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

ABD-Net: Attentive but Diverse Person Re-Identification

Tianlong Chen[1], Shaojin Ding[1*], Jingyi Xie[2*], Ye Yuan[1], Wuyang Chen[1]
Yang Yang[3], Zhou Ren[4], Zhangyang Wang[1]
[1]Texas A&M University, [2]University of Science and Technology of China
[3]Walmart Technology, [4]Wormpex AI Research
{wiwjp619,shjd,ye.yuan,wuyang.chen,atlaswang}@tamu.edu
hsfzxjy@mail.ustc.edu.cn, yang.yang2@walmart.com, zhou.ren@bianlifeng.com
https://github.com/TAMU-VITA/ABD-Net

Abstract

Attention mechanisms have been found effective for person re-identification (Re-ID). However, the learned "attentive" features are often not naturally uncorrelated or "diverse", which compromises the retrieval performance based on the Euclidean distance. We advocate the complementary powers of attention and diversity for Re-ID, by proposing an Attentive but Diverse Network (ABD-Net). ABD-Net seamlessly integrates attention modules and diversity regularizations throughout the entire network to learn features that are representative, robust, and more discriminative. Specifically, we introduce a pair of complementary attention modules, focusing on channel aggregation and position awareness, respectively. Then, we plug in a novel orthogonality constraint that efficiently enforces diversity on both hidden activations and weights. Through an extensive set of ablation study, we verify that the attentive and diverse terms each contributes to the performance boosts of ABD-Net. It consistently outperforms existing state-of-the-art methods on there popular person Re-ID benchmarks.

1. Introduction

Person Re-Identification (Re-ID) aims to associate individual identities across different time and locations. It embraces many applications in intelligent video surveillance. Given a query image and a large set of gallery images, person Re-ID represents each image with a *feature embedding*, and then ranks the gallery images in terms of feature embeddings' similarities to the query. Despite the exciting progress in recent years, person Re-ID remains to be extremely challenging in practical unconstrained scenarios. Common challenges arise from body misalignment, occlusion, background perturbance, view point changes, pose variations and noisy labels, among many others [1].

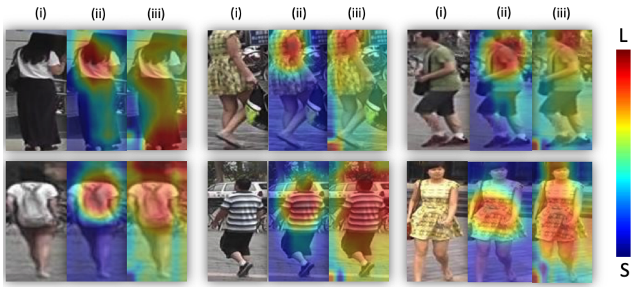


Figure 1. Visualization of attention maps. (i) Original images; (ii) Attentive feature maps; (iii) Attentive but diverse feature maps. In general, diversity is observed to make attention "broader" and to reduce the (incorrect) overfitting of local regions (such as clothes textures) by attention. (L: large values; S: small values)

Substantial efforts have been devoted to addressing those various challenges. Among them, incorporating body part information [2, 3, 4, 5, 6] has empirically proven to be effective in enhancing the feature robustness against body misalignment, incomplete parts, and occlusions. Motivated by such observations, the attention mechanism [7] was introduced to enforce the features to mainly capture the discriminative appearances of human bodies (or certain body parts). Since then, the attention-based models [8, 9, 10, 11, 12] have boosted person Re-ID performance much.

On a separate note, the feature embeddings are used to compute similarities between images, typically based on the Euclidean distance, to return the closest matches. Sun et al. [13] pointed out that correlations among feature embeddings would significantly compromise the matching performance. The low feature correlation property is, however, not naturally guaranteed by attention-based models. Our observation is that those attention-based models are often more prone to higher feature correlations, because intuitively, the attention mechanism tends to have features focus on a more compact subspace (such as foreground instead of

[*] Equal Contribution.

2380-7504/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCV.2019.00844

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

the full image, see Fig.1 for examples).

In view of the above, we argue that a more desirable feature embedding for person Re-ID should be both **attentive** and **diverse**: the former aims to correct misalignment, eliminate background perturbance, and focus on discriminative local parts of body appearances; the latter aims to encourage lower correlation between features, and therefore better matching, and potentially make feature space more comprehensive. We propose an Attentive but Diverse Network (**ABD-Net**), that strives to integrate attention modules and diversity regularization and enforces them throughout the entire network. The main contributions of ABD-Net are outlined as below:

- We incorporate a compound attention mechanism into ABD-Net, consisting of *Channel Attention Module* (CAM) and *Position Attention Module* (PAM). CAM facilitates channel-wise, feature-level information aggregation, while PAM captures the spatial awareness of body and part positions. They are found to be complementary and altogether benefit Re-ID.
- We introduce a novel regularization term called *spectral value difference orthogonality* (SVDO) that directly constrains the conditional number of the weight Gram matrix. SVDO, efficiently implemented, is applied to both activations and weights, and is shown to effectively reduce learned feature correlations.
- We perform extensive experiments on Market-1501 [14], DukeMTMC-Re-ID [15], and MSMT17 [1]. ABD-Net significantly outperforms existing methods, achieving new state-of-the-art on all three popular benchmarks. We also verify that the attentive and diverse terms each contributes to a performance gain, through rigorous ablation studies and visualizations.

2. Related Work

2.1. Person Re-identification: Brief Overview

Person Re-ID has two key steps: obtaining a feature embedding and performing matching under some distance metric [16, 17, 18]. We mainly review the former where both handcrafted features [18, 19, 20, 21] and learned features [22, 23, 4, 24, 25] were studied. In recent years, the prevailing success of convolutional neural networks (CNNs) in computer vision has made person Re-ID no exception. Due to many problem-specific challenges such as occlusion/misalignment, incomplete body parts, as well as background perturbance/view point changes, naively applying CNN backbones to feature extraction may not yield ideal Re-ID performance. Both image-level features and local features extracted from body parts prove to enhance the robustness. Many part-based methods have achieved superior performance [2, 3, 26, 27, 28, 4, 5, 29, 6, 30, 31, 8, 32]. We refer readers to [33] for a more comprehensive review.

