Batch Feature Erasing for Person Re-identification and Beyond

Zuozhuo Dai[1] Mingqiang Chen[1] Siyu Zhu[1] Ping Tan[2]
[1]Alibaba A.I. Labs [2]Simon Fraser University

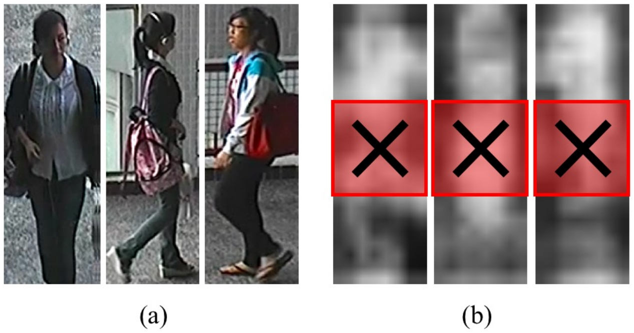
Abstract

This paper presents a new training mechanism called Batch Feature Erasing (BFE) for person re-identification. We apply this strategy to train a novel network with two branches and employing the ResNet-50 as the backbone. The two branches consist of a conventional global branch and a feature erasing branch where the BFE strategy is applied. When training the feature erasing branch, we randomly erase the same region of all the feature maps in a batch. The network then concatenates features from the two branches for person re-identification. Albeit simple, our method achieves state-of-the-art on person re-identification and is applicable to general metric learning tasks in image retrieval problems. For instance, we achieve 75.4% Rank-1 accuracy on the CUHK03-Detect dataset and 83.0% Recall-1 score on the Stanford Online Products dataset, outperforming the existed works by a large margin (more than 6%).

1. Introduction

Person re-identification (re-ID) amounts to identify the same person from multiple detected pedestrian images, typically seen from different cameras without view overlap. It has important applications in surveillance and presents a significant challenge in computer vision. Most of recent works focus on learning suitable feature representation that is robust to pose, illumination, and view angle changes to facilitate person re-ID. Among them, many works [21, 46, 47, 72, 69] seek to localize different body parts and align their associated features, while other works [8, 25, 28, 29, 49, 54, 62] use coarse partitions or attention selection network to improve feature learning.

This paper studies the optimization process in training a neural network for person re-ID. We present a simple yet surprisingly effective training mechanism, Batch Feature Erasing (BFE), which randomly crops away a block of the learned feature map for all images in the same batch. This is similar to the DropBlock in a concurrent work [13], which discards spatially correlated features to generalize drop-out to convolutional layers. Different from the DropBlock, we advocate for batch processing, where all images in the same batch are cropped in a consistent way. When the input images are roughly aligned, like the case of person re-ID, this batch operation is particularly useful for tasks like metric learning. Intuitively, it disables all features associated to a semantic part, e.g. head, leg, or bag etc, in all images in the same batch to force the network to learn a better representation of the remaining features. This idea is exemplified in Figure 1.



(a) (b)

Figure 1: (a) The input images in the same batch. (b) The corresponding feature response maps from the last output of the backbone network (ResNet-50 [14]), where the grayscale intensity visualizes the $l2$-norm of feature vectors. Our Batch Feature Erasing (BFE) will crop away a large block of the feature map in all the images, e.g. the red box. This disables features associated with the bag and abdomen and forces the network to optimize the feature representation of other regions.

This simple training strategy brings significant improvement on person re-ID. The ResNet-50 architecture [14] with hard triplet loss [15] and our Batch Feature Erasing (BFE) achieves 74.4% Rank-1 accuracy on CUHK03-Detect dataset, which is 7.6% higher than the state-of-the-art work [56].

This Batch Feature Erasing can also be adopted in different metric learning schemes, including triplet loss [38, 15], lifted structure loss [33], weighted sampling based margin loss [60], and histogram loss [52]. We test it with the image retrieval tasks on the CUB200-2011 [55], CARS196 [20], In Shop Clothes Retrieval dataset [30] and Stanford online products dataset [44]. The BFE can consistently improve the Rank-1 accuracy of various schemes.

2. Related work

Person re-ID is a challenging task in computer vision due to the large variation of pose, background, illumination, and camera conditions. Historically, people used hand-craft features for person re-identification [4, 9, 26, 27, 31, 32, 35, 36, 64, 75]. Recently, deep learning based methods dominate the Person re-ID benchmarks [5, 40, 48, 69, 71, 77].

The formulation of person re-ID has gradually evolved from a classification problem to a metric learning problem, which aims to find embedding features for input images in order to measure their semantic similarity. The work [74] compares both strategies on the Market-1501 dataset. Current works in metric learning generally focus on the design of loss functions, such as contrastive loss [53], triplet loss [8, 28], lifted structure loss [33], quadruplet loss [6], histogram loss [52], etc. In addition to loss functions, the hard sample mining methods, such as distance weighted sampling [60], hard triplet mining [15] and margin sample mining [61] are also critical to the final retrieval precision. Another work [67] also studies the application of mutual learning in metric learning tasks. In this paper, the proposed Batch Feature Erasing is a general training strategy that is effective in many metric learning formulations with different loss functions.

The human body is highly structured and distinguishing corresponding body parts can effectively determine the identity. Many recent works [28, 49, 51, 54, 56, 59, 65, 67, 68] aggregate salient features from different body parts and global cues for person re-ID. Among them, the part-based methods [8, 49, 56] achieve the state-of-the-art performance, which split a input feature map horizontally into a fixed number of strips and aggregate features from those strips. However, aggregating the feature vectors from multiple branches generally results in a complicated network structure. In comparison, our method involves only a simple network with two branches, one-third the size of the state-of-the-art MGN method [56].

