2019 IEEE/CVF International Conference on Computer Vision (ICCV)

Batch DropBlock Network for Person Re-identification and Beyond

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

Abstract

Since the person re-identification task often suffers from the problem of pose changes and occlusions, some attentive local features are often suppressed when training CNNs. In this paper, we propose the Batch DropBlock (BDB) Network which is a two branch network composed of a conventional ResNet-50 as the global branch and a feature dropping branch. The global branch encodes the global salient representations. Meanwhile, the feature dropping branch consists of an attentive feature learning module called Batch DropBlock, which randomly drops the same region of all input feature maps in a batch to reinforce the attentive feature learning of local regions. The network then concatenates features from both branches and provides a more comprehensive and spatially distributed feature representation. Albeit simple, our method achieves state-of-the-art on person re-identification and it is also applicable to general metric learning tasks. For instance, we achieve 76.4% Rank-1 accuracy on the CUHK03-Detect dataset and 83.0% Recall-1 score on the Stanford Online Products dataset, outperforming the exsiting works by a large margin (more than 6%).

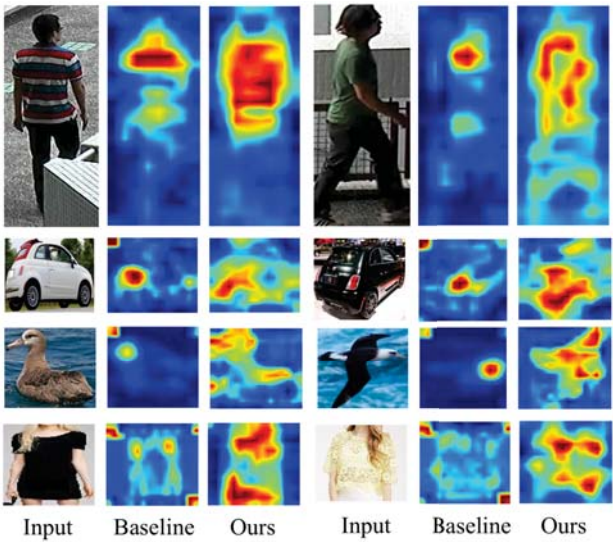

Figure 1: The class activation map on Baseline and BDB Network. Compared with the Baseline, the two-branch structure in BDB Network learns more comprehensive and spatially distributed features consisting of both global and attentive local representations.

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 using convolution neural networks. Because the body parts such as faces, hands and feet are unstable as the view angle changes, the CNN tends to focus on the main body part and the other descriminative body parts are consequently suppressed. To solve this problem, many pose-based works [23, 48, 49, 74, 71] seek to localize different body parts and align their associated features, and other part-based works [8, 27, 30, 31, 51, 56, 64] use coarse partitions or attention selection network to improve feature learning. However, such pose-based networks usually require additional body pose or segment information. Moreover, these networks are designed using specific partition mechanisms, such as a horizontal partition, which is fit for person re-ID but hard to be generalized to other metric learning tasks. The problems above motivate us to propose a simple and generalized network for person re-ID and other metric learning tasks.

In this paper, we propose the Batch DropBlock Network (BDB Network) for the roughly aligned metric learning tasks. The Batch DropBlock Network is a two-branch network consisting of a conventional global branch and a feature dropping branch where the Batch DropBlock, an attentive feature learning module, is applied. The global branch encodes the global feature representations and the feature dropping branch learns local detailed features. Specifically, Batch DropBlock randomly drops the same region of all the feature maps, namely the same semantic body parts, in a batch during training and reinforces the attentive feature learning of the remaining parts. Concatenating the features

2380-7504/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCV.2019.00379

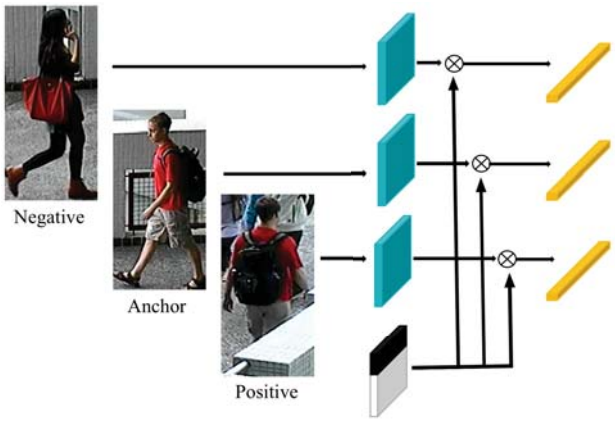

Figure 2: The Batch DropBlock Layer demonstrated on the triplet loss function [40].

of both branches brings a more comprehensive saliency representation rather than few discriminative features. In Figure 1, we use class activation map [84] to visualize the feature attention. We can see that the attention of baseline mainly focuses on the main body part while the BDB network learns more uniformly distributed representations.