2.2. Attention Mechanisms in Person Re-ID

Several studies proposed to integrate attention mechanism into deep models to address the misalignment issue in person Re-ID. Zhao et al. [8] proposed a part-aligned representation based on a part map detector for each predefined body part. Yao et al. [9] proposed a Part Loss Network which defined a loss for each average pooled body part and jointly optimized the summation losses. Si et al. [10] proposed a dual attention matching network based on an inter-class and an intra-class attention module to capture the context information of video sequences for person Re-ID. Li et al. [12] proposed a multi-task learning model that learns hard region-level and soft pixel-level attention jointly to produce more discriminative feature representations. Xu et al. [11] used pose information to learn attention masks for rigid and non-rigid parts, and then combined the global and part features as the final feature embedding.

Our proposed attention mechanism differs from previous methods in several aspects. First, previous methods [8, 9, 11] only use attention mechanisms to extract part-based spatial patterns from person images, which are usually focus in the foregrounds. In contrast, ABD-Net combines spatial and channel clues; besides, our added diversity constraint will avoid the overly correlated and redundant attentive features. Second, our attention masks are directly learned from the data and context, without relying on manually-defined parts, part region proposals, nor pose estimation [8, 9, 11]. Our two attention modules are embedded within a single backbone, making our model lighter-weight than the multi-task learning alternatives [11, 12].

2.3. Diversity via Orthogonality

Orthogonality has been widely explored in deep learning to encourage the learning of informative and diverse features. In CNNs, several studies [34, 35, 36, 37] perform regularization using "hard orthogonality constraints", which typically depends on singular value decomposition (SVD) to strictly constrain their solutions on a Stiefel manifold. The similar idea was first exploited by [13] for person Re-ID, where the authors performed SVD on the weight matrix of the last layer, in an effort to reduce feature correlations. Despite the effectiveness, SVD-based hard orthogonality constraints are computationally expensive, and sometimes appear to limit the learning flexibility.

Recent studies also investigated "softer" orthogonality regularizations by enforcing the Gram matrix of each weight matrix to be close to an identity matrix, under Frobenius norm [38] or spectral norm [39]. We propose a novel spectral value difference orthogonality (SVDO) regularization that directly constrains the conditional number of the Gram matrix. Also contrasting from [13, 38] that apply orthogonality only to CNN weights, we enforce the new regularization on both hidden activations and weights.

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

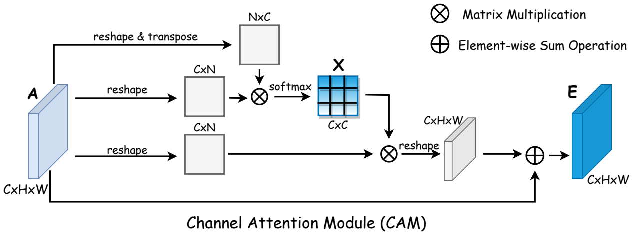


Figure 2. Channel Attention Module (CAM)

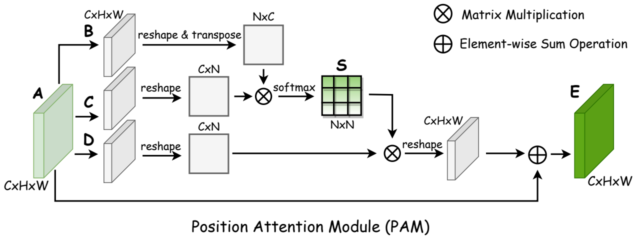


Figure 3. Position Attention Module (PAM)

3. Attentive but Diverse Network

In this section, we first introduce the two attention modules, followed by the new diversity (orthogonality) regularization. We then wrap them up and describe the overall architecture of ABD-Net.

3.1. Attention: Channel-Wise and Position-Wise

The goal of attention for Re-ID is to focus on person-related features while eliminating irrelevant backgrounds. Inspired by the successful idea in segmentation [40], we integrate two complementary attention mechanisms: Channel Attention Module (**CAM**) and Positional Attention Module (**PAM**). The full configurations for CAM and PAM can be found in the supplementary.

3.1.1 Channel Attention Module

The high-level convolutional channels in a trained CNN classifier are well-known to be semantic-related and often category-selective. In the person Re-ID case, we hypothesize that the high-level channels in the person Re-ID case are also "grouped", *i.e.*, some channels share similar semantic contexts (such as foreground person, occlusions, or background) and are more correlated with each other. CAM is designed to group and aggregate those semantically similar channels.

The full structure of CAM is illustrated in Fig.2. Given the input feature maps $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$, where C is the total number of channels and $H \times W$ is the feature map size, we compute the channel affinity matrix $\mathbf{X} \in \mathbb{R}^{C\times C}$, as shown below:

$$x_{ij} = \frac{exp(A_i \cdot A_j)}{\sum_{j=1}^{C} exp(A_i \cdot A_j)},\ i,j \in \{1, \cdots, C\} \quad (1)$$

where x_{ij} represents the impact of channel i on channel j. The final output feature map $\mathbf{E}$ is calculated by equation (2):

$$E_i = \gamma \sum_{j=1}^{C} (x_{ij} A_j) + A_i,\ i \in \{1, \cdots, C\} \quad (2)$$

γ is a hyperparameter to adjust the impact of CAM.

3.1.2 Position Attention Module

In contrast to CAM, Position Attention Module (PAM) is designed to capture and aggregate those semantically related *pixels* in the spatial domain. We depict the structure of PAM in Fig.3. The input feature maps $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ are first fed into convolution layers with batch normalization and ReLU activation to produce feature maps $\mathbf{B, C, D} \in \mathbb{R}^{C\times H\times W}$. Then we compute the pixel affinity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$ where $N = H \times W$. Note that the dimensions of S and X are different, since the former computes correlations between the total N pixels rather than C channels. We generate the final output feature map $\mathbf{E}$ with similar calculation as CAM in Section 3.1.1.

