A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification

Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, Jianyang Gu

***Abstract*—This study explores a simple but strong baseline for person re-identification (ReID). Person ReID with deep neural networks has progressed and achieved high performance in recent years. However, many state-of-the-art methods design complex network structures and concatenate multi-branch features. In the literature, some effective training tricks briefly appear in several papers or source codes. The present study collects and evaluates these effective training tricks in person ReID. By combining these tricks, the model achieves 94.5% rank-1 and 85.9% mean average precision on Market1501 with only using the global features of ResNet50. The performance surpasses all existing global- and part-based baselines in person ReID. We propose a novel neck structure named as batch normalization neck (BNNeck). BNNeck adds a batch normalization layer after global pooling layer to separate metric and classification losses into two different feature spaces because we observe they are inconsistent in one embedding space. Extended experiments show that BNNeck can boost the baseline, and our baseline can improve the performance of existing state-of-the-art methods. Our codes and models are available at: https://github.com/michuanhaohao/reid-strong-baseline**

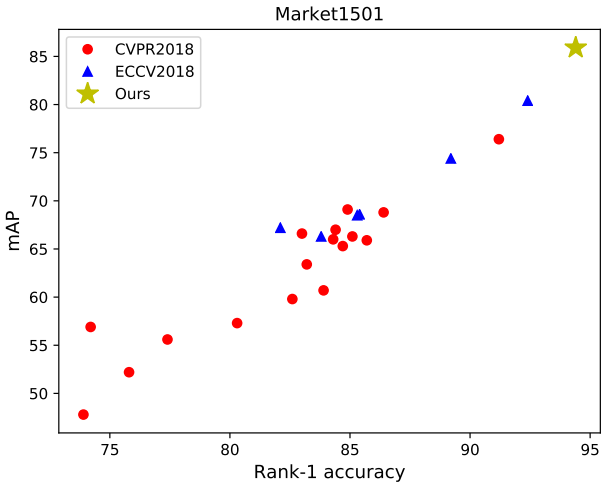
***Index Terms*—Person ReID, Baseline, Tricks, BNNeck, Deep learning.**

I. Introduction

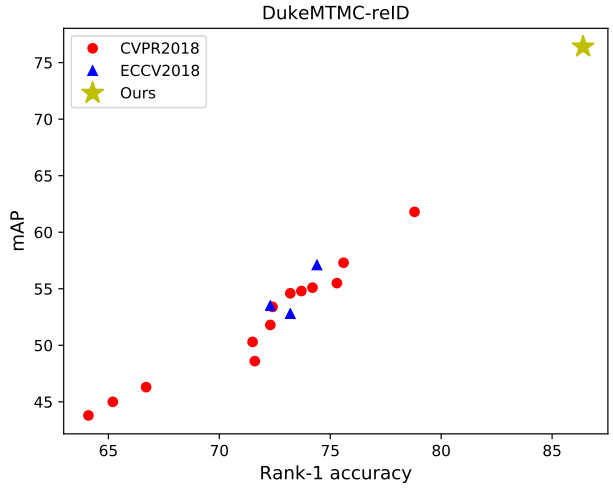
Person re-identification (ReID) is widely applied in video surveillance and criminal investigation applications [2]. Person ReID with deep neural networks has progressed and achieved high performance in recent years [3]–[5]. Apart from many novel and effective ideas being proposed, the improvement of baseline model plays a key role. The importance of baseline model should not be ignored. However, few works [5]–[7] have focused on the design of an effective baseline. The performance of such baselines has gradually become obsolete due to the rapid development of person ReID. In the literature, some effective training tricks or refinements briefly appear in several papers or source codes. In the present study, we design a strong and effective baseline for person ReID by collecting and evaluating such effective training tricks.

This study has three motivations. First, we survey articles published in ECCV2018 and CVPR2018 of the past year. As shown in Fig. 1, most of the previous works were expanded on poor baselines. On Market1501, only two in 23 baselines surpassed 90% rank-1 accuracy. The rank-1 accuracies of four baselines were even lower than 80%. On DukeMTMC-reID [8], all baselines did not surpass 80% rank-1 accuracy or 65% mean average precision (mAP). Achieving improvements on poor baselines cannot strictly demonstrate the effectiveness of some methods. Thus, a strong baseline is crucial in promoting research development.

Second, we discover that some works were unfairly com-



(a) Market1501



(b) DukeMTMC-reID

Fig. 1. Performance of different baselines on Market1501 and DukeMTMC-reID datasets. We compare our strong baseline with other baselines published in CVPR2018 and ECCV2018.

This paper is the extended version of the oral paper [1] accepted by TRMTMCT2019 Workshop in CVPR2019.

Hao Luo, Wei Jiang, Youzhi Gu, Jianyang Gu is with the College of Control Science and Enginneering, Zhejiang University, Hangzhou 310027, China; E-mail: haoluocsc@zju.edu.cn, jiangwei_zju@zju.edu.cn.

Fuxu Liu is with Ping An Technology, Shenzhen, China; E-mail: LIUFUXU641@pingan.com.cn

Xingyu Liao is with Chinese Academy of Sciences, Beijing , China; E-mail: randall@mail.ustc.edu.cn

Shenqi Lai is with Xi'an Jiaotong University, Xi'an, China; E-mail: laishenqi@stu.xjtu.edu.cn

Manuscript received June 19, 2019; revised MMDD, YYYY.

pared with other state-of-the-art methods. The improvements were mainly from training tricks rather than methods themselves. However, the training tricks were understated in the paper; thus, readers ignored them, thereby exaggerating the effectiveness of the method. We suggest that reviewers consider these tricks when commenting on academic papers.

Third, the industry prefers simple and effective models over concatenating many local features in the inference stage. In pursuit of high accuracy, researchers in the academia always combine several local features or utilize semantic information from pose estimation or segmentation models. Nevertheless, such methods bring extra consumption. Large features also greatly reduce the speed of the retrieval process. Thus, we use tricks to improve the capability of the ReID model and only use global features to achieve high performance.

On the basis of the aforementioned considerations, the motivations of designing a strong baseline are summarized as follows:

- For the academia, we survey many works published on top conferences and discover that most of them were expanded on poor baselines. We aim to provide a strong baseline for researchers to achieve high accuracies in person ReID.
- For the community, we aim to provide references to reviewers regarding tricks that will affect the performance of the ReID model. We suggest that reviewers consider these tricks when comparing the performance of different methods.
- For the industry, we aim to provide effective tricks for acquiring improved models without extra consumption.

Many effective training tricks have been presented in papers or open-sourced projects. We collect tricks and evaluate each of them on ReID datasets. After numerous experiments, we select six tricks to introduce in this study. We propose a novel bottleneck structure, namely, batch normalization neck (BNNeck). As classification and metric losses are inconsistent in the same embedding space, BNNeck optimizes these two losses in two different embedding spaces. In addition, person ReID task mainly focuses on ranking performance, such as cumulative match characteristic (CMC) curve and mAP, but ignores the clustering effect, such as intra-class compactness and inter-class separability. However, clustering effect is important to some special tasks, such as object tracking, which must decide on a distance threshold to separate positive samples from negative ones. An easy approach to overcome this problem is to train the model with center loss. Finally, we add the tricks into a widely used baseline to obtain our modified baseline (the backbone is ResNet50), which achieves 94.5% and 85.9% mAP on Market1501.

To determine whether these tricks are generally useful or not, we design extended experiments from three aspects. First, we follow the cross-domain ReID settings in which the models are trained and evaluated on different datasets. Cross-domain experiments can show whether the tricks boost the models or simply suppress overfitting in the training dataset. Second, we evaluate all tricks with different backbones, such as ResNet18, SeResNet50, and IBNNet-50. All backbones achieve improvements from our training tricks. Third, we reproduce some state-of-the-art methods on our modified baseline. Experimental results show that our baseline obtains better performance than those reported in published papers. Although our baseline achieves surprising performance, some methods remain effective on our baseline. Thus, our baseline can be a strong baseline for the ReID community.

As a supplement, we discover that different works select different image sizes and batch size numbers. Therefore, we explore their effects on model performance. The contributions of this study are summarized as follows:

- We collect effective training tricks for person ReID. We evaluate the improvements from each trick on two widely used datasets.
- We observe the inconsistency between ID loss and triplet and propose a novel neck structure, namely, BNNeck.
- We observe that the ReID task ignores intra-class compactness and inter-class separability and claim that center loss can compensate for it.
- We provide a strong ReID baseline, which achieves 94.5% and 85.9% mAP on Market1501. The results are obtained with global features provided by ResNet50 backbone. To our best knowledge, this result is the best performance acquired by global features in person ReID.
- We design extended experiments to demonstrate that our baseline can be a strong baseline for the ReID community.
- As a supplement, we evaluate the influences of image size and batch size number on the performance of ReID models.

II. Related Works

This section focuses on deep learning baseline for person ReID. In addition, existing approaches compared with our strong baseline for deep person ReID are introduced.