Our Batch DropBlock is different from the general DropBlock [14] in two aspects. First, Batch DropBlock is an attentive feature learning module for metric learning tasks while DropBlock is a regularization method for classification tasks. Second, Batch DropBlock drops the same block for a batch of images during a single iteration, while DropBlock [14] erases randomly across different images. Here, 'Batch' means the group of images participating in a single loss calculation during training, for example, a pair for pairwise loss, a triplet for triplet loss and a quadruplet for quadruplet loss. If we erase features randomly as [14], for example, one image keeps head features and another image keeps feet features, the network can hardly find the semantic correspondence, not to mention reinforcing the learning of local attentive representations.

In the experimental section, the ResNet-50 [16] based Batch DropBlock Network with hard triplet loss [17] achieves 72.8% Rank-1 accuracy on CUHK03-Detect dataset, which is 6.0% higher than the state-of-the-art work [58]. Batch DropBlock can also be adopted in different metric learning schemes, including triplet loss [40, 17], lifted structure loss [35], weighted sampling based margin loss [62], and histogram loss [54]. We test it with the image retrieval tasks on the CUB200-2011 [57], CARS196 [22], In Shop Clothes Retrieval dataset [32] and Stanford online products dataset [46]. The BDB Network 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 poses, background, illumination, and camera conditions. Historically, people used hand-craft features for person re-identification [4, 9, 28, 29, 33, 34, 37, 38, 66, 77]. Recently, deep learning based methods dominate the Person re-ID benchmarks [5, 42, 50, 71, 73, 79].

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 [76] 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 [55], triplet loss [8, 30], lifted structure loss [35], quadruplet loss [6], histogram loss [54], etc. In addition to loss functions, the hard sample mining methods, such as distance weighted sampling [62], hard triplet mining [17] and margin sample mining [63] are also critical to the final retrieval precision. Another work [69] also studies the application of mutual learning in metric learning tasks. In this paper, the proposed two-branch BDB Network 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 [30, 51, 53, 56, 58, 61, 67, 69, 70] aggregate salient features from different body parts and global cues for person re-ID. Among them, the part-based methods [8, 51, 58] achieve the state-of-the-art performance, which split an 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 [58].

To handle the imperfect bounding box detection and body part misalignment, many works [27, 42, 43, 44, 78] exploit the attention mechanisms to capture and focus on attentive regions. Saliency weighting [59, 72] is another effective approach to this problem. Inspired by attention models, Zhao et al. [71] propose part-aligned representations for person re-ID. Following the similar ideology, the works [20, 24, 25, 31] 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 [23, 48, 74] fuse different body parts features with the help of pose estimation and human parsing network. However, such methods based on pose estimation and semantic parsing algorithms are only designed for person re-ID tasks while our approach can be applied to other general metric learning tasks.