3.2. Diversity: Orthogonality Regularization

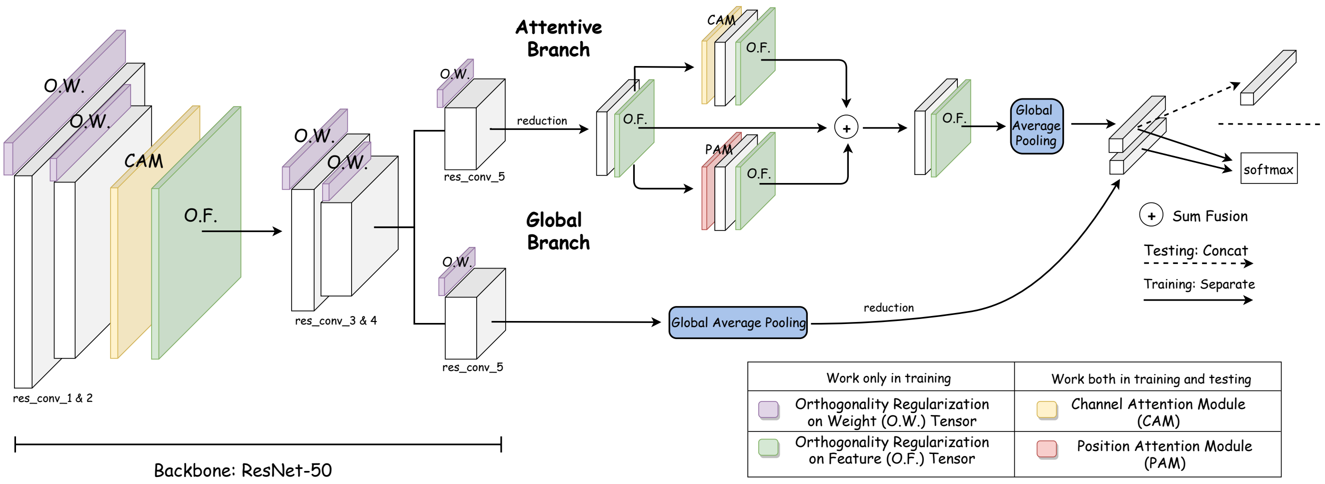
Following [13], we enforce diversity via orthogonality, yet derive a novel orthogonality regularizer term. It is applied to both hidden features and weights, of both convolutional and fully-connected layers. Orthogonality regularizer on feature space (short for **O.F.** hereinafter) is to reduce feature correlations that can directly benefit matching. The orthogonal regularizer on weight (**O.W.**) encourages filter diversity [39] and enhances the learning capacity.

Next, we show the detailed derivation of our orthogonality term on features, while the weight orthogonality can be derived in a similar manner[*]. For feature maps $\mathbf{M} \in \mathbb{R}^{C\times H\times W}$, where C, H, W are the number of channels, feature map's height and width, respectively, we will first reshape $\mathbf{M}$ into a matrix form $\mathbf{F} \in \mathbb{R}^{C\times N}$, with $N = H \times W$.

Many orthogonality methods [34, 35, 36, 37], including the prior work on person Re-ID [13], enforce hard constraints on orthogonality of weights, whose computations rely on SVD. However, computing SVD on high-dimensional matrices is expensive, urging for the development of soft orthogonality regularizers. Many existing soft regularizers [38, 41] restrict the Gram matrix of $\mathbf{F}$ to be close to an identity matrix under Frobenius norm that can avoid the SVD step while being differentiable. However,

[*] For the weight tensor $\mathbf{W}_c \in \mathbb{R}^{S\times H\times C\times M}$ in a convolutional layer, where S, H, C, M are filter's width and height, the number of input and output channels, we follow the convention of [38, 39] to reshape $\mathbf{W}_c$ into a matrix form $\mathbf{F}^* \in \mathbb{R}^{C^*\times M}$, where $C^* = S \times H \times C$.

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.



Work only in training	Work both in training and testing
Orthogonality Regularization on Weight (O.W.) Tensor	Channel Attention Module (CAM)
Orthogonality Regularization on Feature (O.F.) Tensor	Position Attention Module (PAM)

Figure 4. Architecture of ABD-Net: O.W. is applied on all ResNet layers. O.F. is applied after CAM on res_conv_2 and after res_conv_5 in the Attentive Branch. l The feature vectors from both attentive and global branches are concatenated as the final feature embedding.

the gram matrix for an overcomplete **F** cannot reach identity because of rank deficiency, making those regularizers biased. [39] hence introduced the spectral norm-based regularizer that effectively alleviates the bias.

We propose a new option to enforce the orthogonality via directly regularizing the conditional number of $\mathbf{F}\mathbf{F}^T$:

$$\beta||k(\mathbf{F}) - 1||_2^2 \,, \tag{3}$$

where β is the coefficient and $k(\mathbf{F})$ denotes the condition number of **F**, defined as the ratio of maximum singular value to minimum singular value of **F**. Naively solving $k(\mathbf{F})$ will take one full SVD. To make it computationally more tractable, we convert (3) into a spectral value difference orthogonality (SVDO)[†] regularization:

$$\beta||\lambda_1(\mathbf{F}\mathbf{F}^T) - \lambda_2(\mathbf{F}\mathbf{F}^T)||_2^2 \,, \tag{4}$$

where $\lambda_1(\mathbf{F}\mathbf{F}^T)$ and $\lambda_2(\mathbf{F}\mathbf{F}^T)$ denote the largest and smallest eigenvalues of $\mathbf{F}\mathbf{F}^T$, respectively.

We use auto-differentiation to obtain the gradient of SVDO, however, this computation still contains the expensive eigenvalue decomposition (EVD). To bypass EVD, we refer to the power iteration method to approximate the eigenvalues. We start with a random initialized q, and then iteratively perform equation (5) (two times by default):

$$p \leftarrow Xq \,, q \leftarrow Xp \,, \lambda(X) \leftarrow \frac{||q||}{||p||} \,. \tag{5}$$

where **X** in equation (5) is $\mathbf{F}\mathbf{F}^T$ for computing $\lambda_1(\mathbf{F}\mathbf{F}^T)$, and $\mathbf{F}\mathbf{F}^T - \lambda_1\mathbf{I}$ for $\lambda_2(\mathbf{F}\mathbf{F}^T)$. In that way, the computation of SVDO becomes practically efficient.