A. Baseline for Deep Person ReID

Recent studies on person ReID mostly focus on building deep convolutional neural networks (CNNs) to represent the features of person images in an end-to-end learning manner. GoogleNet [9], ResNet [10], DenseNet [11], etc are widely used backbone networks. The baselines can be classified into two main genres in accordance with the loss function, *i.e.* classification loss and metric loss. For classification loss, Zheng *et al.* [6] proposed ID-discriminative embedding (IDE) to train the re-ID model as image classification which is fine-tuned from the ImageNet [12] pre-trained models. Classification loss is also called ID loss in person ReID because IDE is trained by classification loss. However, ID loss requires an extra fully connected (FC) layer to predict the logits of person IDs in the training stage. In the inference stage, such FC layer is removed, and the feature from the last pooling layer is used as the representation vector of the person image.

Different from ID loss, metric loss regards the ReID task as a clustering or ranking problem. The most widely used baseline based on metric learning is training model with triplet loss [13]. A triplet includes there images, *i.e.* anchor, positive, and negative samples. The anchor and positive samples belong

to the same person ID, whereas the negative sample belongs to a different person ID. Triplet loss minimizes the distance from the anchor sample to the positive sample and maximizes the distance from the anchor sample to the negative one. However, triplet loss is greatly influenced by the sample triplets. Inspired by FaceNet [14], Hermans *et al.* proposed an online hard example mining for triplet loss (TriHard loss). Most current methods are expanded on the TriHard baseline. Combining ID loss with TriHard loss is also a popular manner of acquiring a strong baseline [3].

Apart from designing different losses, some works focus on building effective baseline model for deep person ReID. In [7], three good practices were proposed to build an effective CNN baseline toward person ReID. Their most important practice is adding a batch normalization (BN) layer after the global pooling layer. Similar to these models, the baseline uses a global feature for image representation. Sun *et al.* [5] proposed part-based convolutional baseline (PCB). Given an image input, PCB outputs a convolutional descriptor consisting of several part-level features. Both baselines have achieved good performance in person ReID.

B. Some Existing Approaches for Deep person ReID

On the basis of the aforementioned baselines, many methods have been proposed in the past few years. We divide these works into striped-based, pose-guided, mask-guided, attention-based, GAN-based, and re-ranking methods.

Stripe-based methods, which divide the image into several stripes and extract local features for each stripe, play an important role in person ReID. Inspired by PCB, the typical methods includes AlignedReID++ [3], MGN [15], SCPNet [16], etc. Stripe-based local features are effective in boosting the performance of the ReID model. However, they always encounter the problem of pose misalignment.

Pose-guided methods [17]–[20] use an extra pose/skeleton estimation model to acquire human pose information. Pose information can exactly align corresponding parts of two person images. However, an extra model brings additional computation consumption. A trade off between the performance and speed of the model is important.

Mask-guided models [21]–[23] use mask as external cues to remove the background clutters in pixel level and contain body shape information. For example, Song *et al.* [21] proposed a mask-guided contrastive attention model that applies binary segmentation masks to learn features separately from the body and background regions. Kalayeh *et al.* [22] proposed SPReID, which uses human semantic parsing to harness local visual cues for person ReID. Mask-guided models extremely rely on accurate pedestrian segmentation model.

Attention-based methods [24]–[27] involve an attention mechanism to extract additional discriminative features. In comparison with pixel-level masks, attention region can be regraded as an automatically learned high-level 'mask'. A popular model is Harmonious Attention CNN (HA-CNN) model porposed by Li *et al.* [25]. HA-CNN combines the learning of soft pixel and hard regional attentions along with simultaneous optimization of feature representations. An advantage of attention-based models is that they do not require a segmentation model to acquire mask information.

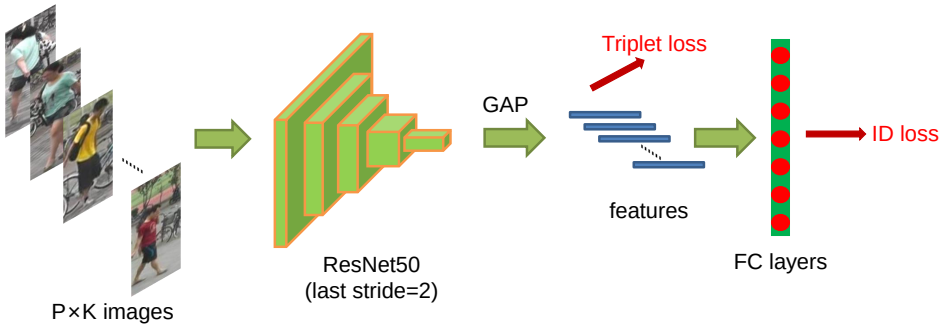
GAN-based methods [28]–[31] address the limited data for person ReID. Zheng *et al.* [28] first used GAN [32] to generate images for enriching ReID datasets. The GAN model randomly generates unlabeled and unclear images. On the basis of [28], PTGAN [29] and CamStyle [30] were proposed to bridge domain and camera gaps for person ReID, respectively. Qian *et al.* [31] proposed PNGAN for obtaining a new pedestrian feature and transforming a person into normalized poses. The final feature is obtained by combining the pose-independent features with original ReID features. With the development of GAN, many ganbased methods have been proposed to generate high quality for supervised and unsupervised person ReID tasks.

Re-ranking methods [33]–[35] are post-processing strategies for image retrieval. In general, person ReID simply uses Euclidean or cosine distances in the retrieval stage. Zhong *et al.* [33] a k-reciprocal encoding method to re-rank the ReID results. Given an image, a k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbors into a single vector, which is used for re-ranking under the Jaccard distance. The final distance is computed as the combination of the original and Jaccard distances. Shen *et al.* [34] proposed a deep group-shuffling random walk (DGRW) network for fully utilizing the affinity information between gallery images in training and testing processes. In the retrieval stage, DGRW can be regarded as a re-ranking method. Re-ranking is a critical step in improving retrieval accuracy.

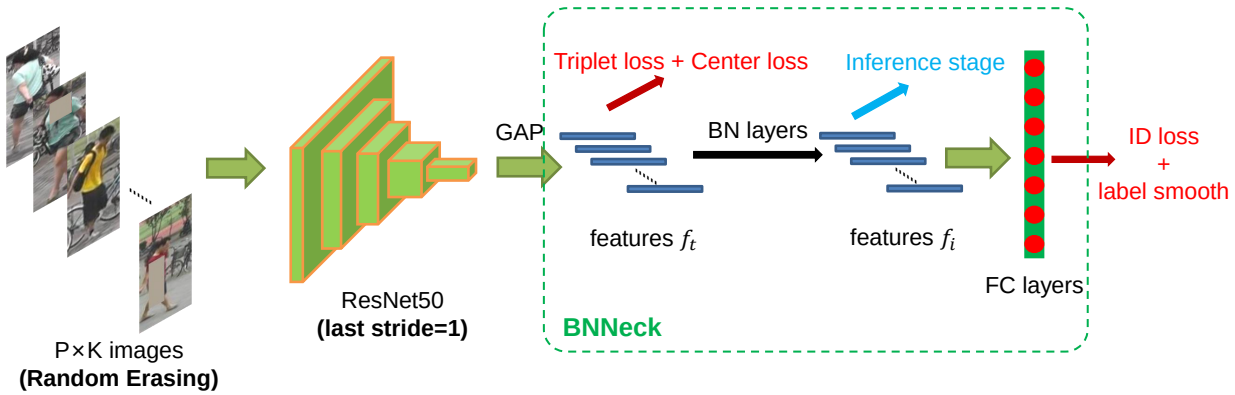
III. STANDARD BASELINE

In this section, a widely used baseline for the academia and industry is introduced. For convenience, such baseline is called standard baseline. The backbone of the standard baseline is ResNet50 [10]. In the training stage, the pipeline includes the following steps:

1) We initialize the ResNet50 with pre-trained parameters on ImageNet and change the dimension of the fully connected layer to N. N denotes the number of identities in the training dataset.
2) We randomly sample P identities and K images of per person to constitute a training batch. Finally, the batch size equals to $B = P \times K$. In this study, we set $P = 16$ and $K = 4$.
3) We resize each image into 256×128 pixels and pad the resized image 10 pixels with zero values. We randomly crop it into a 256×128 rectangular image.
4) Each image is flipped horizontally with 0.5 probability.
5) Each image is decoded into 32-bit floating point raw pixel values in $[0, 1]$. RGB channels are normalized by subtracting 0.485, 0.456, 0.406 and dividing by 0.229, 0.224, 0.225, respectively.
6) The model outputs ReID features f and ID prediction logits p.
7) ReID features f is used to calculate triplet loss [36]. ID prediction logits p is used to calculated cross-entropy loss. The margin m of triplet loss is 0.3.



(a) The pipeline of the standard baseline.