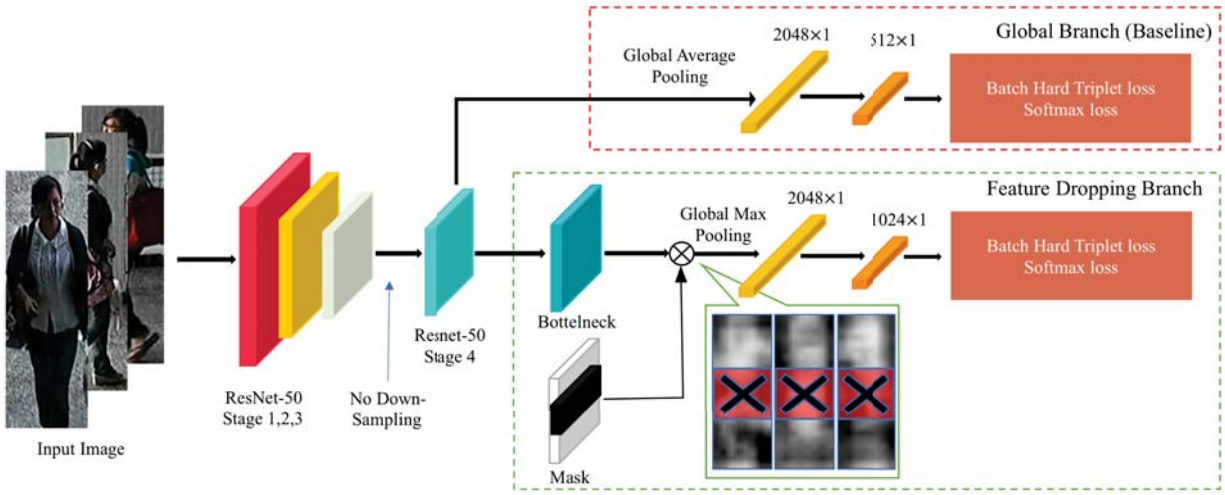

Figure 3: The structure of our Batch DropBlock (BDB) Network with the batch hard triplet loss [17] demonstrated on the person re-ID problem. The global branch is appended after ResNet-50 Stage 4 and the feature dropping branch introduces a mask to crop out a large block in the bottleneck feature map. During training, there are two loss functions for both global branch and feature dropping branch. During testing, the features from both branches are concatenated as the final descriptor of a pedestrian image.

To further improve the retrieval precision, re-ranking strategies [2, 82] and inference with specific person attributes [41] are adopted too. Recent works also introduce synthetic training data [3], adversarially occluded samples [19] and unlabeled samples generated by GAN [80] to remarkably augment the variant of input training dataset. The work in [13] transfers the representations learned from the general classification dataset to address the data sparsity of the person re-ID problems. Some general data augmentation methods such as Random Erasing [82] and Cutout [11] are also generally used. Notably, such policies above can be used jointly with our method.

3. Batch DropBlock (BDB) Network

This section describes the structure and components of the proposed Batch DropBlock Network.

Backbone Network. We use the ResNet-50 [16] as the backbone network for feature extraction as many of the person re-ID networks. For a fair comparison with the recent works [51, 58], 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 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. The performance of Baseline with or without triplet loss on person re-ID datasets are shown in table 1. Our baseline without triplet loss is identical to the baseline used in recent works [51, 58].

Batch DropBlock Layer. Given the feature tensor T computed by backbone network from a single batch of input images, the Batch DropBlock Layer randomly drops the same region of tensor T. All the units inside the dropping area are zeroed out. We visualize the application of Batch DropBlock Layer in the triplet loss function in Figure 2, while it can be adopted in other loss functions [35, 54, 62] as well. The height and width of the erased region varies from task to task. But in general, the dropping region should be big enough to cover a semantic part of input feature map. Unlike DropBlock [14], there is no need to change the keep probability hyper-parameter during training in Batch DropBlock Layer.

Network Architecture. As illustrated in Figure 3, our BDB Network consists of a global branch and a feature dropping branch.

The global branch is commonly used for providing global feature representations in multi-branch network architectures [8, 51, 58]. It also supervises the training for the feature dropping branch and makes the Batch DropBlock layer applied on a well-learned feature map. To demonstrate it, we visualize in Figure 4 the class activation map of the dropping branch trained with and without the global branch. We can see that the features learned by the dropping branch alone are more spatially dispersed with redundant background noise (e.g. at the bottom of Figure 4 (c)). As

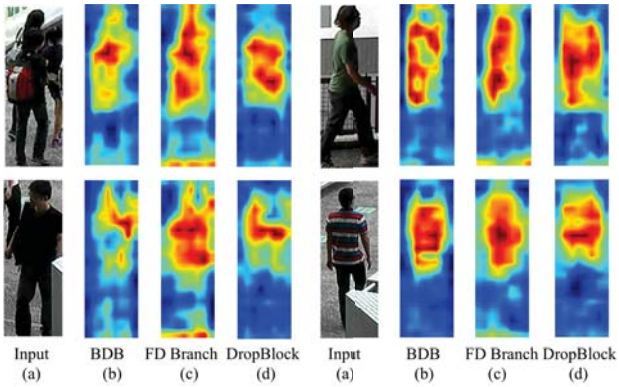

Figure 4: The class activation map of the BDB Network, the feature dropping branch when training alone, and when DropBlock is used in our network. 'FD Branch' means feature dropping branch.

mentioned in [14], dropping a large area randomly on input feature maps may hurt the network learning at the beginning. It therefore uses a scheduled training method which sets the dropping area small initially and gradually increases it to stabilize the training process. In BDB network, we do not need to change the dropping area with the intermediate supervision of the global branch. At the beginning stage of training, when the feature dropping branch could not learn well, the global branch helps the training.