3.3. Network Architecture Overview

The overall architecture of the proposed ABD-Net is shown in Fig.4. ABD-Net is compatible with most common feature extraction backbones, such as ResNet [42], InceptionNet [43], and Densenet [44]. Unless otherwise specified, we use ResNet-50 as the default backbone network due to its popularity in Re-ID [45, 46, 47, 48, 49, 11, 50, 51].

We add a CAM and O.F. on the outputs of res_conv_2 block. The regularized feature map is used as the input of res_conv_3. Next, after the res_conv_4 block, the network splits into a *global branch* and an *attentive branch* in parallel. We apply O.W. on all conv layers in our ResNet-50 backbone, $i.e.$, from res_conv_1 to res_conv_4 and the two res_conv_5 in both branches. The outputs of two branches are concatenated as the final feature embedding.

The *attentive branch* uses the same res_conv_5 layer as that in ResNet-50. The output feature map is then fed into a reduction layer[‡] with O.F. applied yielding a smaller feature map $\mathbf{T}_a$. We feed $\mathbf{T}_a$ into a CAM and a PAM simultaneously, both with O.F. constraints. The outputs from both attentive modules are concatenated with the input $\mathbf{T}_a$, and altogether go through a global average pooling layer, ending up with a k_a-dimension feature vector.

In the *global branch*, after res_conv_5[§], the feature map $\mathbf{T}_g$ is fed into a global average-pooling layer followed by a reduction layer, leading to a k_g-dimension feature vector. The global branch intends to preserve global context information in addition to the attentive branch features.

Eventually, ABD-Net is trained under the loss function L

[†] The reason why we choose to penalize the difference between $\lambda_1(\mathbf{F}\mathbf{F}^T)$ and $\lambda_2(\mathbf{F}\mathbf{F}^T)$ rather than the ratio of them is to avoid numerical instability caused by dividing a very small $\lambda_2(\mathbf{F}\mathbf{F}^T)$, which we find happen frequently in our experiments.

[‡] A reduction layer consists of a linear layer, batch normalization, ReLU, and dropout. See: https://github.com/KaiyangZhou/deep-person-reid

[§] For both two res_conv_5 layers in two branches, we removed the down-sampling layer, in order for larger feature maps.

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

consisting of a cross entropy loss, a hard mining triplet loss, and orthogonal constraints on feature (O.F.) and on weights (O.W.) penalty terms:

$$L = L_{xent} + \beta_{tr} L_{triplet} + \beta_{O.F.} L_{O.F.} + \beta_{O.W.} L_{O.W.} \quad (6)$$

where $L_{O.F.}$ and $L_{O.W.}$ stand for the SVDO penalty term applied to the hidden features and weights, respectively. $\beta_{tr}, \beta_{O.F.}$ and $\beta_{O.W.}$ are hyper-parameters.

4. Experiments

To evaluate ABD-Net, we conducted experiments on three large-scale person re-identification datasets: Market-1501 [14], DukeMTMC-Re-ID [15] and MSMT17 [1]. First, we report a set of ablation study (mainly on Market-1501 and DukeMTMC-Re-ID) to validate the effectiveness of each component. Second, we compare the performance of ABD-Net against existing state-of-the-art methods on all three datasets. Finally, we provide more visualizations and analysis to illustrate how ABD-Net has achieved its effectiveness.

4.1. Datasets

Market-1501 [14] comprises 32,668 labeled images of 1,501 identities captured by six cameras. Following [14], 12,936 images of 751 identities are used for training, while the rest are used for testing. Among the testing data, the test probe set has 3,368 images of 750 identities. The test gallery set also includes 2,793 additional distractors.

DukeMTMC-Re-ID [15] contains 36,411 images of 1,812 identities. These images are captured by eight cameras, among which 1,404 identities appear in more than two cameras and 408 identities (distractors) appear in only one camera. The 1,404 identities are randomly divided, with 702 identities for training and the others for testing. In the testing set, one query image for each ID per camera is chosen for the probe set, while all remaining images including distractors are in the gallery.

MSMT17 [1] is the current largest publicly-available person Re-ID dataset. It has 126,441 images of 4,101 identities captured by a 15-camera network (12 outdoor, 3 indoor). We follow the training-testing split of [1]. The video is collected with different weather conditions at three-time slots (morning, noon, afternoon). All annotations, including camera IDs, weathers and time slots, are available. MSMT17 is **significantly more challenging** than the other two, due to its massive scale, more complex and dynamic scenes. Additionally, the amount of methods that report on this dataset is limited since it is recently released.

4.2. Implementation Details and Evaluation

During training, the input images are re-sized to 384 × 128 and then augmented by random horizontal flip, normalization, and random erasing [52]. The testing images are re-sized to 384 × 128 and augmented only by normalization. In our experiments, the sizes of feature maps $\mathbf{T}_a$ and $\mathbf{T}_g$ are $1024 \times 24 \times 8$, and $2048 \times 24 \times 8$, respectively. We set the dimension of features (k_a, k_g) after global average-pooling both equal to 1024, leading to a 2048-dimensional final feature embedding for matching.

With the ImageNet-pretrained ResNet-50 backbone, we used the two-step transfer learning algorithm [53] to fine-tune the model. First, we freeze the backbone weights and only train the reduction layers, classifiers and all attention modules for 10 epochs with only the cross entropy loss and triplet loss applied. Second, all layers are freed for training for another 60 epochs, with the full loss (6) applied. We set $\beta_{tr} = 10^{-1}$, $\beta_{OF} = 10^{-6}$ and $\beta_{OW} = 10^{-3}$, and the margin parameter for triplet loss $\alpha = 1.2$.

Our network is trained using 2 Tesla P100 GPUs with a batch size of 64. Each batch contains 16 identities, with 4 instances per identity. We use the Adam optimizer with the base learning rate initialized to 3×10^{-4}, then decayed to 3×10^{-5}, 3×10^{-6} after 30, 40 epochs, respectively. The training takes about 4 hours on the Market-1501 dataset.