To handle the imperfect bounding box detection and body part misalignment, many works [25, 40, 41, 42, 76] exploit the attention mechanisms to capture and focus on attentive regions. Saliency weighting [57, 70] is another effective approach to this problem. Inspired by attention models, Zhao et al. [69] propose part-aligned representations for person re-ID. Following the similar ideology, the works [18, 22, 23, 29] have also demonstrated superior performance, which incorporate a regional attention selection sub-network into the person re-ID model. To learn a feature representation robust to pose changes, the pose guided attention methods [21, 46, 72] fuse different body parts features with the help of pose estimation and human parsing network. However, such methods are prone to the possible noise from the pose estimation and semantic parsing algorithms.

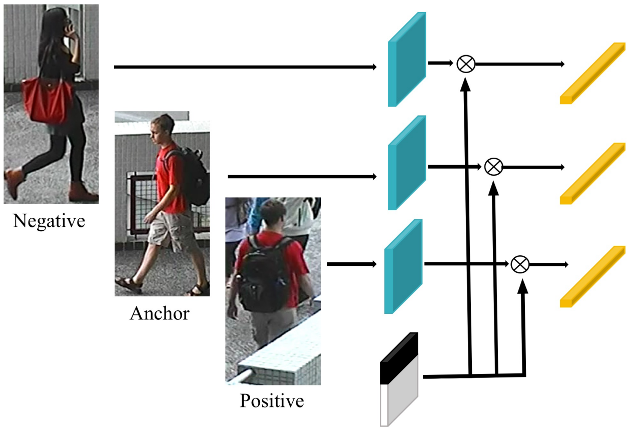


Figure 2: The Batch Feature Erasing Layer demonstrated on the triplet loss function [38].

To further improve the retrieval precision, re-ranking strategies [2, 80] and inference with specific person attributes [39] are adopted too. Recent works also introduce synthetic training data [3], adversarially occluded samples [17] and unlabeled samples generated by GAN [78] to remarkably augment the variant of input training dataset. The work in [12] transfers the representations learned from the general classification dataset to address the data sparsity of the person re-ID problems. Notably, such policies above also can be used jointly with our method.

3. Problem Formulation

Person re-identification is often formulated as an image retrieval problem, which aims to find the most similar image from a large set of candidate images. This retrieval problem can be solved by finding an embedding function $f(\cdot)$ and a metric function $D(\cdot,\cdot)$, where the embedding maps an input image x to a high dimensional feature vector $f(x)$ and the metric $D(f(x), f(y))$ measures the similarity between two embedding vectors $f(x)$ and $f(y)$.

In this paper, we simply take the Euclidean distance as the metric $D(\cdot,\cdot)$ and seek to train a neural network to learn the embedding $f(\cdot)$. We formulate the general image retrieval problem as the following. Given a dataset $\mathcal{X}$ with C classes, we split it into the training set $\mathcal{X}_{train}$ and testing set $\mathcal{X}_{test}$ with C_{train} and C_{test} classes respectively. Notice that there is no overlap between C_{train} and C_{test}. We train the network to learn the embedding $f(\cdot)$ with appropriate

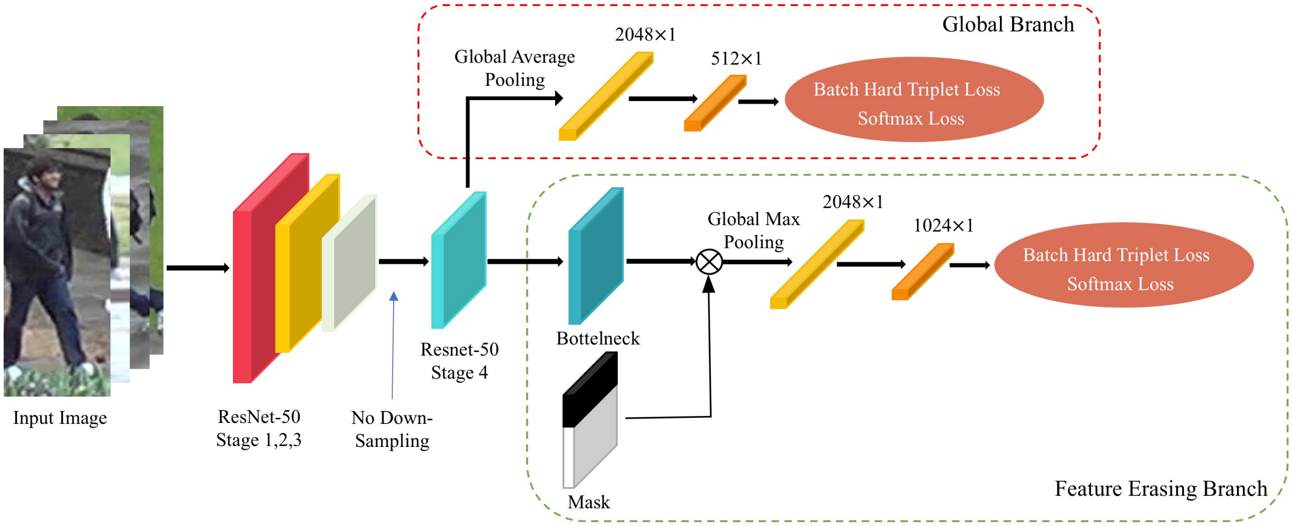


Figure 3: The structure of our Batch Feature Erasing (BFE) network with the batch hard triplet loss [15] demonstrated on the person re-ID problem. The global branch is appended after ResNet-50 Stage 4 and the feature erasing branch introduces a mask to crop out a large block in the bottleneck feature map. During testing, the features from both the global branch and feature erasing branch are concatenated as the final descriptor of a pedestrian image.

