoint-aware attentive multi-view inference for vehicle re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6489–6498.

[7] Y. Zhou, L. Liu, L. Shao, Vehicle re-identification by deep hidden multi-view inference, IEEE Transactions on Image Processing 27 (7) (2018) 3275–3287.

[8] R. R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: Springer European Conference on Computer Vision, 2016, pp. 791–808.

[9] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.

[10] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.

[11] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling and a strong convolutional baseline, in: Springer European Conference on Computer Vision, 2018, pp. 480–496.

[12] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned CNN embedding for person reidentification, ACM Transactions on Multimedia 14 (1) (2018) 1–20.

[13] P. Chhabra, N. K. Garg, M. Kumar, Content-based image retrieval system using ORB and SIFT features, Springer Neural Computing and Applications 32 (7) (2020) 2725–2733.

[14] M. Kumar, P. Chhabra, N. K. Garg, An efficient content based image retrieval system using BayesNet and K-NN, Springer Multimedia Tools and Applications 77 (16) (2018) 21557–21570.

[15] M. Bansal, M. Kumar, M. Kumar, 2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors, Springer Multimedia Tools and Applications 80 (12) (2021) 18839–18857.

[16] Y. Yang, Z. Lei, S. Zhang, H. Shi, S. Z. Li, Metric embedded discriminative vocabulary learning for high-level person representation, in: AAAI Conference on Artificial Intelligence, 2016, pp. 3648–3654.

[17] C.-X. Ren, X.-L. Xu, Z. Lei, A deep and structured metric learning method for robust person re-identification, Elsevier Pattern Recognition 96 (2019) 106995.

[18] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[20] S. Singh, U. Ahuja, M. Kumar, K. Kumar, M. Sachdeva, Face mask detection using yolov3 and faster R-CNN models: COVID-19 environment, Springer Multimedia Tools and Applications 80 (13) (2021) 19753–19768.

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[22] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, Y. Wei, Vehicle re-identification with viewpoint-aware metric learning, in: IEEE International Conference on Computer Vision, 2019, pp. 8282–8291.

[23] A. Kumar, M. Kumar, A. Kaur, Face detection in still images under occlusion and non-uniform illumination, Springer Multimedia Tools and Applications 80 (10) (2021) 14565–14590.

[24] M. Bansal, M. Kumar, M. Kumar, K. Kumar, An efficient technique for object recognition using shi-tomasi corner detection algorithm, Soft Computing 25 (6) (2021) 4423–4432.

[25] S. Gupta, M. Kumar, A. Garg, Improved object recognition results using SIFT and ORB feature detector, Multimedia Tools and Applications 78 (23) (2019) 34157–34171.

[26] A. Serbetci, Y. S. Akgul, End-to-end training of CNN ensembles for person re-identification, Elsevier Pattern Recognition 104 (2020) 107319.

[27] H. Zhao, M. Tian, S. Sun, S. Jing, X. Tang, Spindle Net: Person re-identification with human body region guided feature decomposition and fusion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 907–915.

[28] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose-invariant embedding for deep person re-identification, IEEE Transactions on Image Processing 28 (9) (2019) 4500–4509.

[29] C. Patruno, R. Marani, G. Cicirelli, E. Stella, T. D'Orazio, People re-identification using skeleton standard posture and color descriptors from RGB-D data, Elsevier Pattern Recognition 89 (2019) 77–90.

[30] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 384–393.

[31] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, HydraPlus-Net: Attentive deep features for pedestrian analysis, in: IEEE International Conference on Computer Vision, 2017, pp. 350–359.

[32] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, Attention driven person re-identification, Elsevier Pattern Recognition 86 (2019) 143–155.

[33] X. Liu, L. Wu, M. Tao, H. Ma, A deep learning-based approach to progressive vehicle re-identification for urban surveillance, in: Springer European Conference on Computer Vision, 2016, pp. 869–884.

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[35] X. Pan, P. Luo, J. Shi, X. Tang, Two at once: Enhancing learning and generalization capacities via IBN-Net, in: Springer European Conference on Computer Vision, 2018, pp. 464–479.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, ImageNet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[37] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015, pp. 1–11.

[38] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.

[39] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Springer European Conference on Computer Vision, 2016, pp. 499–515.

[40] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, in: IEEE International Conference on Computer Vision, 2019, pp. 3702–3712.

[41] B. Chen, W. Deng, J. Hu, Mixed high-order attention network for person re-identification, in: IEEE International Conference on Computer Vision, 2019, pp. 371–381.

[42] Z. Zheng, X. Yang, Z. Yu, et al, Joint discriminative and generative learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2138–2147.

[43] Y. Bai, Y. Luo, F. Gao, S. Wang, Y. Wu, L.-Y. Duan, Group-sensitive triplet embedding for vehicle reidentification, IEEE Transactions on Multimedia 20 (9) (2018) 2385–2399.

[44] B. He, J. Li, Y. Zhao, Y. Tian, Part-regularized near-duplicate vehicle re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3997–4005.

[45] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, X. Yang, PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data, in: IEEE International Conference on Computer Vision, 2019, pp. 211–220.

[46] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, R. Chellappa, A dual-path model with adaptive attention for vehicle re-identification, in: IEEE International Conference on Computer Vision, 2019, pp. 6132–6141.

[47] Y. Sun, L. Zheng, W. Deng, S. Wang, SVDNet for pedestrian retrieval, in: IEEE International Conference on Computer Vision, 2017, pp. 3800–3808.

[48] X. Chang, T. M. Hospedales, T. Xiang, Multi-level factorisation net for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2109–2118.

[49] W. Yang, H. Huang, Z. Zhang, X. Chen, S. Zhang, Towards rich feature discovery with class activation maps augmentation for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1389–1398.

[50] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, Z. Wang, ABD-Net: Attentive but diverse person re-identification, in: IEEE International Conference on Computer Vision, 2019, pp. 8351–8361.

[51] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, Deep-person: Learning discriminative deep features for person re-identification, Elsevier Pattern Recognition 98 (2020) 107036.

[52] D. Li, Z. Zhang, X. Chen, H. Ling, K. Huang, A richly annotated dataset for pedestrian attribute recognition, arXiv preprint arXiv:1603.07054 (2016).