We adopt standard Re-ID metrics: top-1 accuracy, and the mean Average Precision (mAP). We consider mAP to be a more reliable indicator for Re-ID performance.

4.3. Ablation Study of ABD-Net

To verify the effects of attention modules and orthogonality regularization in ABD-Net, we incrementally evaluate each module on Market-1501 and DukeMTMC-Re-ID. We choose ResNet-50 [¶] with the cross entropy loss (XE) as the baseline. Nine variants are then constructed on top of the baseline[‖]: **a)** baseline (XE) + PAM; **b)** baseline (XE) + CAM; **c)** baseline (XE) + PAM + CAM; **d)** baseline (XE) + O.F.; **e)** baseline (XE) + O.W.; **f)** baseline (XE) + O.F. + O.W.; **g)** baseline + SVD layer (similar to SVD-Net [13]); **h)** ABD-Net (XE), that sets $\beta_{tr} = 0$ in (6); and **i)** ABD-Net, that uses the full loss (6).

Table 1 presents the ablation study results, from which several observations could be drawn:

- Using either PAM or CAM improves the baseline on both datasets. The combination of the two different attention mechanisms gains further improvements, demonstrating their complementary power over utilizing either alone.

- Using either O.F. or O.W. consistently outperforms the baseline on both datasets, and their combination leads to further gains which validates the effectiveness of

[¶] For the fairness of ablation study, we use two duplicated branches with the same res_conv_5 like the structure in ABD-Net as shown in Fig.4. Data augmentation and dropout are applied.

[‖] Note that (1) CAM is used in two places of ABD-Net; (2) ABD-Net adopts O.F. + O.W. + PAM + CAM.

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

Table 1. Ablation Study of ABD-Net on Market-1501. O.F. and O.W.: Orthogonality Regularization on Features and Weights; PAM and CAM: Position and Channel Attention Modules.

Method	Market-1501		DukeMTMC	
	top1	mAP	top1	mAP
baseline (XE)	91.50	77.40	82.80	66.40
baseline (XE) + PAM	92.10	78.10	83.80	67.00
baseline (XE) + CAM	91.80	78.00	84.30	67.60
baseline (XE) + PAM + CAM	92.70	78.50	84.40	67.90
baseline (XE) + O.F.	92.90	82.10	84.90	71.30
baseline (XE) + O.W.	92.50	78.50	83.70	67.40
baseline (XE) + O.F. + O.W.	93.20	82.30	85.30	72.20
baseline + SVD layer	90.80	76.90	79.40	62.50
ABD-Net (XE)	94.90	85.90	87.30	76.00
ABD-Net	95.60	88.28	89.00	78.59

our orthogonality regularizations. We also observe that the proposed SVDO-based O.W. empirically performs better than the SVD layer, potentially because SVD layer acts as a "hard constraint" and hence restricts the learning capability of the ResNet-50 backbone.

- By combining "attention" and "diversity", ABD-Net (XE) sees further boosts. For example, on Market-1501, ABD-Net (XE) outperforms the "no attention" counterpart (baseline (XE) + O.F. + O.W.) by a margin of 1.50% (top-1)/3.60% (mAP), and it outperforms "no diversity" counterpart (baseline (XE) + O.F. + O.W.) by 2.20% (top-1)/7.40% (mAP). Moreover, there are further performance improvements when we enforce diversity in the attention mechanism. Finally, the full ABD-Net further benefits from adding triplet loss.

4.4. Comparison to State-of-the-art Methods

We compare ABD-Net against the state-of-the-art methods on Market-1501, DukeMTMC-Re-ID and MSMT17, as shown in Tables 2, 3, and 4, respectively. For fair comparison, no post-processing such as re-ranking [54] or multi-query fusion [55] was used for our methods.

ABD-Net has clearly yielded overall state-of-the-art performance on all datasets. Specifically, on DukeMTMC-Re-ID, ABD-Net obtains 89.00% top-1 accuracy and 78.59% mAP, which significantly outperforms all existing methods. On MSMT17, ABD-Net presents a clear winner case too. On Market-1501, its top-1 accuracy (95.60%) slightly lags behind Local CNN [48] (95.90%) and MGN [47] (95.70%); yet ABD-Net clearly surpasses all existing methods in terms of mAP (88.28%, outperforming the closest competitor [48] by a large margin of 0.88%).

Specifically, we emphasize the comparison between ABD-Net and existing attention-based methods (marked by $*$ in the Tables 2 3). As shown in Table 2 and 3, ABD-Net achieves at least 2.40% top-1 and 5.98% mAP improvement on Market-1501, compared to the closest attention-based prior work CA^3Net [51]. On DukeMTMC, the margin becomes 3.40% for top-1 and 6.40% for mAP. We also considered SVDNet [13] and HA-CNN [50] which also proposed to generate diverse and uncorrelated feature embeddings. ABD-Net surpasses both with significant top-1 and mAP improvement. Overall, our observations endorse the superiority of ABD-Net by combing "attentive" and "diverse".

Table 2. Comparison to state-of-the-art methods on Market-1501. Red denotes our performance, and Blue denotes the best performance reported by existing methods: the same hereinafter.

Method	Market-1501	
	top1	mAP
BOW [55] (2015 ICCV)	44.42	20.76
Re-Rank [54] (2017 CVPR)	77.11	63.63
SSM [56] (2017 CVPR)	82.21	68.80
SVDNet(RE) [52] (2017 CVPR)	87.08	71.31
AWTL [57] (2018 CVPR)	84.20	68.03
DSR [58] (2018 CVPR)	83.68	64.25
MLFN [59] (2018 CVPR)	90.00	74.30
Deep CRF [60] (2018 CVPR)	93.50	81.60
Deep KPM [61] (2018 CVPR)	90.10	75.30
HAP2S [62] (2018 ECCV)	84.20	69.76
SGGNN [63] (2018 ECCV)	92.30	82.08
Part-aligned [31] (2018 ECCV)	91.70	79.60
PCB [64] (2018 ECCV)	93.80	81.60
SNL [45] (2018 ACM MM)	88.27	73.43
HDLF [46] (2018 ACM MM)	93.30	79.10
‡ MGN [47] (2018 ACM MM)	95.70	86.90
‡ Local CNN [48] (2018 ACM MM)	95.90	87.40
* MGCAM [49] (2018 CVPR)	83.79	74.33
* AACN [11] (2018 CVPR)	85.90	66.87
* HA-CNN [50] (2018 CVPR)	91.20	75.70
* CA^3Net [51] (2018 CVPR)	93.20	80.00
* Mancs [65] (2018 ECCV)	93.10	82.30
* A^3M [66] (2018 ACM MM)	86.54	68.97
• SPReID [67] (2018 CVPR)	93.68	83.36
$*$ ⋄ DuATM [68] (2018 CVPR)	91.42	76.62
ABD-Net	95.60	88.28