The **feature dropping branch** then applies the Batch DropBlock 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. The purpose of the feature dropping branch is to learn multiple attentive feature regions instead of only focusing on the major discriminative region. Figure 4 also visualizes the class activation map of feature dropping branch with DropBlock or Batch DropBlock. One can see the features learned by DropBlock miss some attentive part features (e.g. legs in Figure 4 (d)) and the salient representations from Batch DropBlock have more accurate and clearer contours. An intuitive explanation is that, by blocking the same roughly aligned regions, we reinforce the attentive feature learning of the rest parts with semantic correspondences.

The BDB Network uses global average pooling (GAP) on the global branch, the same as the original ResNet-50 network [16]. Notably, we use global max pooling (GMP) in feature dropping branch, because GMP encourages the network to identify comparatively weak salient features after the most descriminative part is dropped. The strong feature is easy to be selected while the weak feature is hard to be distinguished from other low values. When the strong feature is dropped, GMP could encourage the network to strength the weak features. For GAP, low values except the weak features would still impact the results.

Also noteworthy is the ResNet bottleneck block [16] which applies a stack of convolution layers on feature map T. 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.

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

Loss function. The loss function is the sum of soft margin batch-hard triplet loss [17] and softmax loss on both the global branch and feature dropping branch.

4. Experiments

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

4.1. Person re-ID Experiments

4.1.1 Datasets and Settings

We test three generally used person re-ID datasets including Market-1501 [75], DukeMTMC-reID [39, 80], and CUHK03 [26] datasets. We also follow the same strategy used in recent works [17, 51, 58] 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 and normalization. In Batch DropBlock layer, we set the erased height ratio r_h to 0.3 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

Method	CUHK03-Label		CUHK03-Detect		DukeMTMC-reID		Market1501	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
IDE [76]	22.2	21.0	21.3	19.7	67.7	47.1	72.5	46.0
PAN [81]	36.9	35.0	36.3	34.0	71.6	51.5	82.8	63.4
SVDNet [50]	-	-	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 [27]	44.4	41.0	41.7	38.6	80.5	63.8	91.2	75.7
SVDNet+Era [83]	49.4	45.0	48.7	37.2	79.3	62.4	87.1	71.3
TriNet+Era [83]	58.1	53.8	55.5	50.7	73.0	56.6	83.9	68.7
DaRe [60]	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 [51]	-	-	61.3	54.2	81.9	65.3	92.4	77.3
PCB + RPP [51]	-	-	62.8	56.7	83.3	69.2	93.8	81.6
MGN [58]	68.0	67.4	66.8	66.0	88.7	**78.4**	**95.7**	**86.9**
Baseline	52.6	49.9	51.1	47.9	81.0	62.8	91.6	77.1
Baseline+Triplet	67.4	61.5	63.6	60.0	83.8	68.5	93.1	80.6
BDB	73.6	71.7	72.8	69.3	86.8	72.1	94.2	84.3
BDB+Cut	**79.4**	**76.7**	**76.4**	**73.5**	**89.0**	76.0	95.3	86.7

Table 1: The comparison with the existing person re-ID methods. 'Era' means Random Erasing [83]. 'Cut' means Cutout [11].

experiments are conducted in a single-query setting without re-ranking[2, 82] for simplicity.

4.1.2 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 [17] to avoid margin parameters. We use the Adam optimizer [21] with the base learning rate initialized to 1e-3 with a linear warm-up [15] in first 50 epochs, then decayed to 1e-4 after 200 epochs, and further decayed to 1e-5 after 300 epochs. The whole training procedure has 400 epochs and takes approximately 1.5 hours.

4.1.3 Comparison with State-of-the-Art

The statistical comparison between our BDB Network and the state-of-the-art methods on CUHK03, DukeMTMC-reID and Market-1501 datasets is shown in Table 1. It shows that our method achieves state-of-the-art performance on both CUHK03 and DukeMTMC-reID datasets. Remarkably, our method achieves the largest improvement over previous methods on CUHK03-Detect dataset, which is the most challenging dataset. For Market1501 datasets, our model achieves comparative performance to MGN [58]. 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 BDB Network.

Some sample query results are illustrated in Figure 5. We can see that, given a back view person image, BDB Network can even retrieve the front view and side view images of the same person.

Figure 5: The top-4 ranking list for the query images on CUHK03-Label dataset from the proposed BDB Network. The correct results are highlighted by green borders and the incorrect results by red borders.