(b) The pipeline of our modified baseline.

Fig. 2. Pipelines of the standard baseline and our modified baseline.

8) Adam method is adopted to optimize the model. The initial learning rate is 0.00035 and is decreased by 0.1 at the 40th epoch and 70th epoch. Training epochs total 120.

Fig. 2a presents the framework of the standard baseline, and additional details are available in our open source code.

IV. Our Strong Baseline and Training Tricks

This section introduces some effective training tricks in person ReID. Our proposed BNNeck structure is discussed in detail. The intra-class compactness and inter-class separability problem for person ReID is also raised. Most tricks can be expanded on the standard baseline without changing the model architecture. Fig. 2b shows training strategies and model architecture.

A. Warmup Learning Rate

Learning rate has a great effect on the performance of a ReID model. Standard baseline is initially trained with a large and constant learning rate. In [37], a warmup strategy was applied to bootstrap the network for enhanced performance. In practice, we spend 10 epochs, thereby linearly increasing the learning rate from 3.5×10^{-5} to 3.5×10^{-4}, as shown in Fig. 3. The learning rate is decayed to 3.5×10^{-5} and 3.5×10^{-6} at 40th and 70th epochs, respectively. The learning rate $lr(t)$ at epoch t is compute as follows:

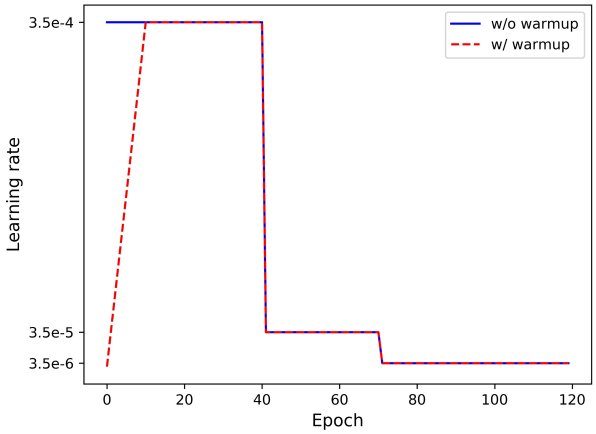


Fig. 3. Comparison of learning rate schedules. With warmup strategy, the learning rate is linearly increased in the first 10 epochs.

$$lr(t) = \begin{cases} 3.5 \times 10^{-4} \times \frac{t}{10} & \text{if } t \leq 10 \\ 3.5 \times 10^{-4} & \text{if } 10 < t \leq 40 \\ 3.5 \times 10^{-5} & \text{if } 40 < t \leq 70 \\ 3.5 \times 10^{-6} & \text{if } 70 < t \leq 120 \end{cases} \quad (1)$$

B. Random Erasing Augmentation

In person ReID, persons in the images are sometimes occluded by other objects. To address the occlusion problem

and improve the generalization capability of ReID models, Zhong *et al.* [38] proposed a new data augmentation approach, namely, random erasing augmentation (REA). In practice, for an image I in a mini-batch, the probability of REA undergoing random erasing is p_e, and the probability of remaining unchanged is $1-p_e$. REA randomly selects a rectangular region I_e with size (W_e, H_e) in image I, and erases its pixels with random values. Assuming the area of image I and region I_e are $S = W \times H$ and $S_e = W_e \times H_e$, respectively, we denote $r_e = \frac{S_e}{S}$ as the area ratio of erasing rectangle region. In addition, the aspect ratio of region I_e is randomly initialized between r_1 and r_2. To determine a unique region, REA randomly initializes a point $\mathcal{P} = (x_e, y_e)$. If $x_e + W_e \leq W$ and $y_e + H_e \leq H$, then we set the region $I_e = (x_e, y_e, x_e + W_e, y_e + H_e)$ as the selected rectangle region. Otherwise we repeat the above process until an appropriate I_e is selected. With the selected erasing region I_e, each pixel in I_e is assigned to the mean value of image I.

In this study, we set hyper-parameters to $p = 0.5, 0.02 < S_e < 0.4, r_1 = 0.3, r_2 = 3.33$, respectively. Some examples are shown in Fig. 4.

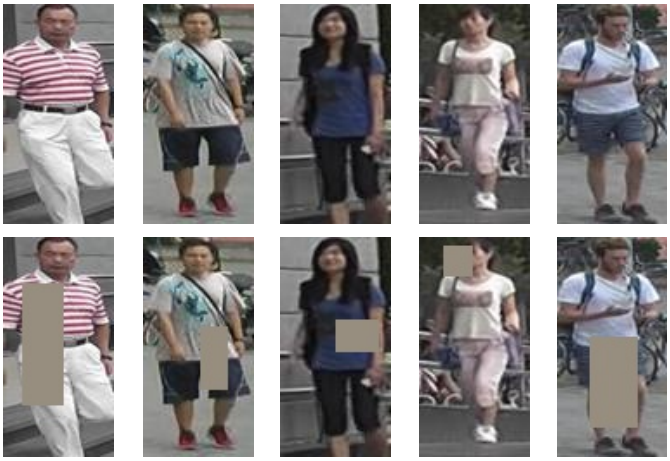


Fig. 4. Examples of random erasing augmentation. The first row shows the five original training images. The second row presents the processed images.

C. Label Smoothing

The IDE [6] network is a basic baseline in person ReID. The last layer of IDE, which outputs the ID prediction logits of images, is a fully connected layer with a hidden size equal to the number of persons N. Given an image, we denote y as truth ID label and p_i as ID prediction logits of class i. The cross-entropy loss is computed as follows:

$$L(ID) = \sum_{i=1}^{N} -q_i \log(p_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases} \quad (2)$$

As the category of classification is determined by the person ID, we call such loss function as ID loss in this study.

Nevertheless, person ReID can be a one-shot learning task because person IDs of the testing set do not appear in the training set. The ReID model must be prevented from overfitting training IDs. Label smoothing (LS) proposed in [39] is a widely used method to prevent overfitting for a classification task. The construction of q_i is changed to:

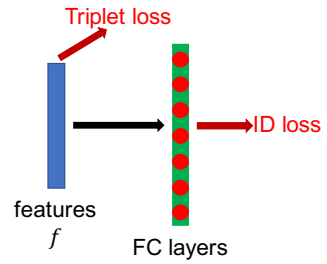
$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon & \text{if } i = y \\ \varepsilon / N & \text{otherwise,} \end{cases} \quad (3)$$

where ε is a small constant to encourage the model to be less confident on the training set. In this study, ε is set to be 0.1. When the training set is not large, LS can significantly improve the model performance.

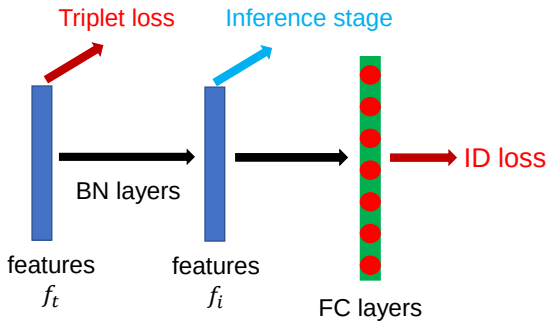
D. Last Stride

A high spatial resolution always enriches feature granularity. In [5], Sun *et al.* removed the last spatial down-sampling operation in the backbone network to increase the size of the feature map. For convenience, we denote the last spatial down-sampling operation in the backbone network as the last stride. The last stride of ResNet50 is set to be 2. When fed into an image with 256×128 size, the backbone of ResNet50 outputs a feature map with a spatial size of 8×4. If last stride is changed from 2 to 1, then we can obtain a feature map with increased spatial size (16×8). This manipulation only slightly increases the computation cost and does not involve extra training parameters. However, an increased spatial resolution brings significant improvement.

E. BNNeck



(a) Neck of the standard baseline.



(b) Designed BNNeck. In the inference stage, we select f_i following the BN layer to perform the retrieval.

Fig. 5. Comparison between standard neck and our designed BNNeck.

Most works combined ID and triplet losses to train ReID models. Fig. 5(a) shows that both losses constrain the same feature f in the standard baseline.. However, the targets of these two losses are inconsistent in the embedding space.