loss functions from image batches sampled in $\mathcal{X}_{train}$. During testing, we first compute the feature embedding for all images in $\mathcal{X}_{test}$ to build a database of embedding vectors $\mathcal{F}_{test}$. Then for each test image q in $\mathcal{X}_{test}$, we compute its embedding f_q and search the K nearest neighbors (KNN) in the embedding database $\mathcal{F}_{test}$. The embedding of the query image itself is excluded from these K nearest neighbors. For each returned KNN result, we consider it is correct if its class is the same as the query image's class.

Person re-ID can be regarded as image retrieval where each person corresponds to a class. In addition, each person image has a camera ID indicating the camera it is captured from. In person re-ID, challenging query images are selected for each person in the testing set $\mathcal{X}_{test}$ to form the query dataset $\mathcal{X}_{query}$. The gallery dataset $\mathcal{X}_{gallery}$ consists of the remaining testing images. In other words, a person re-ID dataset $\mathcal{X}$ is split to $\mathcal{X}_{train}$, $\mathcal{X}_{query}$, and $\mathcal{X}_{gallery}$. The training set $\mathcal{X}_{train}$ contains C_{train} identities, while $\mathcal{X}_{query}$ and $\mathcal{X}_{gallery}$ contain the same C_{test} identities. The training process of our person re-ID is exactly the same as general image retrieval. During testing, we compute the embedding database $\mathcal{F}_{gallery}$ for gallery dataset and query KNN results for each query image in $\mathcal{X}_{query}$. To make the problem more challenging, we only consider a result is correct when it has the same person identity and different camera ID from the query image.

4. Batch Feature Erasing (BFE) Network

This section describes how the Batch Feature Erasing (BFE) is adopted in a neural network for person re-ID and image retrieval.

Backbone Network. We use the ResNet-50 [14] as the backbone network for feature extraction as many of the person re-ID networks. Following the recent works [49, 56], we also modify the backbone ResNet-50 slightly, in which the down-sampling operation at the beginning of stage 4 is not employed. In this way, we get a larger feature map of size $2048 \times 24 \times 8$.

ResNet-50 Baseline. On top of this backbone network, we append a branch denoted as **global branch**. Specifically, after stage 4 of ResNet-50, we employ global average pooling to get a 2048-dimensional feature vector, the dimension of which is further reduced to 512 for both the triplet loss and softmax loss through a 1×1 convolution layer, a batch normalization layer, and a ReLU layer. We denote the backbone network together with the global branch as **ResNet-50 Baseline** in the following sections. Our baseline achieves 90.4% Rank-1 accuracy and 75.5% mAP on the Market-1501 dataset, which outperforms IDE [74] by a large margin in both Rank-1 (+17.9%) accuracy and mAP (+29.5%).

Batch Feature Erasing Layer. Given the feature tensor T computed by backbone network from a single batch of input images, the BFE Layer randomly erases the same region of tensor T. All the units inside the erased area are zeroed out. The detailed algorithm is shown in Algorithm 1. We also visualize the application of BFE Layer in the the triplet loss function in Figure 2, while it can be adopted in other loss functions [33, 52, 60] as well. The height and width of the erased region varies from task to task. But in general,

	CUHK03-Label		CUHK03-Detect		DukeMTMC-reID		Market1501	
Method	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
IDE [74]	22.2	21.0	21.3	19.7	67.7	47.1	72.5	46.0
PAN [79]	36.9	35.0	36.3	34.0	71.6	51.5	82.8	63.4
SVDNet [48]	-	-	41.5	37.3	76.7	56.8	82.3	62.1
DPFL [7]	43.0	40.5	40.7	37.0	79.2	60.0	88.9	73.1
HA-CNN [25]	44.4	41.0	41.7	38.6	80.5	63.8	91.2	75.7
SVDNet+Era [81]	49.4	45.0	48.7	37.2	79.3	62.4	87.1	71.3
TriNet+Era [81]	58.1	53.8	55.5	50.7	73.0	56.6	83.9	68.7
DaRe [58]	66.1	61.6	63.3	59.0	80.2	64.5	89.0	76.0
GP-reid [1]	-	-	-	-	85.2	72.8	92.2	81.2
PCB [49]	-	-	61.3	54.2	81.9	65.3	92.4	77.3
PCB + RPP [49]	-	-	62.8	56.7	83.3	69.2	93.8	81.6
MGN [56]	68.0	67.4	66.8	66.0	**88.7**	**78.4**	**95.7**	**86.9**
BFE$^{256+512}$	**75.4**	**71.2**	**74.4**	**70.8**	86.8	71.5	93.5	82.8
BFE	75.0	70.9	72.1	67.9	**88.7**	75.8	94.4	85.0

Table 1: The comparison with the existing person re-ID methods on CUHK03, DukeMTMC-reID and Market1501 datasets. BFE$^{256+512}$ means the feature dimension of global branch is 256 and the feature dimension of feature erasing branch is 512.