4.1.4 Ablation Studies

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

Benefit of Global Branch and Feature Dropping Branch. Without the global branch, the BDB Network still performs

Method	Rank-1	mAP
Global Branch (Baseline)	93.1	80.6
Feature Dropping Branch	93.6	83.3
Both Branches (BDB)	94.2	84.3
Feature Dropping Branch + Cut	88.0	75.7
BDB + Cut	95.3	86.7

Table 2: The effect of global branch and feature dropping branch on Market-1501 dataset. 'Cut' means Cutout [11] augmentation.

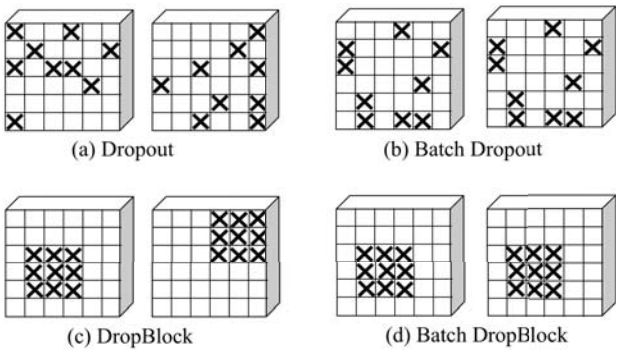

Figure 6: The comparison with Dropout methods on two feature maps within the same batch.

better than the baseline as illustrated in Table 2. Adding the global branch could further improve the performance. The motivation behind the two-branch structure in the BDB Network is that it learns both the most salient appearance clues and fine-grained discriminative features. This suggests that the two branches reinforce each other and are both important to the final performance.

Comparison with Dropout and DropBlock. Dropout [47] drops values of input tensor randomly and is a widely used regularization technique to prevent overfitting. We replace the Batch DropBlock layer with various Dropout methods and compare their performance in Table 3. SpatialDropout [52] 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 Batch DropBlock and Batch Dropout is that Batch DropBlock zeroes a large contiguous area while Batch Dropout zeroes some isolated features. DropBlock [14] means for a batch of input tensor, every tensor randomly drops a contiguous region. The difference between Batch DropBlock and DropBlock is that Batch DropBlock drops the same region for every input tensor within a batch while DropBlock crops out different regions. These Dropout methods are visualized in Figure 6. As shown in Table 3, Batch DropBlock is more effective than these various Dropout strategies in the person re-ID tasks.

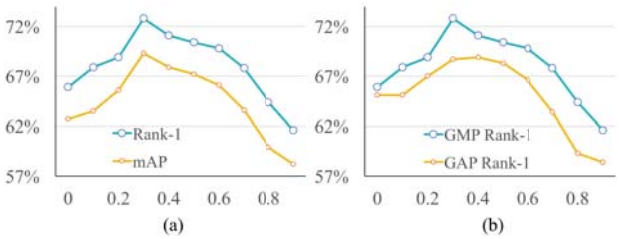

Figure 7: (a) The effects of erased height ratio on mAP and CMC scores. The erased width ratio is fixed to 1.0. (b) The comparison of global average pooling and global max pooling on the feature dropping branch under different height ratio settings. The statistics are analyzed on the CUHK03-Detect dataset.

Global Average Pooling (GAP) vs Global Max Pooling (GMP) in Feature Dropping Branch. As shown in Figure 7 (b), the Rank-1 accuracy of the feature dropping branch with GMP is consistently superior to that with GAP. We therefore demonstrate the importance of Max Pooling for a robust convergence and increased performance on the feature dropping branch.

Benefit of Triplet Loss The BDB Network is trained using both triplet loss and softmax loss. The triplet loss is a vital part of BDB Network since the Batch DropBlock layer has effect only when considering relationship between images. In table 4, 'Baseline + Dropping' is the BDB Network without triplet loss. We can see that the triplet loss significantly improves the performance.

Impact of Batch DropBlock Layer Hyper-parameters. Figure 7 (a) studies the impact of erased height ratio on the performance of the BDB Network. Here, the erased width ratio is fixed to 1.0 in all the person Re-ID experiments. We can see that the best performance is achieved when height erased ratio is 0.3, which is the setting for BDB Network in person re-ID experiments.

Relationship with Data Augmentation methods. A natural question about BDB Network is could BDB Network still benefit from image erasing data augmentation methods such as Cutout [11] and Random Erasing [83] since they perform similar operations? The answer is yes. Because the BDB Network contains a global branch which sees the complete feature map and it can benefit from Cutout or Random Erasing. To verify it, we apply image erasing augmentation on BDB Network with or without the global branch in Table 2. We can see Cutout performs bad without the global branch. Table 5 shows BDB Network performs well with data augmentation methods. As can be seen, 'BDB + Cut' or 'BDB + RE' are significantly better than 'Baseline + Cut', 'Baseline + RE', or 'BDB'.