* This also exploits attention mechanisms.
• This is with a **ResNet-152** backbone.
⋄ This is with a **DenseNet-121** backbone.
‡ Official codes are not released. We report the numbers in the original paper, which are better than our re-implementation.

4.5. Visualizations[**]

Attention Pattern Visualization: We conduct a set of attention visualizations[††] on the final output feature maps of

[**] To fairly evaluate the contribution of our proposed attentive mechanism and diversity regularization, we exclude the effect of triplet loss, and only compare the following three methods: the baseline (XE), baseline (XE) + PAM + CAM, and ABD-Net (XE).

[††] Grad-CAM visualization method [73]: https://github.com/utkuozbulak/pytorch-cnn-visualizations; RAM visualization method [74] for testing images. More results can be found in the supplementary.

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

Table 3. Comparison to state-of-the-art methods on DukeMTMC.

Method	DukeMTMC-Re-ID	
	top1	mAP
BOW [55] (2015 ICCV)	25.13	12.17
SVDNet [13] (2017 ICCV)	76.70	56.80
SVDNet(RE) [52] (2017 CVPR)	79.31	62.44
FMN [69] (2017 CVPR)	74.51	56.88
PAN [70] (2018 TCSVT)	71.59	51.51
AWTL(2-stream) [57] (2018 CVPR)	79.80	63.40
Deep-person [71] (2018 CVPR)	80.90	64.80
MLFN [59] (2018 CVPR)	81.20	62.80
GP-Re-ID [72] (2018 CVPR)	85.20	72.80
PCB [64] (2018 ECCV)	83.30	69.20
Part-aligned [31] (2018 ECCV)	84.40	69.30
‡ MGN [47] (2018 ACM MM)	88.70	78.40
‡ Local CNN [48] (2018 ACM MM)	82.23	66.04
* AACN [11] (2018 CVPR)	76.84	59.25
* HA-CNN [50] (2018 CVPR)	80.50	63.80
* CA^3Net [51] (2018 CVPR)	84.60	70.20
* Mancs [65] (2018 ECCV)	84.90	71.80
• SPReID [67] (2018 CVPR)	85.95	73.34
∗ ⋄ DuATM [68] (2018 CVPR)	78.74	62.26
ABD-Net	89.00	78.59

* This also exploits attention mechanisms.
• This is with a **ResNet-152** backbone.
⋄ This is with a **DenseNet-121** backbone.
‡ Official codes are not released. We report the numbers in the original paper, which are better than our re-implementation.

Table 4. Comparison to state-of-the-art methods on MSMT17.

Method	MSMT17		
	top1	top5	mAP
PDC [5] (2017 ICCV)	58.00	73.60	29.70
GLAD [29] (2017 ACM MM)	61.40	76.80	34.00
ABD-Net	82.30	90.60	60.80

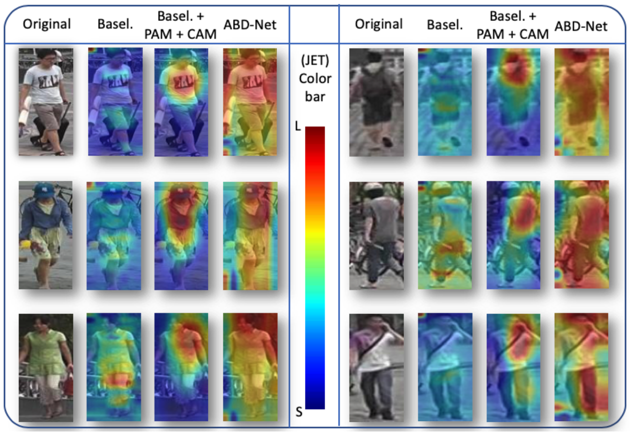


Figure 5. Visualization of attention maps from Baseline, Baseline + PAM + CAM and ABD-Net (XE). As shown in column four and eight, the diverse attention map from ABD-Net almost span over the whole person rather than overfit in some local regions.

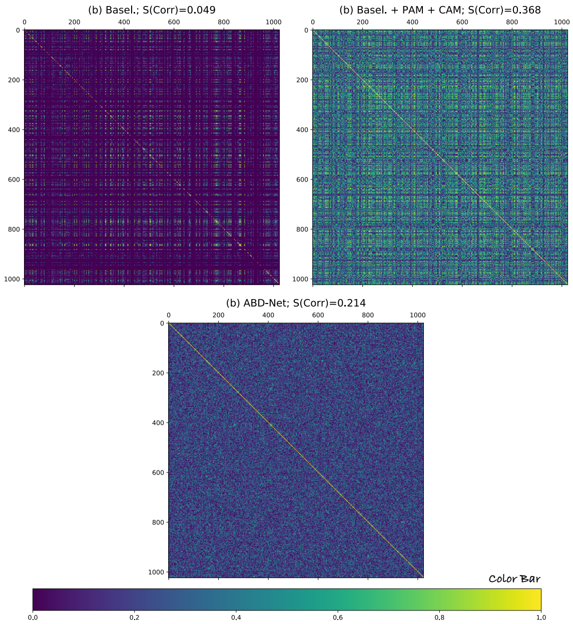


Figure 6. Visualization of correlation matrix between channels from Baseline, Baseline + PAM + CAM and ABD-Net (XE). Brighter color indicates larger correlation. In clockwise order from top left image, attention brings feature embeddings high correlation, diversity reduces the redundancy and further improve the discriminative.

the baseline (XE), baseline (XE) + PAM + CAM, and ABD-Net (XE), as shown in Fig.5. We notice that the feature maps from the baseline show little attentiveness. PAM + CAM enforces the network to focus more on the person region, but the attention regions can sometimes overly emphasize some local regions (*e.g.*, clothes), implying the risk of overfitting person-irrelevant nuisances. Most channels focus on the similar region may also cause a high correlation in the feature embeddings. In contrast, the attention of ABD-Net (XE) can strike a better balance: it focuses on more of the local parts of the person's body while still being able to eliminate the person from backgrounds. The attention patterns now differ more from person to person, and the feature embeddings become more decorrelated and diverse.