the erased region should be big enough to cover a semantic part of input feature map.

```
Input: $T$, input tensor of size B × C × H × W
       $r_h$, ratio of erased height
       $r_w$, ratio of erased width
Result: $T'$, erased tensor
if training then
    $H_e \leftarrow r_h \times H, W_e \leftarrow r_w \times W$
    $x_e \leftarrow Rand(0, H - H_e)$
    $y_e \leftarrow Rand(0, W - W_e)$
    $M \leftarrow Ones(H, W)$
    $M[x_e : x_e + H_e, y_e : y_e + W_e] \leftarrow 0$
    $T' \leftarrow T \times M$
else
    $T' \leftarrow T$
end
```

Algorithm 1: The Batch Feature Erasing Layer

Structure of the BFE Network. Our BFE Network consists of a global branch and a feature erasing branch as illustrated in Figure 3. The global branch is introduced for two purposes. First, it provides global feature representations. Second, it supervises the training for the Feature Erasing Branch. The **feature erasing branch** then applies the Batch Feature Erasing Layer on feature map T and provides the batch erased feature map T'. Afterwards, we apply global max pooling to get the 2048-dimensional feature vector. Finally, the dimension of a feature vector is reduced from 2048 to 1024 for both triplet and softmax losses. Dropout [45] is not used in our network.

The ResNet bottleneck block [14] which applies a stack of convolution layers on feature map T is critical. Without it, the global average pooling layer and the global max pooling layer would be applied simultaneously on T, making the network hard to converge. According to our observation, average pooling is unstable after the Feature Erasing Layer, we therefore use global max pooling on the feature erasing branch instead of the average pooling.

Then, features from the global branch and the feature erasing branch are concatenated as the embedding vector of an pedestrian image. Here, the following three points are worth noting. 1) The BFE Layer is parameter free and will not increase the network size. 2) The BFE Layer can be easily adopted in other metric learning tasks beyond person re-ID. 3) The BFE hyper-parameters are tunable without changing network structure for different tasks.

Loss function. The loss function is the sum of soft margin batch-hard triplet loss [15] and softmax loss on both the global branch and feature erasing branch. Specifically, the soft margin batch-hard triplet loss l_{SBH} is defined as

$$l_{SBH}(X) = \sum_{i=1}^{P} \sum_{a=1}^{K} \log(1 + \exp(l_{BH}(x_a^i))), \qquad (1)$$

$$l_{BH}(x_a^i) = \underbrace{\max_{p=1 \dots K} D(f_\theta(x_a^i), f_\theta(x_p^i))}_{\text{hardest positive}} - \underbrace{\min_{\substack{j=1 \dots P \\ n=1 \dots K \\ n \neq j}} D(f_\theta(x_a^i), f_\theta(x_n^j))}_{\text{hardest negative}}, \qquad (2)$$

where P is the number of distinct persons and K is the number of images for each person so there are $P \times K$ triplets in a batch. $l_{BH}(\cdot)$ denotes the batch-hard triplet loss. For each anchor image x_a^i, we select the image from the same identity i with the maximum distance as the positive image x_p^i, and select the image from a different identity j with the minimum distance as the negative image x_n^j. Therefore, x_a^i, x_p^i, and x_a^j form a triplet. $l_{SBH}(\cdot)$ is the sum of the soft margin batch-hard triplet loss for all the triplets in a batch. $D(\cdot, \cdot)$ denotes the Euclidean distance function and f_θ is the feature embedding learned by the BFE Network.

Method	Rank-1	mAP
Triplet	90.3	77.0
Softmax	89.4	76.4
Triplet + Softmax	94.4	85.0

Table 2: The effect of the joint training method (Triplet + Softmax) on Rank-1 accuracy (%) and mAP (%). The statistics are collected from the Market-1501 dataset.

Method	Rank-1	mAP
Global Branch (Baseline)	90.4	75.5
Feature Erasing Branch	88.7	71.9
Both Branches	94.4	85.0

Table 3: The effect of global branch and feature erasing branch on Rank-1 accuracy (%) and mAP (%). We collect the statistics from the Market-1501 dataset.

5. Experiments

We verify our BFE Network on the benchmark person re-ID data-sets. BFE Network with different metric learning loss functions is also tested on the standard image retrieval datasets.

5.1. Person re-ID Experiments

5.1.1 Datasets and Settings

We test three generally used person re-ID datasets including Market-1501 [73], DukeMTMC-reID [37, 78], and CUHK03 [24] datasets. We also follow the same strategy used in recent works [15, 49, 56] to generate training, query, and gallery data. Notice that the original CUHK03 dataset is divided into 20 random training/testing splits for cross validation which is commonly used in hand-craft feature based methods. The new partition method adopted in our experiments further splits the training and gallery images, and selects challenging query images for evaluation. Therefore, CUHK03 dataset becomes the most challenging dataset among the three.

During training, the input images are re-sized to 384 × 128 and then augmented by random horizontal flip, normalization, and random erasing [81]. In BFE Layer, we set the erased height ratio r_h to 0.5 and erased width ratio r_w to 1.0. The same setting is used in all the person re-ID datasets. The testing images are re-sized to 384 × 128 and only augmented with normalization.

For each query image, we rank all the gallery images in decreasing order of their Euclidean distances to the query images and compute the Cumulative Matching Characteristic (CMC) curve. We use Rank-1 accuracy and mean average precision (mAP) as the evaluation metrics. Results with the same identity and the same camera ID as the the query image are not counted. It is worth noting that all the experiments are conducted in a single-query setting without re-ranking[2, 80] for simplicity.