4.2. Image Retrieval Experiments

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

Method	Rank-1	mAP
SpatialDropout[52]	60.5	56.8
Dropout [47]	65.3	62.2
Batch Dropout	65.8	62.9
DropBlock [14]	70.6	67.7
Batch DropBlock	72.8	69.3

Table 3: The Comparison with other Dropout methods on the CUHK03-Detect dataset.

Method	CUHK03-Detect		Market1501	
	Rank-1	mAP	Rank-1	mAP
Baseline	51.1	47.9	91.6	77.1
Baseline + Triplet	63.6	60.0	93.1	80.6
Baseline + Dropping	60.9	57.2	93.8	80.5
Baseline + Triplet + Dropping (BDB Network)	72.8	69.3	94.2	84.3

Table 4: Ablation studies of the effective components of BDB network on CUHK03-Detect and Market1501 datasets. 'Dropping' means the feature dropping branch.

Method	CUHK03-Detect		Market1501	
	Rank-1	mAP	Rank-1	mAP
Baseline	63.6	60.0	93.1	80.6
Baseline + RE	70.6	65.9	93.3	81.5
Baseline + Cut	67.7	64.2	93.5	82.0
Baseline + RE + Cut	70.7	65.9	93.1	82.0
BDB	72.8	69.3	94.2	84.3
BDB + RE	75.9	72.6	94.4	85.0
BDB + Cut	**76.4**	**73.5**	**95.3**	**86.7**

Table 5: The comparison with data augmentation methods. 'RE' means Random Erasing [83]. 'Cut' means Cutout [11].

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 6: The statistics of the image retrieval datastes including CARS196 [22], CUB200-2011 [57], Stanford online products(SOP) [35], and In-Shop Clothes retrieval dataset [32]. 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.

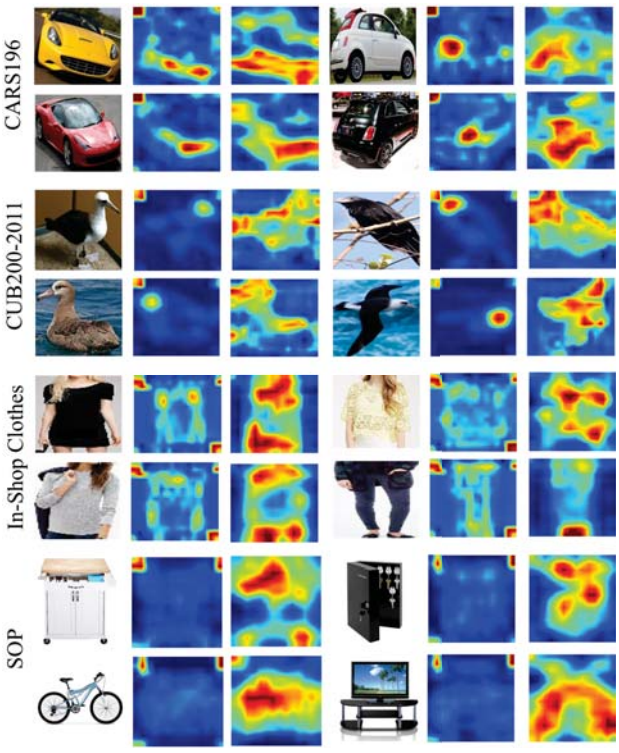

Figure 8: The class activation map of Baseline and BDB Network on CARS196, CUB200-2011, In-Shop Clothes retrieval and SOP datasets.

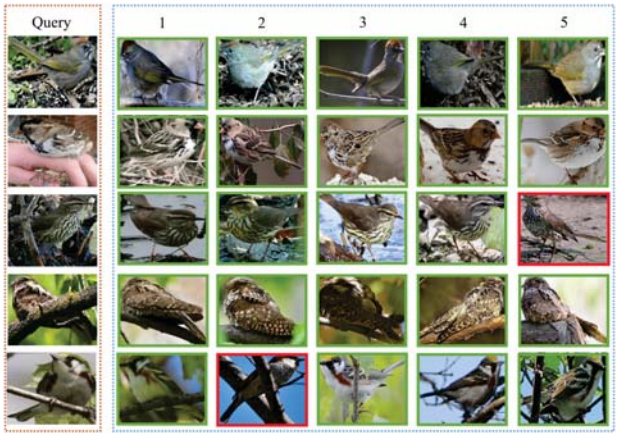

Figure 9: The top-5 ranking list for the query images on CUB200-2011 dataset from BDB Network. The green and red borders respectively denote the correct and incorrect results.