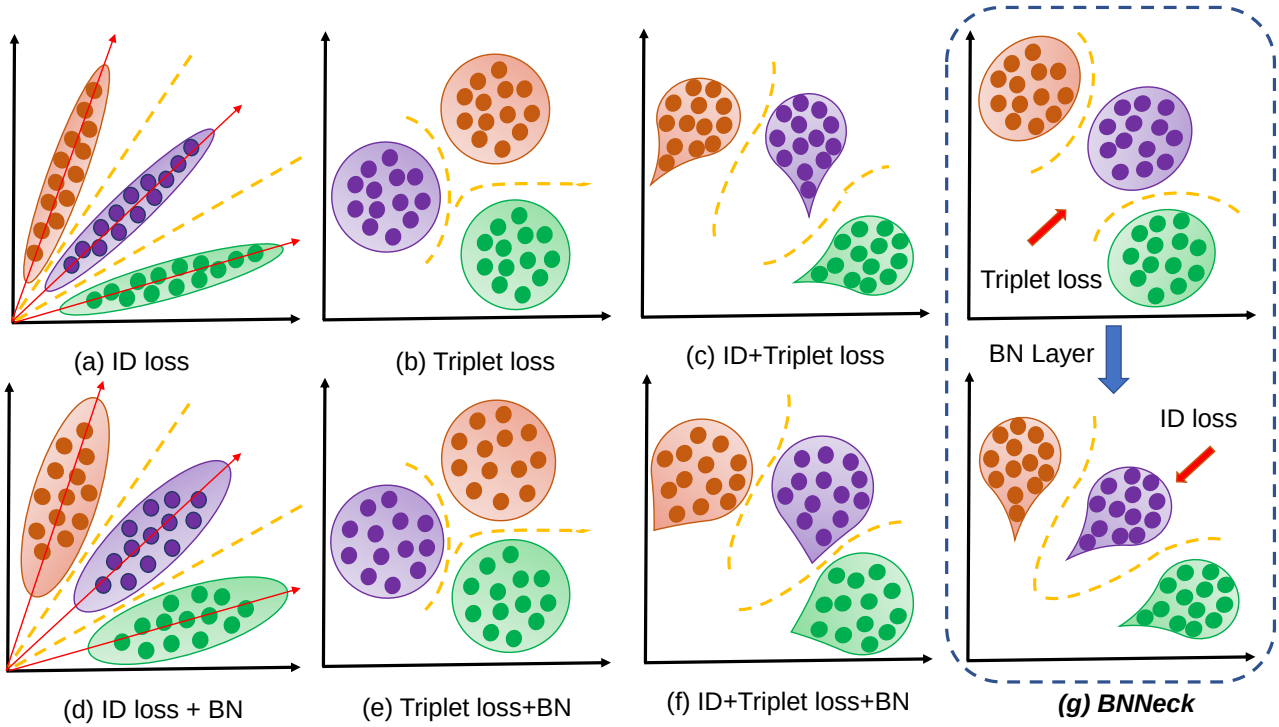


Fig. 6. Two-dimensional visualization of sample distribution in the embedding space supervised by different losses and neck structures. (a∼g) correspond to (a∼g) in Fig. 10. Points of different colors represent embedding features from different person IDs. The yellow dotted lines stand for decision surfaces. For better understanding, we make some overexpression compared to Fig. 9.

Fig. 6(a) presents that ID loss constructs several hyperplanes to separate the embedding space into different subspaces. The features of each class are distributed affinely in different subspaces. Cosine distance is more suitable than Euclidean distance for the model optimized by ID loss in the inference stage. However, as shown in 6(b), triplet loss enhances intra-class compactness and inter-class separability in the Euclidean space. Inter-class distance sometimes is smaller than intra-class distance because triplet loss cannot provide globally optimal constraint. A widely used method is to combine ID and triplet losses to train the model. This approach allows the model to learn additional discriminative features. Nevertheless, for image pairs in the embedding space, ID loss optimizes the cosine distances whereas triplet loss focuses on the Euclidean distances. If we use both losses to optimize a feature space simultaneously, then their goals may be inconsistent. During training, a possible problem is that one loss is reduced, whereas the other loss oscillates or even increases, as shown in Fig.8. Finally, triplet loss may influence the clear decision surfaces of ID loss, and ID loss may reduce the intra-class compactness of triplet loss. The feature distribution is tadpole shaped. Therefore, directly combining these two losses can boost the performance, but it is not the best way.

Xiong *et al.* [7] added a BN layer between feature and ID loss, which is same as Fig. 10(d). The authors claimed that the BN layer overcomes the overfitting and boosts the performance of IDE baseline. However, we consider that the BN layer can smoothen the feature distribution in the embedding space. For ID loss (Fig. 6(a)), the BN layer will enhance the intra-class compactness. The BN layer can improve the performance of ID loss because the features close to the affine center lack clear decision surfaces and are difficult to distinguish. Nevertheless, such layer increases the cluster radius of intra-class feature for triplet loss. Thus, the decision surfaces of 6(e)(f) are stricter than those of Fig. 6(b)(c).

To overcome this problem, we design a structure, namely, BNNeck, as shown in Fig. 5(b). BNNeck adds a BN layer after features and before classifier FC layers. The BN and FC layers are initialized through Kaiming initialization proposed in [40]. The feature before the BN layer is denoted as f_t. We let f_t pass through a BN layer to acquire the feature f_i. In the training stage, f_t and f_i are used to compute triplet and ID losses, respectively. Fig. 5(g) shows that f_t not only can keep a compact distribution from but also acquires ID knowledge from ID loss. Affected by the BN layer and ID loss, the distribution of f_i is tadpole shaped. In comparison with 5(c), f_i has clear decision surfaces because of the weaker influence of the triple loss. Additional details are introduced in Section V-D.

In the inference stage, we select f_i to perform the person ReID task. Cosine distance metric can achieve better performance than Euclidean distance metric. Experimental results in Table. I show that BNNeck can improve the performance of the ReID model by a large margin.

F. Center Loss

Person ReID is always regarded as a retrieval/ranking task. The evaluation protocols, *i.e.* CMC curve and mAP, are

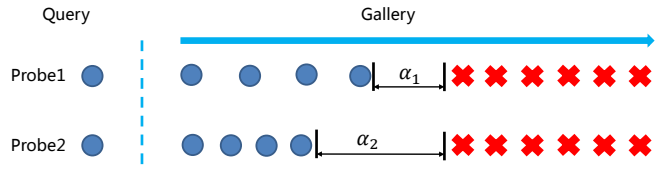


Fig. 7. Visualized demonstration that ranking task ignores the intra-class compactness of positive pairs. Blue circles and red crosses represent positive and negative samples, respectively. In the direction of the arrow, the feature distance of two samples is increasing. Although two cases can acquire the same ranking results, probe2 is easy for the tracking task that must decide on a threshold to separate positive and negative samples.

determined by the ranking results but ignore the clustering effect. However, for some ReID applications, such as tracking task, an important step is to decide on a distance threshold to separate positive and negative objects. As shown in Fig. 7, two cases can acquire the same ranking results, but probe2 is easy for the tracking task because of its intra-class compactness of positive pairs.

Focusing on relative distance, triplet loss is computed as:

$$L_{Tri} = [d_p - d_n + \alpha]_+, \tag{4}$$

where d_p and d_n are feature distances of positive and negative pairs. α is the margin of triplet loss, and $[z]_+$ equals $max(z, 0)$. In this study, α is set to 0.3. However, the triplet loss only considers the difference between d_p and d_n and ignores their absolute values. For instance, when $d_p = 0.3$ and $d_n = 0.5$, the triplet loss is 0.1. For another case, when $d_p = 1.3$ and $d_n = 1.5$, the triplet loss also is 0.1. Triplet loss is determined by two randomly sampled person IDs. Ensuring that $d_p < d_n$ in the entire training dataset is difficult. In addition, intra-class compactness is ignored.

To compensate for the drawbacks of the triplet loss, we involve center loss [41] intraining, simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers. The center loss function is formulated as follows:

$$\mathcal{L}_C = \frac{1}{2}\sum_{j=1}^{B}\left\|\boldsymbol{f}_{t_j} - \boldsymbol{c}_{y_j}\right\|_2^2, \tag{5}$$

where y_j is the label of the jth image in a mini-batch. $\boldsymbol{c}_{y_j}$ denotes the y_ith class center of deep features, and B is the batch size number. The formulation effectively characterizes the intra-class variations. Minimizing center loss increases intra-class compactness. Our model includes three losses as follows:

$$L = L_{ID} + L_{Triplet} + \beta L_C \tag{6}$$

where β is the balanced weight of center loss. In our baseline, β is set to be 0.0005.

V. Experiment

A. Datasets

We evaluate our models on Market1501 [42] and DukeMTMC-reID [8] datasets, because both datasets are widely used and large scale. Following the previous works, we use rank-1 accuracy and mAP for evaluation on both datasets.

Market1501 contains 32,217 images of 1,501 labeled persons of six camera views. The training set has 12,936 images from 751 identities, and the testing set has 19,732 images from 750 identities. In testing, 3,368 hand-drawn images from 750 identities are used as queries to retrieve the matching persons in the database. Single-query evaluation is used in this study.

DukeMTMC-reID is a new large-scale person ReID dataset and collects 36,411 images from 1,404 identities of eight camera views. The training set has 16,522 images from 702 identities, and the testing set has 19,889 images from other 702 identities. Single-query evaluation is used in this study.