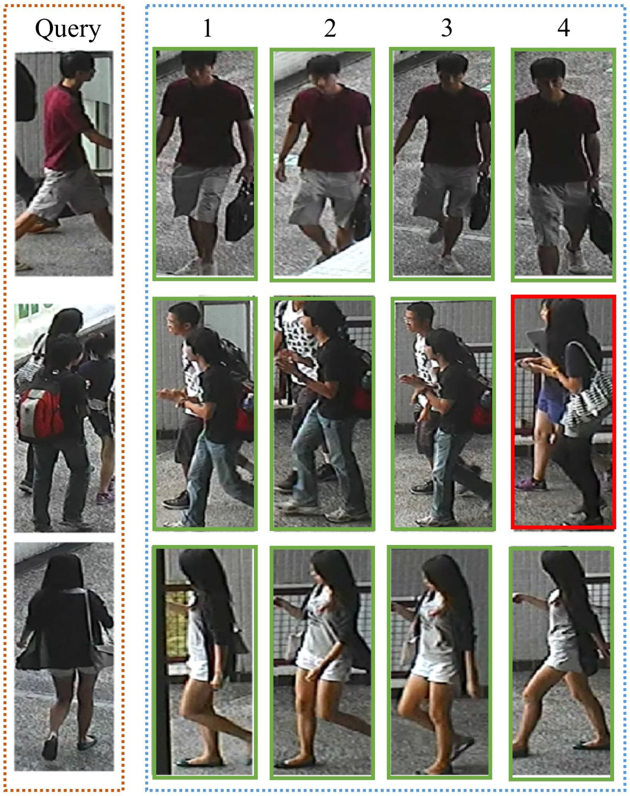


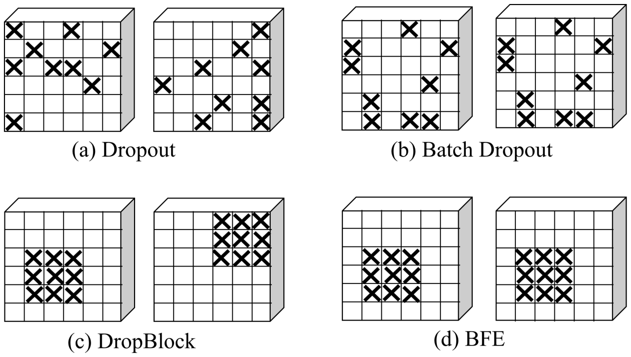
Figure 4: The top-4 ranking list for the query images on CUHK03-Label dataset from the proposed BFE Network. The correct results are highlighted by green borders and the incorrect results by red borders.

5.1.2 Training

Our network is trained using 4 GTX1080 GPUs with a batch size of 128. Each identity contains 4 instance images in a batch, so there are 32 identities per batch. The backbone ResNet-50 is initialized from the ImageNet [10] pre-trained model. We use the batch hard soft margin triplet loss [15] to avoid margin parameters. We use the Adam optimizer[19] with the base learning rate initialized to 1e-3, then decayed to 1e-4 after 100 epochs, and further decayed to 1e-5 after 300 epochs. The whole training procedure has 600 epochs and takes approximately 2 hours.

5.1.3 Comparison with State-of-the-Art

The statistical comparison between our BFE Network and the state-of-the-art methods on CUHK03, DukeMTMC-reID and Market-1501 datasets is shown in Table 1. We


(a) Dropout

(b) Batch Dropout

(c) DropBlock

(d) BFE

Figure 5: The comparison with Dropout methods on two feature maps within the same batch.

Method	Rank-1	mAP
Dropout [45]	72.1	67.1
SpatialDropout[50]	72.7	68.5
Batch Dropout	71.7	67.1
DropBlock [13]	60.4	56.8
BFE	74.4	70.8

Table 4: The Comparison with other Dropout methods on the CUHK03-Detect dataset.

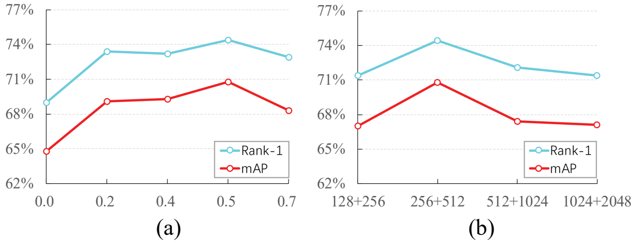
also show the results of BFE$^{256+512}$, which generates a more compact feature embedding. It shows that our method achieves state-of-the-art performance on both CUHK03 and DukeMTMC-reID datasets and comparative performance on Market-1501 dataset. Remarkably, our method achieves the largest improvement (i.e., 7.6% in Rank-1 accuracy) over previous methods on CUHK03-Detect dataset, which is the most challenging dataset. For DukeMTMC-reID and Market1501 datasets, our model only achieves comparative performance to MGN [56]. However, it is worth to point out that MGN benefits from a much lager and more complex network which generates 8 feature vectors with 8 branches supervised by 11 loss functions. The model size (i.e., number of parameters) of MGN is three times of BFE Network.

Some sample query results are illustrated in Figure 4. We can see that, given a back view person image, BFE can even retrieve the front view and side view images of the same person.

5.1.4 Ablation Studies

We perform extensive experiments on Market-1501 and CUHK03 datasets to analyze the effectiveness of each component and impact of hyper parameters in our method.