Feature De-correlation: We study the correlation matrix between the channel outputs produced by Baseline, Baseline + PAM + CAM and ABD-Net (XE) [‡‡]. The feature embedding before the global average pooling is reshaped into $\mathbf{F} \in \mathbb{R}^{C\times N}$, where $N = H \times W$. Then, we visualize the correlation coefficient matrix for $\mathbf{F}$, denoted as **Corr** $\in \mathbb{R}^{C\times C}$[§§] in Fig.6, and also compute the average of all correlation coefficients in each setting. The baseline fea-

[‡‡] Here we used a random testing image as the example and we offer more results in the supplementary.

[§§] We take the absolute value for correlation coefficients.

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

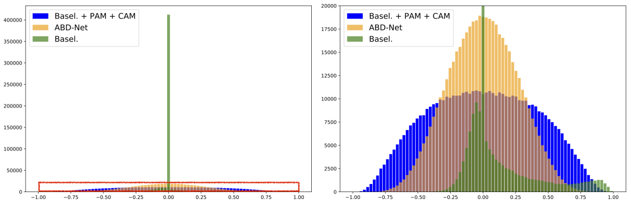


Figure 7. Histogram of correlation from Baseline, Baseline + PAM + CAM and ABD-Net (XE). A more skewed distribution indicates better de-correlated feature embeddings. (b) is a zoom-in view of the red box area in (a).

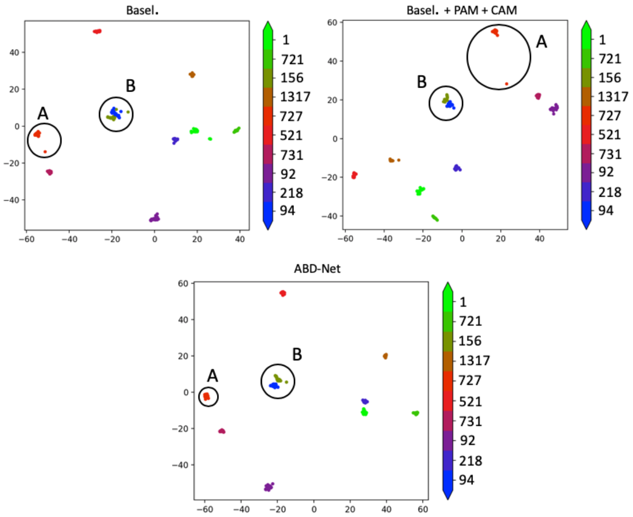


Figure 8. t-SNE visualization of feature distributions, from Baseline, Baseline + PAM + CAM and ABD-Net (XE). Ten identities are randomly selected from the Market-1501 and their IDs are listed on the right side of graphs. Circles A and B contain the features from IDs 521, 94 and 156, respectively.

ture embeddings reveal low correlations (**0.049** in average) in off-diagonal elements. After applying PAM and CAM, the feature correlations become much larger (**0.368** in average), supporting our hypothesis that the attention mechanism tends to encourage more "focused" and thus highly correlated features. However, with our orthogonality regularization, the feature correlations in ABD-Net (XE) are successfully suppressed (**0.214** in average) compared to the attention-only case. The feature histogram plots in Fig.7 also certify the same observation.

Feature Embeddings Distributions: Fig.8 shows the t-SNE visualization on feature distributions from Baseline, Baseline + PAM + CAM and ABD-Net (XE) using t-SNE. Compared with Baseline, although attentive features from Baseline + PAM + CAM make ID 94 and ID 156 in cycle B slightly distinguishable, ABD-Net enlarges the intra-class distance of ID 521 in cycle A. It makes the features from ID 94 and ID 156 more discriminative, meanwhile the features from ID 521 also lie in a compact region.

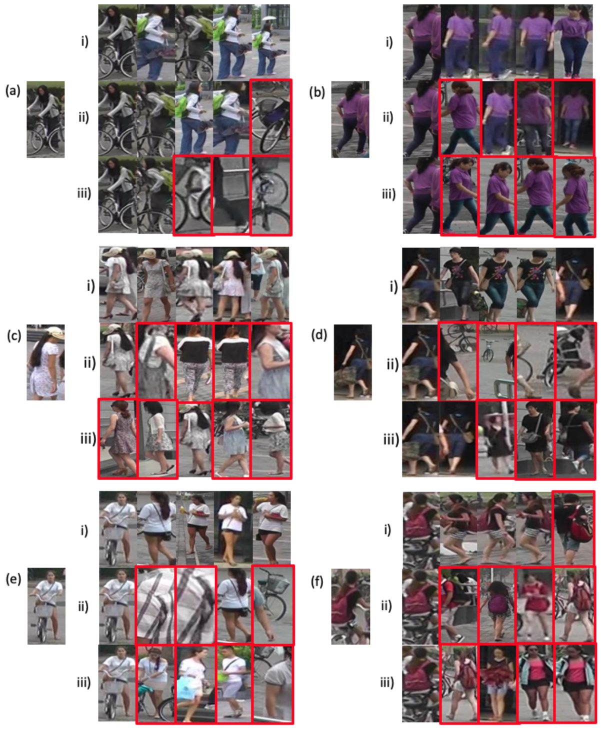


Figure 9. Six Re-ID examples of ABD-Net (XE), Baseline + PAM + CAM and Baseline on Market-1501. Left: query image. Right: i): top-5 results of ABD-Net (XE). ii): top-5 results of Baseline + PAM + CAM. iii): top-5 results of Baseline. Images in red boxes are negative results. Attentive but diverse feature embeddings boost the retrieval preformance.