B. Influences of Each Trick (Same domain)

Model	Market1501		DukeMTMC	
	r = 1	mAP	r = 1	mAP
Baseline-S	87.7	74.0	79.7	63.7
+warmup	88.7	75.2	80.6	65.1
+REA	91.3	79.3	81.5	68.3
+LS	91.4	80.3	82.4	69.3
+stride=1	92.0	81.7	82.6	70.6
+BNNeck	94.1	85.7	86.2	75.9
+center loss	94.5	85.9	86.4	76.4

TABLE I
PERFORMANCE OF DIFFERENT MODELS IS EVALUATED ON MARKET1501 AND DUKEMTMC-REID DATASETS. BASELINE-S REPRESENTS THE STANDARD BASELINE INTRODUCED IN SECTION III.

The standard baseline introduced in section III achieves 87.7% and 79.7% rank-1 accuracies on Market1501 and DukeMTMC-reID, respectively. The performance of standard baseline is similar to most baselines reported in other papers. Warmup strategy, random erasing augmentation, LS, stride change, BNNeck, and center loss are individually added to the model training process. The designed BNNeck boosts performance to a greater extent than other tricks, especially on DukeMTMC-reID. Finally, with these tricks, the baseline acquires 94.5% rank-1 accuracy and 85.9% mAP on Market1501. On DukeMTMC-reID, the baseline reaches 86.4% rank-1 accuracy and 76.4% mAP. Thus, these training tricks boost the performance of the standard baseline by over 10% mAP. To achieve such improvement, we only involve an extra BN layer and do not increase training time.

C. Influences of Each Trick (Cross domain)

To explore the effectiveness further, we present the results of cross-domain experiments in Table. II. In overview, three tricks, namely, warmup strategy, LS, and BNNeck, greatly boost the cross-domain performance of ReID models. Stride change and center loss seem to have no influence on the performance. However, REA harms the models in cross-domain ReID task. When our modified baseline is trained without REA, it achieves 41.4% and 54.3% rank-1 accuracies on Market1501 and DukeMTMC-reID datasets, respectively. The performance surpasses those of the standard baseline by a large margin. We infer that by REA masking the regions of training images, the model learns additional knowledge in the source domain and performs poorly in the target domain.

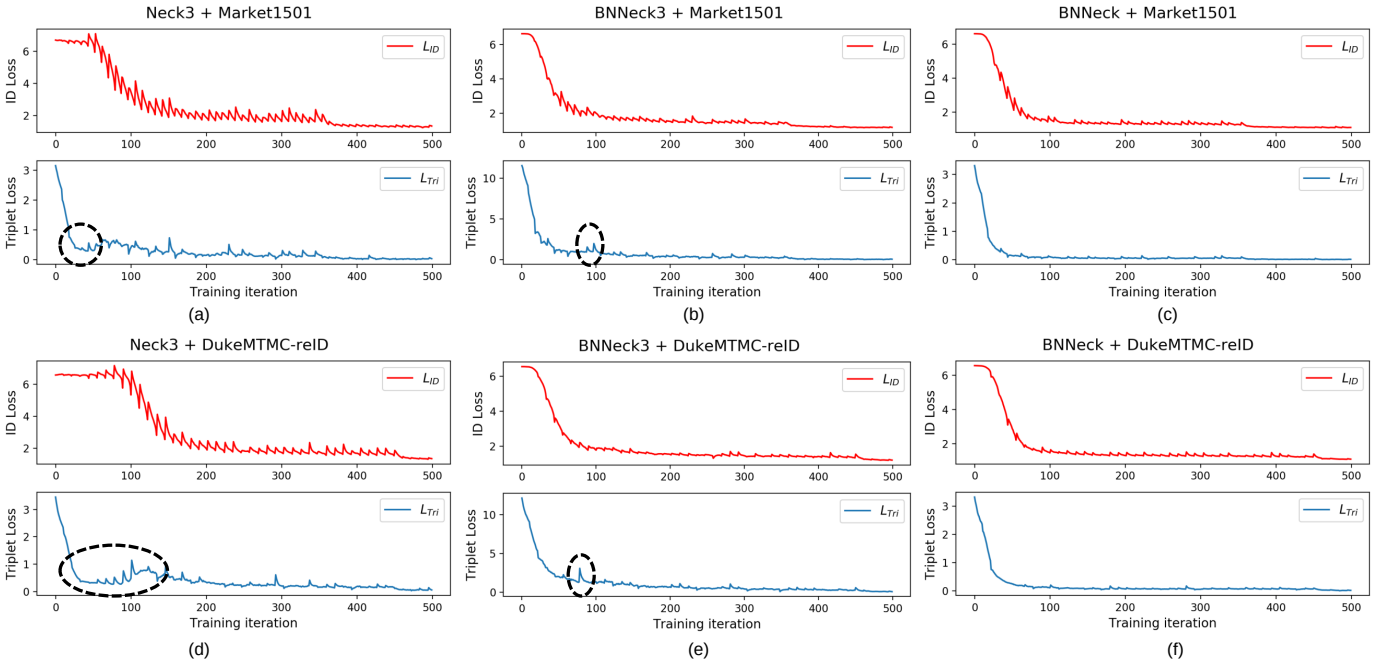


Fig. 8. ID and triplet loss curves of different models on Market1501 and DukeMTMC-reID datasets. We train the models with Neck3, BNNeck3, and our proposed BNNeck, respectively. Black ovals mark the inconsistency between ID and triplet losses. We show that BNNeck suppresses the inconsistency and smoothens the triplet loss curve.

Model	M→D		D→M	
	r = 1	mAP	r = 1	mAP
Baseline	24.4	12.9	34.2	14.5
+warmup	26.3	14.1	39.7	17.4
+REA	21.5	10.2	32.5	13.5
+LS	23.2	11.3	36.5	14.9
+stride=1	23.1	11.8	37.1	15.4
+BNNeck	26.7	15.2	47.7	21.6
+center loss	27.5	15.0	47.4	21.4
-REA	41.4	25.7	54.3	25.5

TABLE II
PERFORMANCE OF DIFFERENT MODELS EVALUATED ON CROSS-DOMAIN DATASETS. M→D MEANS THAT WE TRAIN THE MODEL ON MARKET1501 AND EVALUATE IT ON DUKEMTMC-REID.

Finally, our baseline achieves good performance and can be used as a strong baseline for cross-domain ReID task.

D. Analysis of BNNeck

Feature	Metric	Market1501		DukeMTMC	
		r = 1	mAP	r = 1	mAP
f	Neck1	89.4	77.5	78.9	65.3
f	Neck2	91.0	80.9	82.5	69.4
f	Neck3	92.0	81.7	82.6	70.6
f_i	BNNeck1	93.1	83.9	85.2	74.0
f_i	BNNeck2	90.3	79.1	82.5	67.9
f_i	BNNeck3	92.5	81.8	83.3	71.9
f_t	BNNeck	**94.2**	85.5	85.7	74.4
f_i	BNNeck	94.1	**85.7**	**86.2**	**75.9**

TABLE III
ABLATION STUDY OF DIFFERENT NECK STRUCTURES IN FIG. 10.

1) Different neck structures: To discuss the effectiveness of our BNNeck, we design several different neck structures, as shown as Fig. 10. In addition, some ablation studies also are analysed in Table III. Neck3 outperforms Neck1 and Neck2. In addition, BNNeck2 is worse than Neck2, but BNNeck1 is better than Neck1. Our BNNeck achieves the best performance on two benchmarks. In summary, we present the following observations/conclusions. 1) Without the BN layer, integrating ID and triplet losses is better than only using one loss. 2) The BN layer is effective for ID loss but is invalid for triplet loss. 3) Our BNNeck that sets triplet loss before the BN layer is a reasonable neck structure.

2) Inconsistency between ID loss and Triplet loss: To verify that ID and triplet losses are inconsistent in the same feature space, we train the models with Neck3, BNNeck3, and our proposed BNNeck. Fig. 10 shows that these three neck structures use ID and triplet losses to optimize the same feature. Fig. 8 presents the training loss curves of 500 iterations. In Fig. 8a and 8d, the triplet loss initially increases and then decays in the loss curves marked by black ovals, showing a clear confrontation between triplet and ID losses. In comparison with Neck3, BNNeck3 adds a BN layer after f. In Figs. 8b and 8e, the BN layer weakens but does not eliminate the inconsistency. However, for BNNeck in Figs. 8c and 8f, the inconsistency is suppressed, and the triplet loss curves are smoothened. In conclusion, the BN layer can weaken the inconsistency between the losses, and separating them into two different feature spaces is important.