Benefit of Triplet and Softmax Joint Training. The BFE Network is trained using both triplet loss and softmax loss. Table 2 shows the performance on Market-1501 dataset given different combinations of losses. We can see that the joint training method significantly improves the performance measured by Rank-1 accuracy and mAP.



(a)

(b)

Figure 6: (a) The effects of erased height ratio on mAP and CMC scores. The erased width ratio is fixed to 1.0. (b) The effects of feature dimensions on mAP and CMC scores. The statistics are analyzed on the CUHK03-Detect dataset.

Benefit of Global and Feature Erasing Branches. Without the global branch, the BFE Network performs a little bit worse than the baseline network as illustrated in Table 3. In this case, the BFE Layer plays the role as a regularization method, similar to DropBlock [13]. The motivation behind the two-branch structure in the BFE Network is that it learns both the most salient appearance clues from the global branch and fine-grained discriminative features from the feature erasing branch. The two branches reinforce each other and are both important to the final performance.

Comparison with Dropout Strategies. Dropout [45] drops values of input tensor randomly and is a widely used regularization technique to prevent overfitting. We replace the BFE Layer with various Dropout methods and compare their performance in Table 4. SpatialDropout [50] randomly zeroes whole channels of the input tensor. The channels to zero-out are randomized on every forward call. Here, Batch Dropout means we select random spatial positions and drops all input features in these locations. The difference between BFE and Batch Dropout is that BFE zeroes a large contiguous area while Batch Dropout zeroes some isolated features. DropBlock [13] means for a batch of input tensor, every tensor randomly drops a contiguous region. The difference between BFE and DropBlock is that BFE drops the same region for every input tensor within a batch while DropBlock crops out different regions. These Dropout methods are visualized in Figure 5. As shown in Table 4, BFE is more effective than these various Dropout strategies in the person re-ID tasks.

Impact of BFE Layer Hyper-parameters. Figure 6 (a) studies the impact of erased height ratio on the performance of the BFE Network. Here, the erased width ratio is fixed to 1.0 in all the person Re-ID experiments. We can see that as the erased height ratio ranges from 0.2 to 0.7, the performance measured by Rank-1 accuracy and mAP is obvi-

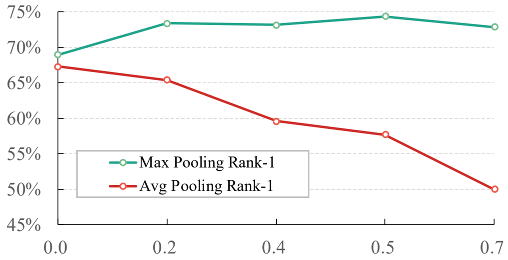


Figure 7: The comparison of global average pooling and global max pooling on the feature erasing branch under different height ratio settings. The statistics are analyzed on the CUHK03-Detect dataset.

ously superior to that of the non-erasing network (erased ratio equals to 0.0). It also suggests that the BFE Layer is robust to various erased ratio settings.

Impact of Embedding Dimensions. A feature descriptor with a higher dimension is more expressive but may suffer from overfitting during evaluation. Typically, BFE Network uses 512+1024 dimensional embedding where 512-dimension for the feature vector from the global branch and 1024-dimension for the feature vector from the feature erasing branch. In Figure 6 (b), we compare the performance of different feature dimensions, such as 128+256 dimension, 256+512 dimension, 512+1024 dimension, and 1024+2048 dimension. Table 1 and Figure 6 show that the optimal dimension differs in datasets. CUHK03 prefers lower dimension since it suffers from overfitting while other datasets prefer a higher dimension. In general, 512+1024 dimension is a reasonable setting.

Average Pooling vs Max Pooling in Feature Erasing Branch. The BFE Network uses global average pooling on the global branch, the same as the original ResNet-50 network [14]. In Figure 7, we compare the performance of feature erasing branch with different pooling methods. We find that the Rank-1 accuracy of the feature erasing branch with Max Pooling first increases and then gradually becomes stable. It is also consistently superior to that with Average Pooling. The empirical evidence therefore demonstrates the importance of Max Pooling to a robust convergence and increased performance on the feature erasing branch.

5.2. Image Retrieval Experiments

The BFE Network structure can be applied directly on image retrieval problems.

5.2.1 Datasets and Settings

Our method is evaluated on the commonly used image retrieval datasets including CUB200-2011 [55], CARS196 [20], Stanford online products (SOP) [33], and In-Shop Clothes retrieval [30] datasets. For CUB200-2011 and CARS196, the cropped datasets are used since our BFE Network requires input images to be roughly aligned. The experimental setup is the same as that in [33]. We show the statistics of the four image retrieval datasets in Table 5.

Dataset	CARS	CUB	SOP	Clothes
# images	16,185	11,788	120,053	52,712
# classes	196	200	22,634	11,735
# training class	98	100	11,318	3,997
# training image	8,054	5,864	59,551	25,882
# testing class	98	100	11,316	3,985
# testing image	8,131	5,924	60,502	26,830

Table 5: The statistics of the image retrieval datastes including CARS196 [20], CUB200-2011 [55], Stanford online products(SOP) [33], and In-Shop Clothes retrieval dataset [30]. Notice that the test set of In-Shop Clothes retrieval dataset is further split to query dataset with 14,218 images and gallery dataset with 12,612 images.