Re-ID Qualitative Visual Results: Fig.9 shows Re-ID visual examples of ABD-Net (XE), Baseline + PAM + CAM and Baseline on Market-1501. They indicate that ABD-Net succeeds in finding more true positives than Baseline + PAM + CAM model, even when the persons in the images are under significant view changes and appearance variations.

5. Conclusion

This paper proposes a novel Attentive but Diverse Network (ABD-Net) to learn more representative, robust, discriminative feature embeddings for person Re-ID. ABD-Net demonstrates its state-of-the-art performance through extensive experiments where the ablations and visualizations show that each added component substantially contributes to its final performance. In the future, we will generalize the design concept of ABD-Net to other computer vision tasks.

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

References

[1] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 5

[2] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008. 1, 2

[3] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010. 1, 2

[4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. 1, 2

[5] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 3980–3989. IEEE, 2017. 1, 2, 7

[6] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017. 1, 2

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1

[8] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3239–3248, 2017. 1, 2

[9] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Deep representation learning with part loss for person re-identification. *arXiv preprint arXiv:1707.00798*, 2017. 1, 2

[10] 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. *arXiv preprint arXiv:1803.09937*, 2018. 1, 2

[11] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *arXiv preprint arXiv:1805.03344*, 2018. 1, 2, 4, 6, 7

[12] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, volume 1, page 2, 2018. 1, 2

[13] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 3, 5, 6, 7

[14] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2, 5

[15] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *The European Conference on Computer Vision (ECCV)*, September 2016. 2, 5

[16] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006. 2

[17] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith. Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3610–3617, 2013. 2

[18] Sameh Khamis, Cheng-Hao Kuo, Vivek K Singh, Vinay D Shet, and Larry S Davis. Joint learning for attribute-consistent person re-identification. In *European Conference on Computer Vision*, pages 134–146. Springer, 2014. 2

[19] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012. 2

[20] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013. 2

[21] Bingpeng Ma, Yu Su, and Frédéric Jurie. Bicov: a novel image representation for person re-identification and face verification. In *British Machive Vision Conference*, pages 11–pages, 2012. 2

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 2

[23] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–151, 2014. 2

[24] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[25] Wentong Liao, Michael Ying Yang, Ni Zhan, and Bodo Rosenhahn. Triplet-based deep similarity learning for person re-identification. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 385–393, 2017. 2

[26] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence represen-

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

tation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 2

[27] Andy J Ma, Pong C Yuen, and Jiawei Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3567–3574, 2013. 2

[28] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668, 2013. 2

[29] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 420–428. ACM, 2017. 2, 7

[30] Fuqing Zhu, Xiangwei Kong, Liang Zheng, Haiyan Fu, and Qi Tian. Part-based deep hashing for large-scale person re-identification. *IEEE Transactions on Image Processing*, 26(10):4806–4817, 2017. 2

[31] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. *arXiv preprint arXiv:1804.07094*, 2018. 2, 6, 7

[32] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017. 2

[33] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2

[34] Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. The numpy array: a structure for efficient numerical computation. *CoRR*, abs/1102.1523, 2011. 2, 3

[35] Mehrtash Harandi and Basura Fernando. Generalized backpropagation,\'{E} tude de cas: Orthogonality. *arXiv preprint arXiv:1611.05927*, 2016. 2, 3

[36] Mete Ozay and Takayuki Okatani. Optimization on submanifolds of convolution kernels in cnns, 2016. 2, 3

[37] Lei Huang, Xianglong Liu, Bo Lang, Adams Wei Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks, 2017. 2, 3

[38] Di Xie, Jiang Xiong, and Shiliang Pu. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 2, 3

[39] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In *Advances in Neural Information Processing Systems*, pages 4266–4276, 2018. 2, 3, 4

[40] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*, 2018. 3

[41] Hongyu Xu, Zhangyang Wang, Haichuan Yang, Ding Liu, and Ji Liu. Learning simple thresholded features with sparse support recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 3

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[43] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 4

[44] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[45] Kai Li, Zhengming Ding, Kunpeng Li, Yulun Zhang, and Yun Fu. Support neighbor loss for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1492–1500. ACM, 2018. 4, 6

[46] Mingyong Zeng, Chang Tian, and Zemin Wu. Person re-identification with hierarchical deep learning feature and efficient xqda metric. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1838–1846. ACM, 2018. 4, 6

[47] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 274–282. ACM, 2018. 4, 6, 7

[48] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. Local convolutional neural networks for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1074–1082. ACM, 2018. 4, 6, 7

[49] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. 4, 6

[50] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 4, 6, 7

[51] Jiawei Liu, Zheng-Jun Zha, Hongtao Xie, Zhiwei Xiong, and Yongdong Zhang. Ca3net. *2018 ACM Multimedia Conference on Multimedia Conference - MM 18*, 2018. 4, 6, 7

[52] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 5, 6, 7

[53] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016. 5

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.

[54] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 6

[55] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 6, 7

[56] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2530–2539, 2017. 6

[57] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6036–6046, 2018. 6, 7

[58] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7073–7082, 2018. 6

[59] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018. 6, 7

[60] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2018. 6

[61] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6886–6895, 2018. 6

[62] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 188–204, 2018. 6

[63] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 486–504, 2018. 6

[64] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 6, 7

[65] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, 2018. 6, 7

[66] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu. Attribute-aware attention model for fine-grained representation learning. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 2040–2048. ACM, 2018. 6

[67] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018. 6, 7

[68] 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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 6, 7

[69] Guodong Ding, Salman Khan, Zhenmin Tang, and Fatih Porikli. Let features decide for themselves: Feature mask network for person re-identification, 2017. 7

[70] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification, 2017. 7

[71] Xiang Bai, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. Deep-person: Learning discriminative deep features for person re-identification, 2017. 7

[72] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person re-identification, 2018. 7

[73] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6

[74] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

Authorized licensed use limited to: University of Exeter. Downloaded on May 08,2020 at 00:27:33 UTC from IEEE Xplore. Restrictions apply.