3) Visualization of feature distribution: To analyze the distribution of the different features in Fig. 10, we train models in MNIST dataset. The visualization has considerable noise because the number of person IDs on ReID benchmark

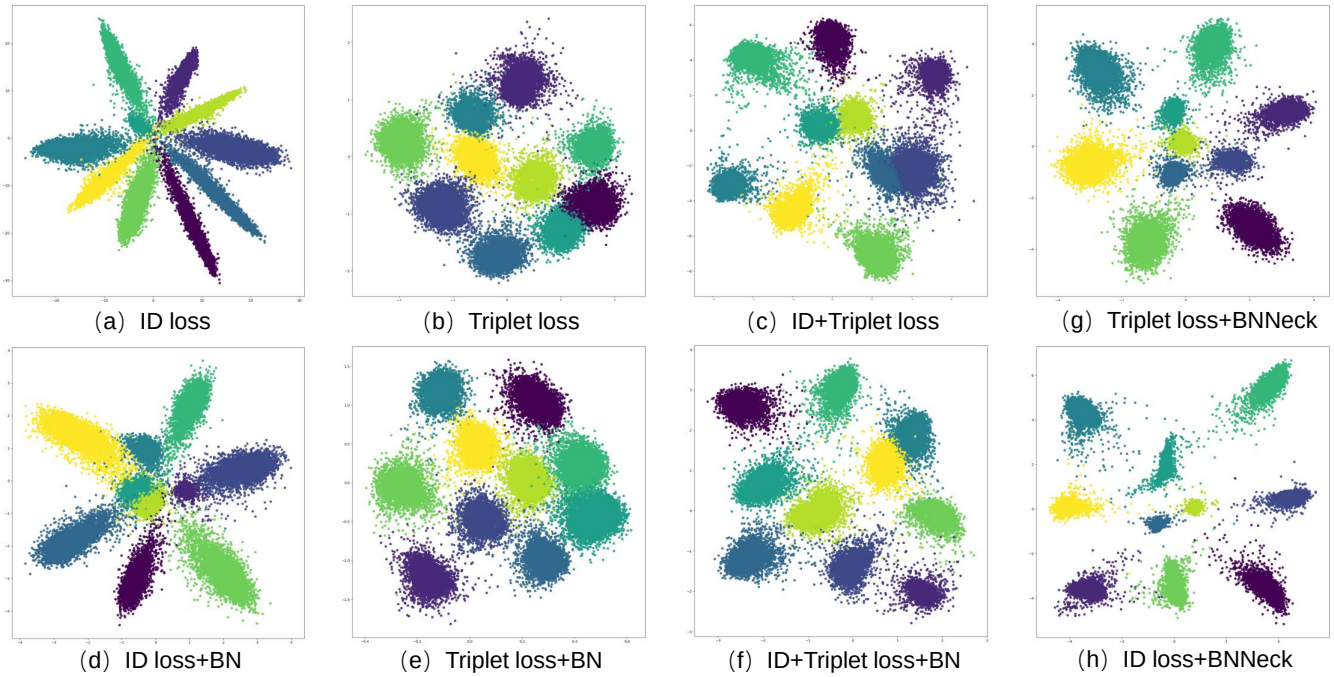


Fig. 9. 2D visualization of feature distribution in the embedding space supervised by different losses and neck structures on MNIST dataset. (a∼f) correspond to (a∼f) in Figs. 10. (g) and (h) are related to Fig. 10(g). The feature dimension is set to 2 for the best view. The BN layer will smoothen the feature.

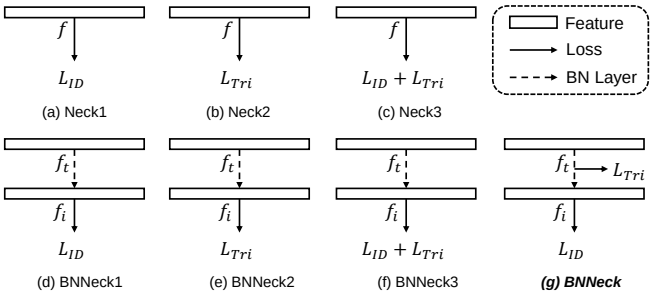


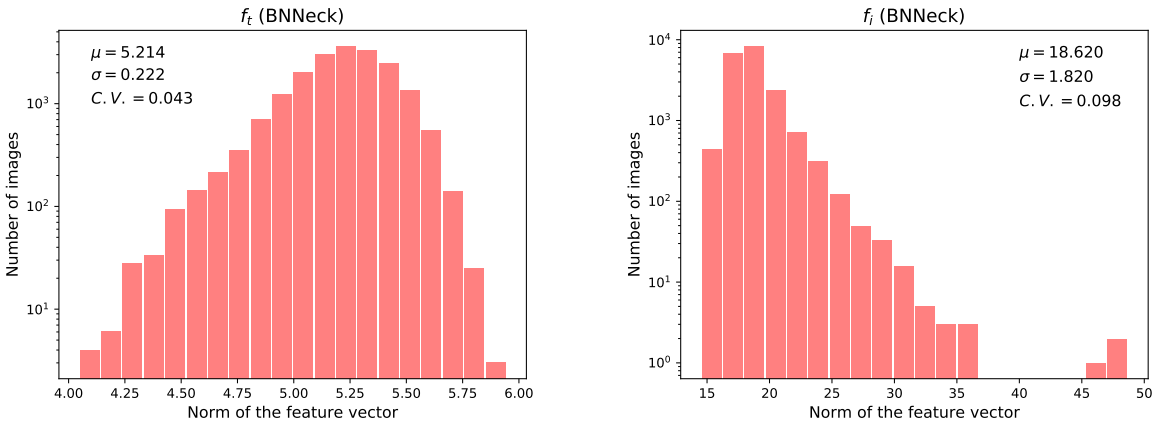
Fig. 10. Different neck structures for ablation study. (a∼c) are standard neck structures, and (d∼f) add an additional BN layer. Different losses includes L_{ID}, L_{Tri}, and $L_{ID} + L_{Tri}$. (g) is our proposed BNNeck that separates triplet and ID losses into different feature spaces.

is large, and the number of images from each person ID is small. By contrast, MNIST only has 10 categories, and each category consists of thousands of samples, making the feature distribution clear and robust. Fig. 9 shows the results. ID and triplet losses have two different feature distributions. When integrating these two losses in Fig. 9c, the clustered distribution is stretched to be tadpole shaped. The distributions of (dsimf) are more gaussian than those of (asimc) because of the BN effect. Figs. 9g and 9h show that our BNNeck separates triplet and ID losses into two different feature spaces. The feature distribution of triplet loss remains clustered, and that of ID loss has clear decision surfaces similar to Figs. 9a and 9b.

We summarize our conclusions or observations as follows: 1) The feature distributions of ID and triplet losses are affined and clustered, *i.e.*, they are inconsistent. 2) The feature distribution of ID+Triplet loss is tadpole shaped. 3) The BN layer can smoothen/normalize the feature distribution and enhance the intra-class compactness for ID loss but reduce it for triplet loss. 4) We separate triplet and ID losses into two different and suitable feature spaces.

4) Two feature space of BNNeck: Although the results on MNIST in Fig. 9 can efficiently support our conclusion, image classification and person ReID are two different tasks. We perform statistical analysis on the norm distribution of f_t and f_i in BNNeck on Market1501 dataset. The mean value μ and standard deviation σ of feature norm are calculated. To analyze the separability of feature distribution, Coefficient of Variation $C.V. = \mu/\sigma$ is also present. As shown in Fig. 11, f_i and f_t are distributed differently in the feature space. f_t is compactly and gaussian distributed in an annular space because it is directly optimized by triplet loss. However, we consider f_i as a tadpole-shaped distribution because ID loss stretches intra-class distribution. The maximum value of f_i is 48.70, whereas the μ is 18.62. $C.V.$ of f_t is 0.043, but $C.V.$ of f_i reaches 0.98, which demonstrates that f_i is distributed more discretely than f_t. In conclusion, BNNeck provide two different and suitable feature spaces for triplet loss and ID loss.

5) Metric space for BNNeck: We evaluate the performance of two different features (f_t and f_i) with Euclidean and cosine distance metrics. All models are trained without center loss in Table. IV. We observe that cosine distance metric performs better than Euclidean distance metric for f_t. As ID loss directly constrains the features followed the BN layer, f_i can be clearly separated by several hyperplanes. The cosine distance can measure the angle between feature vectors; thus, cosine distance metric is more suitable than Euclidean distance metric



(a) f_t (before BN) (b) f_i (after BN)

Fig. 11. Histograms of feature norm for f_t and f_i in BNNeck on Market1501 dataset. μ, σ, $C.V.$ are mean value, standard deviation, and Coefficient of Variation.

Feature	Metric	Market1501 r = 1	Market1501 mAP	DukeMTMC r = 1	DukeMTMC mAP
f (w/o BNNeck)	Euclidean	92.0	81.7	82.6	70.6
f_t	Euclidean	94.2	85.5	85.7	74.4
f_t	Cosine	94.2	85.7	85.5	74.6
f_i	Euclidean	93.8	83.7	86.6	73.0
f_i	Cosine	**94.1**	**85.7**	**86.2**	**75.9**

TABLE IV
ABLATION STUDY OF BNNECK. f (W/O BNNECK) IS BASELINE WITHOUT BNNECK. BNNECK INCLUDES FEATURES f_t AND f_i. WE EVALUATE THEIR PERFORMANCE WITH EUCLIDEAN AND COSINE DISTANCE.