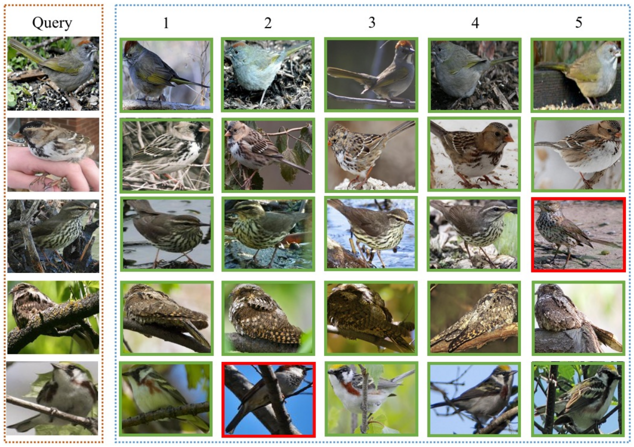


Figure 8: The top-5 ranking list for the query images on CUB200-2011 data-set from our BFE Network. The green and red borders respectively denote the correct and incorrect results.

The training images are padded and resized to 256 × 256 while the aspect ratio is fixed, and then cropped to 224 × 224 randomly. During testing, CUB200-2011, In-Shop Clothes retrieval dataset, and SOP images are padded on the shorter side and then scaled to 256 × 256, while CARS196 images are scaled to 256 × 256 directly. The erasing height ratio and width ratio are both set to 0.5 in the BFE Layer. We use the standard Recall@K metric to measure the image retrieval performance.

K	1	2	4	8
PDDM Triplet [16]	50.9	62.1	73.2	82.5
PDDM Quadruplet [16]	58.3	69.2	79.0	88.4
HDC [66]	60.7	72.4	81.9	89.2
Margin [60]	63.9	75.3	84.4	90.6
ABE-8 [18]	70.6	79.8	86.9	92.2
BFE	**74.1**	**83.6**	**89.8**	**93.6**

(a) CUB200-2011 (cropped) dataset

K	1	2	4	8
PDDM Triplet [16]	46.4	58.2	70.3	80.1
PDDM Quadruplet [16]	57.4	68.6	80.1	89.4
HDC [66]	83.8	89.8	93.6	96.2
Margin [60]	86.9	92.7	95.6	97.6
ABE-8 [18]	93.0	95.9	97.5	98.5
BFE	**94.3**	**96.8**	**98.3**	**98.9**

(b) CARS196 (cropped) dataset

K	1	10	20	30	40
FasionNet [30]	53.0	73.0	76.0	77.0	79.0
HDC [66]	62.1	84.9	89.0	91.2	92.3
DREML [63]	78.4	93.7	95.8	96.7	-
HTL [11]	80.9	94.3	95.8	97.2	97.4
A-BIER [34]	83.1	95.1	96.9	97.5	97.8
ABE-8 [18]	87.3	**96.7**	**97.9**	98.2	98.5
BFE	**89.1**	96.3	97.6	**98.5**	**99.1**

(c) In-Shop Clothes Retrieval dataset

K	1	10	100	1000
LiftedStruct [33]	62.1	79.8	91.3	97.4
N-Pairs [43]	67.7	83.8	93.0	97.8
Margin [60]	72.7	86.2	93.8	98.0
HDC [66]	69.5	84.4	92.8	97.7
A-BIER [34]	74.2	86.9	94.0	97.8
ABE-8 [18]	76.3	88.4	94.8	98.2
BFE	**83.0**	**93.3**	**97.3**	**99.2**

(d) Stanford online products dataset

Table 6: The comparison on Recall@$K(\%)$ scores with other state-of-the-art metric learning methods on CUB200-2011 (cropped), CARS196 (cropped), In-Shop Clothes Retrieval, and Stanford online products datasets.

K	1	5	10	20
Baseline + LiftedStruct [33]	66.8	88.5	93.4	96.3
BFE + LiftedStruct [33]	71.4	89.7	93.9	96.3
Baseline + Margin [60]	65.7	88.1	93.1	96.4
BFE + Margin [60]	72.0	90.8	94.4	97.0
Baseline + Histogram [52]	64.6	87.2	93.0	96.4
BFE + Histogram [52]	73.1	90.7	94.2	96.9
Baseline + Hard Triplet [15]	69.5	89.5	94.0	96.8
BFE + Hard Triplet [15]	**74.1**	**91.0**	**94.7**	**97.1**

Table 7: The BFE network performance on the other standard loss functions of metric learning methods. The statistics are based on the CUB200-2011 (cropped) dataset. "Baseline" refers to the ResNet-50 Baseline defined in section 4.

5.2.2 Comparison with State-of-the-Art

Table 6 shows that our BFE Network achieves the best Recall@1 scores on all the experimental image retrieval datasets. In particular, the BFE Network achieves an obvious improvement (+3.5%) on the small scale CUB200-2011 dataset which is also the most challenging one. On the large scale Stanford online products dataset which contains $22,634$ classes with $120,053$ product images, our BFE network surpasses the state-of-the-art by 6.7%. We can see that our BFE Network is applicable on both small and large scale datasets.

Figure 8 visualizes sample retrieval results of CUB200-2011 (cropped) dataset. We can see that our BFE Network is in some terms robust to the variance in illumination, poses and occlusions.

5.2.3 Adapt to Other Metric Learning Methods

Table 7 shows that our BFE Network can also be used with other standard metric learning loss functions, such as lifted structure loss[33], weighted sampling margin loss[60], and histogram loss[52] to boost their performance. For a fair comparison, we re-implement the above loss functions on our ResNet-50 Baseline and BFE Network to evaluate their performances. Here, the only difference between ResNet-50 Baseline and BFE Network is that the BFE Network has an additional feature erasing branch. Although the ResNet-50 Baseline outperforms the results reported in the work [60] (+1.8%), the BFE Network can still improve the result by a large margin (+7.7%). We can therefore conclude that the proposed BFE Network can be easily generalized to other standard loss functions in metric learning.