4.2.1 Datasets and Settings

Our method is evaluated on the commonly used image retrieval datasets including CUB200-2011 [57], CARS196 [22], Stanford online products (SOP) [35], and In-Shop Clothes retrieval [32] datasets. For CUB200-2011 and CARS196, the cropped datasets are used since our BDB Network requires input images to be roughly aligned. The experimental setup is the same as that in [35]. We show the statistics of the four image retrieval datasets in Table 6.

The training images are padded and resized to 256 $\times$ 256 while the aspect ratio is fixed, and then cropped to 224 $\times$ 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 $\times$ 256, while CARS196 images are scaled to 256 $\times$ 256 directly. The dropping height ratio and width ratio are both set to 0.5 in the Batch DropBlock Layer. We use the standard Recall@K metric to measure the image retrieval performance.

4.2.2 Comparison with State-of-the-Art

Table 7 shows that our BDB Network achieves the best Recall@1 scores on all the experimental image retrieval datasets. In particular, the BDB Network achieves an obvious improvement (+3.5%) on the small scale CUB200-

K	1	2	4	8
PDDM Triplet [18]	50.9	62.1	73.2	82.5
PDDM Quadruplet [18]	58.3	69.2	79.0	88.4
HDC [68]	60.7	72.4	81.9	89.2
Margin [62]	63.9	75.3	84.4	90.6
ABE-8 [20]	70.6	79.8	86.9	92.2
BDB	**74.1**	**83.6**	**89.8**	**93.6**

(a) CUB200-2011 (cropped) dataset

K	1	2	4	8
PDDM Triplet [18]	46.4	58.2	70.3	80.1
PDDM Quadruplet [18]	57.4	68.6	80.1	89.4
HDC [68]	83.8	89.8	93.6	96.2
Margin [62]	86.9	92.7	95.6	97.6
ABE-8 [20]	93.0	95.9	97.5	98.5
BDB	**94.3**	**96.8**	**98.3**	**98.9**

(b) CARS196 (cropped) dataset

K	1	10	20	30	40
FasionNet [32]	53.0	73.0	76.0	77.0	79.0
HDC [68]	62.1	84.9	89.0	91.2	92.3
DREML [65]	78.4	93.7	95.8	96.7	-
HTL [12]	80.9	94.3	95.8	97.2	97.4
A-BIER [36]	83.1	95.1	96.9	97.5	97.8
ABE-8 [20]	87.3	**96.7**	**97.9**	98.2	98.5
BDB	**89.1**	96.3	97.6	**98.5**	**99.1**

(c) In-Shop Clothes Retrieval dataset

K	1	10	100	1000
LiftedStruct [35]	62.1	79.8	91.3	97.4
N-Pairs [45]	67.7	83.8	93.0	97.8
Margin [62]	72.7	86.2	93.8	98.0
HDC [68]	69.5	84.4	92.8	97.7
A-BIER [36]	74.2	86.9	94.0	97.8
ABE-8 [20]	76.3	88.4	94.8	98.2
BDB	**83.0**	**93.3**	**97.3**	**99.2**

(d) Stanford online products dataset

Table 7: 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 [35]	66.8	88.5	93.4	96.3
BDB + LiftedStruct [35]	71.4	89.7	93.9	96.3
Baseline + Margin [62]	65.7	88.1	93.1	96.4
BDB + Margin [62]	72.0	90.8	94.4	97.0
Baseline + Histogram [54]	64.6	87.2	93.0	96.4
BDB + Histogram [54]	73.1	90.7	94.2	96.9
Baseline + Hard Triplet [17]	69.5	89.5	94.0	96.8
BDB + Hard Triplet [17]	**74.1**	**91.0**	**94.7**	**97.1**

Table 8: The BDB 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 3.

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 BDB network surpasses the state-of-the-art by 6.7%. We can see that our BDB Network is applicable on both small and large scale datasets.

Figure 9 visualizes sample retrieval results of CUB200-2011 (cropped) dataset. In Figure 1, we also present the class activation maps of Baseline and our BDB network on the CARS196 and CUB200-2011 data-sets. We can see that our two-branch network encodes more comprehensive features with attentive detail features. This helps to explain why our BDB Network is in some terms robust to the variance in illumination, poses and occlusions.