Feature	β	Market1501 r = 1	Market1501 mAP	Market1501 R	DukeMTMC r = 1	DukeMTMC mAP	DukeMTMC R
f_t	0	94.2	85.5	0.407	85.7	74.4	0.424
	0.0005	93.9	85.7	0.405	86.5	75.1	0.420
	0.005	94.2	85.7	0.394	86.2	75.4	0.417
	0.05	94.4	85.4	0.365	86.4	74.9	0.403
	0.5	92.6	81.1	**0.311**	85.5	72.2	**0.363**
f_i	0	94.1	85.7	0.590	86.2	75.9	0.568
	0.0005	94.5	85.9	0.589	86.4	76.4	0.564
	0.005	94.3	85.9	0.595	86.8	76.4	0.566
	0.05	94.3	85.7	0.592	86.7	76.5	0.560
	0.5	94.1	84.7	0.593	87.4	76.9	0.554

TABLE V
EVALUATION WITH DIFFERENT WEIGHTS OF CENTER LOSS β. R IS THE RATIO OF INTRA-CLASS DISTANCE TO INTER-CLASS DISTANCE.

for f_i. However, f_t is simultaneously close to triplet loss and constrained by ID loss. The two types of metrics achieve similar performance for f_t.

Overall, BNNeck significantly improve the performance of ReID models. We select f_i with cosine distance metric to perform the retrieval in the inference stage.

E. Analysis of Center loss

We discuss the influence of center loss on intra-class compactness. We consider that average intra-class distance cannot fully represent the intra-class compactness because it ignores inter-class distance. For convenience, the average intra-class and inter-class distances are denoted as D_p and D_n, respectively. Inspired by [43], the ratio of D_p to D_n is used to measure the clustering effect of feature distribution. The ratio is computed as $R = D_p/D_n$. We set β to different values and evaluate rank-1, mAP, and R of the models. Table V presents the results.

For the feature f_t constrained directly by center loss, R decreases as β increases. With β increasing from 0 to 0.5, R is reduced from 0.407 to 0.311 on Market1501 and from 0.424 to 0.363 on DukeMTMC-reID. Hence, center loss can improve intra-class compactness and inter-class separability, thereby bringing a clear boundary between positive and negative samples. When β is set to 0.5, f_t can acquire the best clustering effect but obtains the worse rank-1 and mAP accuracies. However, the BN layer destroys such clustering effect. For feature f_i, the value of R is almost not influenced by β. On the basis of these observations, we arrive at the following conclusions: (1) Center loss boosts intra-class compactness and inter-class separability. (2) The BN layer can destroy the effect of center loss. (3) Increasing the weight of center loss may reduce ranking performance.

Baseline	Loss	Market1501 r = 1	Market1501 mAP	DukeMTMC r = 1	DukeMTMC mAP
IDE [6]	ID	79.5	59.9	-	-
TriNet [36]	Tri	84.9	69.1	-	-
AWTL [44]	Tri	89.5	75.7	79.8	63.4
GP [7]	ID	91.7	78.8	83.4	68.8
PCB [5]	ID	92.3	77.4	81.7	66.9
Our	ID+Tri	**94.5**	**85.9**	**86.4**	**76.4**

TABLE VI
COMPARISON OF DIFFERENT BASELINES ON MARKET1501 AND DUKEMTMC-REID DATASETS. ID AND TRI STANDS FOR ID LOSS AND TRIPLET-BASED LOSS, RESPECTIVELY.

F. Comparison to Other Baselines

We compare our strong baseline with other effective baselines, such as IDE [6], TriNet [36], AWTL [44] and PCB [5]. PCB is a part-based baseline for person ReID. Table VI

Type	Method	N_f	Market1501		DukeMTMC	
			r = 1	mAP	r = 1	mAP
Pose-guided	GLAD [18]	4	89.9	73.9	-	-
	PIE [20]	3	87.7	69.0	79.8	62.0
	PSE [19]	3	78.7	56.0	-	-
Mask-guided	SPReID [22]	5	92.5	81.3	84.4	71.0
	MaskReID [23]	3	90.0	75.3	78.8	61.9
Stripe-based	AlignedReID++ [3]	1	90.6	77.7	81.2	67.4
	SCPNet [16]	1	91.2	75.2	80.3	62.6
	PCB+RPP [5]	6	93.8	81.6	83.3	69.2
	Pyramid [45]	1	92.8	82.1	-	-
	Pyramid [45]	21	95.7	88.2	89.0	79.0
	BFE [46]	2	94.5	85.0	88.7	75.8
	MGN [15]	1	89.8	78.5	-	-
	MGN [15]	8	95.7	86.9	88.7	78.4
Attention-based	Mancs [4]	1	93.1	82.3	84.9	71.8
	DuATM [24]	1	91.4	76.6	81.2	62.3
	HA-CNN [25]	4	91.2	75.7	80.5	63.8
GAN-based	Camstyle [30]	1	88.1	68.7	75.3	53.5
	PN-GAN [31]	9	89.4	72.6	73.6	53.2
Global feature	IDE [6]	1	79.5	59.9	-	-
	SVDNet [47]	1	82.3	62.1	76.7	56.8
	TriNet [36]	1	84.9	69.1	-	-
	AWTL [44]	1	89.5	75.7	79.8	63.4
	Ours	1	**94.5**	**85.9**	**86.4**	**76.4**

TABLE VII
COMPARISON OF STATE-OR-THE-ART METHODS. N_f IS THE NUMBER OF FEATURES USED IN THE INFERENCE STAGE. RK IS k-RECIPROCAL RE-RANKING METHOD [33]

presents the performance of these baselines. The experimental results show that our baseline outperforms IDE, TriNet, and AWTL by a large margin. PCB integrates multi-part features and GP uses effective tricks, and both of them achieves great performance. However, our baseline surpasses them by over 7.1% mAP on both datasets. To our best knowledge, our baseline is the strongest baseline.

G. Comparison to State-of-the-Arts

We compare our strong baseline with state-of-the-art methods in Table. VII. All methods have been divided into different types. Pyramid [45] achieves surprising performance on two datasets, but it concatenates 21 local features of different scales. When only the global feature is utilized, Pyramid obtains 92.8% rank-1 accuracy and 82.1% mAP on Market1501. Our strong baseline can reach 94.5% rank-1 accuracy and 85.9% mAP on Market1501. BFE [46] obtains similar performance to our strong baseline, but it combines features of two branches. Among all methods that only use global features, our strong baseline outperforms AWTL [44] by more than 10% mAP on both Market1501 and DukeMTMC-reID. To our best knowledge, our baseline achieves the best performance when only global features are used.

H. Baseline Meets State-of-the-Arts

We reproduce some popular state-of-the-art methods with our strong baseline. Given numerous outstanding methods are available, we cannot try all of them and select only several typical models such as k-reciprocal re-ranking [33], PCB [5], AligedReID++ [3], CamStyle [30], and MGN [15]. For a fair comparison, we use the same losses as the paper reported to train the models. For instance, AlignedReID++ only uses ID and triplet losses, and we do not use center loss to reproduce it. However, as k-reciprocal re-ranking is a post-processing method of global features, three losses are used to improve its performance. Table VIII shows the details and results, wherein the values in parentheses are the results reported by authors in their papers. In addition, we present the performance of the baselines (with BNNeck) trained by different losses as a reference.

Our baseline boosts the performance of k-reciprocal re-ranking, PCB, AligedReID++, and CamStyle by a large margin. The mAP of k-reciprocal re-ranking achieves +30.6% on Market1501, demonstrating that the performance of baselines is important for methods. In addition, our MGN achieves similar performance to [15] because its accuracies are too high to improve, and [15] uses BNNeck1 structure. Integrating multiple part features can reduce the effect of global features and limit the effectiveness of baselines for PCB and MGN. However, PCB and MGN still obtain better performance than Baseline2, *i.e.*, part-based methods are effective for our baseline. However, CamStyle(Our) outperforms CamStyle [30] but not Baseline1. Our baseline can be a strong baseline for the ReID community because it can boost the performance of some methods, and other methods based on it may be ineffective. To some extent, our baseline efficiently filters effective methods.

I. Performance of Different Backbones

All aforementioned models apply ResNet50 as backbones for clear ablation studies and comparison with other methods.