6. Conclusion

In this paper, we propose the Batch Feature Erasing (BFE) to improve the optimization in training a neural network for person re-ID and other general metric learning tasks. The corresponding BFE Network, which adopts this proposed training mechanism, leverages a global branch to embed salient representations and a feature erasing branch to learn detailed features. Extensive experiments on both person re-ID datasets and image retrieval datasets show that the BFE Network can make significant improvement on person re-ID and other general image retrieval benchmarks.

References

[1] J. Almazan, B. Gajic, N. Murray, and D. Larlus. Re-id done right: towards good practices for person re-identification. *arXiv:1801.05339*, 2018. 4

[2] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017. 2, 5

[3] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. In *CVIU*, 2018. 2

[4] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by HPE signature. In *ICCV*, 2010. 2

[5] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang. Group consistent similarity learning via deep CRF for person re-identification. In *CVPR*, 2018. 2

[6] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017. 2

[7] Y. Chen, X. Zhu, S. Gong, et al. Person re-identification by deep learning multi-scale representations. In *ICCV*, 2018. 4

[8] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 1, 2

[9] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *ECCV*, 2014. 2

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[11] W. Ge, W. Huang, D. Dong, and M. R. Scott. Deep metric learning with hierarchical triplet loss. In *ECCV*, 2018. 8

[12] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. In *BigMM*, 2018. 2

[13] G. Ghiasi, T.-Y. Lin, and Q. V. Le. Dropblock: A regularization method for convolutional networks. *arXiv:1810.12890*, 2018. 1, 6

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 3, 4, 7

[15] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017. 1, 2, 3, 4, 5, 8

[16] C. Huang, C. C. Loy, and X. Tang. Local similarity-aware deep feature embedding. In *NIPS*, 2016. 8

[17] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang. Adversarially occluded samples for person re-identification. In *CVPR*, 2018. 2

[18] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon. Attention-based ensemble for deep metric learning. In *ECCV*, 2018. 2, 8

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[20] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV*, 2013. 2, 7

[21] V. Kumar, A. M. Namboodiri, M. Paluri, and C. Jawahar. Pose-aware person recognition. In *CVPR*, 2017. 1, 2

[22] X. Lan, H. Wang, S. Gong, and X. Zhu. Deep reinforcement learning attention selection for person re-identification. In *BMVC*, 2017. 2

[23] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 2

[24] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 5

[25] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 1, 2, 4

[26] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 2

[27] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2

[28] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 2017. 1, 2

[29] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 1, 2

[30] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2, 7, 8

[31] A. J. Ma, P. C. Yuen, and J. Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *ICCV*, 2013. 2

[32] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 2

[33] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 2, 3, 7, 8

[34] M. Opitz, G. Waltner, H. Possegger, and H. Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *arXiv:1801.04815*, 2018. 8

[35] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 2

[36] A. Perina, V. Murino, M. Cristani, M. Farenzena, and L. Bazzani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[37] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 5

[38] 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*, pages 815–823, 2015. 1, 2

[39] A. Schumann and R. Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPRW*, 2017. 2

[40] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, 2018. 2

[41] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015. 2

[42] 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 *CVPR*, 2018. 2

[43] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. 8

[44] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy. Deep metric learning via facility location. In *CVPR*, 2017. 2

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014. 4, 6

[46] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017. 1, 2

[47] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 1

[48] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017. 2, 4

[49] 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 *ECCV*, 2018. 1, 2, 3, 4, 5

[50] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015. 6

[51] E. Ustinova, Y. Ganin, and V. Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *AVSS*, 2017. 2

[52] E. Ustinova and V. Lempitsky. Learning deep embeddings with histogram loss. In *NIPS*, 2016. 2, 3, 8

[53] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 2

[54] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 1, 2

[55] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 7

[56] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. *arXiv:1804.01438*, 2018. 1, 2, 3, 4, 5, 6

[57] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014. 2

[58] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018. 4

[59] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017. 2

[60] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl. Sampling matters in deep embedding learning. In *ICCV*, 2017. 2, 3, 8

[61] Q. Xiao, H. Luo, and C. Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv:1710.00478*, 2017. 2

[62] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017. 1

[63] H. Xuan, R. Souvenir, and R. Pless. Deep randomized ensembles for metric learning. In *ECCV*, 2018. 8

[64] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*, 2014. 2

[65] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv:1707.00798*, 2017. 2

[66] Y. Yuan, K. Yang, and C. Zhang. Hard-aware deeply cascaded embedding. In *ICCV*, 2017. 8

[67] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv:1711.08184*, 2017. 2

[68] 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 *CVPR*, 2017. 2

[69] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 1, 2

[70] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 2

[71] F. Zheng and L. Shao. Learning cross-view binary identities for fast person re-identification. In *IJCAI*, 2016. 2

[72] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv:1701.07732*, 2017. 1, 2

[73] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5

[74] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv:1610.02984*, 2016. 2, 3, 4

[75] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *PAMI*, 2013. 2

[76] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, 2015. 2

[77] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person reidentification. In *TOMM*, 2017. 2

[78] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, 2017. 2, 5

[79] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. In *TCSVT*, 2018. 4

[80] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 2, 5

[81] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv:1708.04896*, 2017. 4, 5