4.2.3 Adapt to Other Metric Learning Methods

Table 8 shows that our BDB Network can also be used with other standard metric learning loss functions, such as lifted structure loss[35], weighted sampling margin loss[62], and histogram loss[54] to boost their performance. For a fair comparison, we re-implement the above loss functions on our ResNet-50 Baseline and BDB Network to evaluate their performances. Here, the only difference between ResNet-50 Baseline and BDB Network is that the BDB Network has an additional feature dropping branch. For weighted sampling margin loss, although the ResNet-50 Baseline outperforms the results reported in the work [62] (+1.8%), the BDB Network can still improve the result by a large margin (+7.7%). We can therefore conclude that the proposed BDB Network can be easily generalized to other standard loss functions in metric learning.

5. Conclusion

In this paper, we propose the Batch DropBlock to improve the optimization in training a neural network for person re-ID and other general metric learning tasks. The corresponding BDB 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 BDB Network can make significant improvement on person re-ID and other general image retrieval benchmarks.

References

[1] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person re-identification. *arXiv:1801.05339*, 2018.

[2] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017.

[3] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. In *CVIU*, 2018.

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

[5] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep CRF for person re-identification. In *CVPR*, 2018.

[6] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017.

[7] Yanbei Chen, Xiatian Zhu, Shaogang Gong, et al. Person re-identification by deep learning multi-scale representations. In *ICCV*, 2018.

[8] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

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

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[12] Weifeng Ge, Weilin Huang, Dengke Dong, and Matthew R Scott. Deep metric learning with hierarchical triplet loss. In *ECCV*, 2018.

[13] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. In *BigMM*, 2018.

[14] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *arXiv:1810.12890*, 2018.

[15] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017.

[18] Chen Huang, Chen Change Loy, and Xiaoou Tang. Local similarity-aware deep feature embedding. In *NIPS*, 2016.

[19] Hungfu Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and K. Huang. Adversarially occluded samples for person re-identification. In *CVPR*, 2018.

[20] Wonsik Kim, Bhavya Goyal, Kunal Chawla, Jungmin Lee, and Keunjoo Kwon. Attention-based ensemble for deep metric learning. In *ECCV*, 2018.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV*, 2013.

[23] Vijay Kumar, Anoop M Namboodiri, Manohar Paluri, and CV Jawahar. Pose-aware person recognition. In *CVPR*, 2017.

[24] Xu Lan, Hangxiao Wang, Shaogang Gong, and Xiatian Zhu. Deep reinforcement learning attention selection for person re-identification. In *BMVC*, 2017.

[25] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[26] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[27] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[28] 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.

[29] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[30] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 2017.

[31] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017.

[32] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[33] Andy J Ma, Pong C Yuen, and Jiawei Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *ICCV*, 2013.

[34] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

[35] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.

[36] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

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

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

[39] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[41] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPRW*, 2017.

[42] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, 2018.

[43] Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015.

[44] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018.

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

[46] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. Deep metric learning via facility location. In *CVPR*, 2017.

[47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[48] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[49] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.

[50] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.

[51] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling(and a strong convolutional baseline). In *ECCV*, 2018.

[52] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015.

[53] Evgeniya Ustinova, Yaroslav Ganin, and Victor Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *AVSS*, 2017.

[54] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In *NIPS*, 2016.

[55] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[56] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[57] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[58] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. *arXiv:1804.01438*, 2018.

[59] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014.

[60] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018.

[61] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017.

[62] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *ICCV*, 2017.

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

[64] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.

[65] Hong Xuan, Richard Souvenir, and Robert Pless. Deep randomized ensembles for metric learning. In *ECCV*, 2018.

[66] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *ECCV*, 2014.

[67] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Deep representation learning with part loss for person re-identification. *arXiv:1707.00798*, 2017.

[68] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. In *ICCV*, 2017.

[69] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv:1711.08184*, 2017.

[70] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[71] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[72] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[73] Feng Zheng and Ling Shao. Learning cross-view binary identities for fast person re-identification. In *IJCAI*, 2016.

[74] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *arXiv:1701.07732*, 2017.

[75] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[76] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv:1610.02984*, 2016.

[77] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *PAMI*, 2013.

[78] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, 2015.

[79] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. In *TOMM*, 2017.

[80] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[81] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. In *TCSVT*, 2018.

[82] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[83] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv:1708.04896*, 2017.

[84] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.