Models with different backbones, such as ResNet, SeResNet, SeResNeXt, and IBNNet, are evaluated because backbones have a great influence on their performance. As shown in Table IX, deep and large backbones can achieve high performance. For example, ResNet101 outperforms ResNet18 by 2.8% and 9.3% in Rank-1 and mAP accuracy on Market1501, respectively. In addition, the channel attention of SeNet and group convolution of ResNeXt can enhance the performance by a slight margin. IBN-Net50 [48], which replaces the BN layers with instance BN layers for ResNet50, is also effective for our baseline. Specifically, IBN-Net50-a is suitable for standard ReID task and obtains 95.0% and 90.1% rank-1 accuracies on Market1501 and DukeMTMC-reID, respectively. However, IBN-Net50-b achieves 50.1% rank-1 and 29.8% mAP for M→D and 61.7% rank-1 and 32.0% mAP (D→M).

For comparison, IBN-Net50-a achieves 40.0% rank-1 and 25.1% mAP for M→D and 52.9% rank-1 and 25.1% mAP (D→M). In conclusion, IBN-Net-a and IBN-Net50-b are suitable for the same domain task and the cross-domain task, respectively.

VI. SUPPLEMENTARY EXPERIMENTS

We observe that some previous works were conducted with different batch size numbers or image sizes. In this section, we explore their effects on model performance as a supplement.

Method	Reference	Market1501 r = 1	Market1501 mAP	DukeMTMC r = 1	DukeMTMC mAP	Loss
Baseline1	BNNeck1	93.1	83.9	85.2	74.0	L_{ID}
Baseline2	BNNeck	94.1	85.7	86.2	75.9	L_{ID}, L_{Tri}
Baseline3	BNNeck	94.5	85.9	86.4	76.4	L_{ID}, L_{Tri}, L_C
k-reciprocal [7]	CVPR17	95.4(77.1)	94.2(63.6)	90.3(-)	89.1(-)	L_{ID}, L_{Tri}, L_C
PCB [5]	ECCV18	94.0(92.3)	84.0(77.4)	88.6(81.7)	77.2(66.1)	L_{ID}, L_{Tri}
AligedReID++ [3]	PR19	94.3(91.8)	86.5(79.1)	86.5(82.1)	76.9(69.7)	L_{ID}, L_{Tri}
CamStyle [30]	TIP19	93.3(88.1)	81.0(68.7)	80.3(75.3)	60.1(53.5)	L_{ID}
MGN [15]	ACMMM19	95.3(95.7)	86.3(86.9)	89.2(88.7)	78.9(78.4)	L_{ID}, L_{Tri}

TABLE VIII
PERFORMANCE OF SOME STATE-OF-THE-ART METHODS REPRODUCED BY OUR STRONG BASELINE. THE VALUES IN PARENTHESES ARE THE RESULTS REPORTED BY AUTHORS.

Backbone	Market1501 r = 1	Market1501 mAP	DukeMTMC r = 1	DukeMTMC mAP
ResNet18	91.7	77.8	82.5	68.8
ResNet34	92.7	82.7	86.4	73.6
ResNet50	94.5	85.9	86.4	76.4
ResNet101	94.5	87.1	87.6	77.6
SeResNet50	94.4	86.3	86.4	76.5
SeResNet101	94.6	87.3	87.5	78.0
SeResNeXt50	94.9	87.6	88.0	78.3
SeResNeXt101	95.0	88.0	88.4	79.0
IBN-Net50-a	**95.0**	**88.2**	**90.1**	**79.1**
IBN-Net50-b	93.5	83.9	86.4	73.5

TABLE IX
PERFORMANCE OF OUR BASELINE WITH DIFFERENT BACKBONE.

Batch Size $P \times K$	Market1501 r = 1	Market1501 mAP	DukeMTMC r = 1	DukeMTMC mAP
8×3	92.6	79.2	84.4	68.1
8×4	92.9	80.0	84.7	69.4
8×6	93.5	81.6	85.1	70.7
8×8	93.9	82.0	85.8	71.5
16×3	93.8	83.1	86.8	72.1
16×4	93.8	83.7	86.6	73.0
16×6	94.0	82.8	85.1	69.9
16×8	93.1	81.6	86.7	72.1
32×3	94.5	84.1	86.0	71.4
32×4	93.2	82.8	86.5	73.1

TABLE X
PERFORMANCE OF REID MODELS WITH DIFFERENT NUMBERS OF BATCH SIZE.

A. Influences of the Number of Batch Size

The mini-batch of triplet loss includes $B = P \times K$ images. P and K denote the number of different persons and the number of different images per person, respectively. A mini-batch can only contain up to 128 images in one GPU; thus, we cannot perform the experiments with $P = 32, K = 6$ or $P = 32, K = 8$. We remove center loss to find the relation between triplet loss and batch size clearly. Table. X presents the results. However, conclusions do not specifically show the effect of B on performance. A slight trend is that a large batch size is beneficial for model performance. We infer that a large K helps mine hard positive pairs, whereas a large P helps mine hard negative pairs.

Image Size	Market1501 r = 1	Market1501 mAP	DukeMTMC r = 1	DukeMTMC mAP
256×128	93.8	83.7	86.6	73.0
224×224	94.2	83.3	86.1	72.2
384×128	94.0	82.7	86.4	73.2
384×192	93.8	83.1	87.1	72.9

TABLE XI
PERFORMANCE OF REID MODELS WITH DIFFERENT IMAGE SIZES.

B. Influences of Image Size

We feed training images of different sizes and train models without center loss with the setting $P = 16, K = 4$ As shown in Table XI, four models achieve similar performances on both datasets. In our opinion, the image size is not a strictly important factor for the performance of ReID models.

VII. CONCLUSIONS AND OUTLOOKS

In this study, we propose a strong baseline for person ReID with only adding an extra BN layer for standard baseline. Our strong baseline achieves 94.5% rank-1 accuracy and 85.9% mAP on Market1501. To our best knowledge, this result is the best performance achieved by the global features of a single backbone. We evaluate each trick of our baseline on same- and cross-domain ReID tasks. In addition, some state-of-the-art methods can be effectively extended on our baseline. We hope that this work can promote ReID research in the academia and industry.

We observe the inconsistency between ID and triplet losses in previous ReID baselines. To address this problem, we propose a BNNeck to separate both losses into two different feature spaces. Extended experiments show that the BN layer can enhance and reduce the intra-class compactness for ID and triplet losses, respectively. Furthermore, ID loss is suitable for optimizing the feature.

We emphasize that the evaluation of ReID task ignores the clustering effect of representation features. However, the clustering effect is important to some ReID applications, such as tracking task wherein an important step is deciding on a distance threshold to separate positive and negative objects. A simple way to address this problem is using center loss to train the model. Center loss can boost the clustering effect of features, but may reduce the ranking performance of ReID models.

In the future, we will explore additional tricks and effective methods based on this strong baseline. In comparison with face recognition, person ReID still has room for further exploration. In addition, some confusions remain, such as why REA reduces the cross-domain performance in our baseline. Points wherein the conclusion is unclear are worth researching.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China (No. 61633019) and the Science Foundation of Chinese Aerospace Industry (JCKY2018204B053).

REFERENCES

[1] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[2] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Transactions on Multimedia*, 2019.

[3] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "Alignedreid++: Dynamically matching local information for person re-identification," *Pattern Recognition*, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320319302031

[4] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 365–381.

[5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[6] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, p. 13, 2018.

[7] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. T. Zhou, "Good practices on building effective cnn baseline model for person re-identification," in *Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*, vol. 11069. International Society for Optics and Photonics, 2019, p. 110690I.

[8] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[13] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[15] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 274–282.

[16] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, and W. Jiang, "Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification," *arXiv preprint arXiv:1810.06996*, 2018.

[17] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1077–1085.

[18] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 420–428.

[19] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," *IEEE Transactions on Image Processing*, 2019.

[21] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[22] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[23] L. Qi, J. Huo, L. Wang, Y. Shi, and Y. Gao, "Maskreid: A mask based deep ranking neural network for person re-identification," *arXiv preprint arXiv:1804.03864*, 2018.

[24] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5363–5372.

[25] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2285–2294.

[26] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 369–378.

[27] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2119–2128.

[28] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[29] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.

[30] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camstyle: A novel data augmentation method for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1176–1190, 2019.

[31] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 650–667.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

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

[34] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2265–2274.

[35] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2553–2566, 2016.

[36] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[37] X. Fan, W. Jiang, H. Luo, and M. Fei, "Spherereid: Deep hypersphere manifold embedding for person re-identification," *Journal of Visual Communication and Image Representation*, 2019.
[38] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.
[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
[41] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.
[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Computer Vision, IEEE International Conference*, 2015.
[43] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5409–5418.
[44] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6036–6046.
[45] F. Zheng, X. Sun, X. Jiang, X. Guo, Z. Yu, and F. Huang, "A coarse-to-fine pyramidal model for person re-identification via multi-loss dynamic training," *arXiv preprint arXiv:1810.12193*, 2018.
[46] Z. Dai, M. Chen, S. Zhu, and P. Tan, "Batch feature erasing for person re-identification and beyond," *arXiv preprint arXiv:1811.07130*, 2018.
[47] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3800–3808.
[48] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 464–479.