Orthogonal Center Learning with Subspace Masking for Person Re-Identification

Weinong Wang · Wenjie Pei* · Qiong Cao · Shu Liu · Xiaoyong Shen · Yu-Wing Tai*

Abstract Person re-identification aims to identify whether pairs of images belong to the same person or not. This problem is challenging due to large differences in camera views, lighting and background. One of the mainstream in learning CNN features is to design loss functions which reinforce both the class separation and intra-class compactness. In this paper, we propose a novel Orthogonal Center Learning method with Subspace Masking for person re-identification.

We make the following contributions: (i) we develop a center learning module to learn the class centers by simultaneously reducing the intra-class differences and inter-class correlations by orthogonalization; (ii) we introduce a subspace masking mechanism to enhance the generalization of the learned class centers; and (iii) we devise to integrate the average pooling and max pooling in a regularizing manner that fully exploits their powers. Extensive experiments show that our proposed method consistently outperforms the state-of-the-art methods on the large-scale ReID datasets including Market-1501, DukeMTMC-ReID, CUHK03 and MSMT17.

Keywords Person re-identification · Orthogonal Center Learning · Subspace Masking · average pooling · max pooling.

1 Introduction

The task of person re-identification over images is to identify the same person in different shooting environments such as camera views, person poses and lighting conditions. It is widely applied to surveillance, person tracking sport or other scenarios in which a substantial amount of people may involve. Hence, a robust person re-identification algorithm is required to cope with a large number of person classes.

The state-of-the-art methods for person re-identification focus on either improving the structure of feature learning modules Chang et al (2018); Li et al (2018); Sun et al (2018), or designing more effective loss functions Chen et al (2017); Hadsell et al (2006); Hermans et al (2017)

*Wenjie Pei and Yu-Wing Tai are joint corresponding authors.

W. Wang · W. Pei · Q. Cao · S. Liu · X. Shen · Y. Tai
Youtu X-lab, Tencent
E-mail: weinong.wang@hotmail.com, wenjiecoder@gmail.com, freyaqcao@tencent.com, iushuhust@gmail.com, goodshenxy@gmail.com, yuwingtai@tencent.com

as we do in this work. A typical way of designing loss functions is to combine softmax loss and triplet loss together since their advantages are complementary: softmax loss defines the optimization as a classification problem and tries to classify each individual sample correctly while the triplet loss aims to maximize the relative distance between same-class pairs and different-class pairs.

With a new perspective, the center loss Wen et al (2016) aims to minimize the distances between samples of the same class. It is originally proposed for face recognition but is straightforward to be applied to person re-identification Jin et al (2017); Xiao et al (2019) due to the similar task setting: both are open-set identification tasks (the classes in test set may not appear in training set) with large number of classes. In this paper, we propose a novel orthogonal center learning module to further boost the feature learning procedure. Different from center loss, we formulate the learning objective functions by not only minimizing the distance between each sample to its corresponding center, but also maximizing the separability between samples from different classes. Specifically, we propose to leverage the orthogonalization to reduce the inter-class correlations.

Orthogonal regularization has been widely explored to improve the performance and training efficiency either by easing the gradient vanishing/explosion in Recurrent Neural Networks (RNNs) Arjovsky et al (2016); Vorontsov et al (2017) or stabilizing the distribution of activations for Convolutional networks (CNNs) Bansal et al (2018); Huang et al (2018). It is also used to reduce the correlations among learned features Sun et al (2017); Xie et al (2017); Zhang et al (2017). Inspired by this observation, we propose to apply orthogonalization to decorrelate the class centers which can potentially yield better separability among samples from different classes. Besides, the orthogonality regularization also encourages the full exploitation of the embedding space of class centers.

To further improve the generalization of the class centers and unleash their full potential, we propose a subspace masking mechanism in the center learning module. Specifically, we randomly mask some units of a center embedding to make them disabled and learn the center with the rest of the activated units during training. Thus, this masking mechanism encourages class centers to be representative in their subspaces, which in turn results in more generalizable class centers in full space in test time.

Our proposed center learning module works jointly with the softmax loss and triplet loss and the whole model can be trained in an end-to-end manner. In practice, we parameterize the class centers to involve them into the optimization of the whole model, which is in contrast to the classical center loss: alternately update the class centers and optimize the model parameters. To reduce the computation complexity and mitigate the potential overfitting, the global pooling or max pooling are typically applied in the last layer of convolutional networks. Both global and max poolings have their own merits. We devise a regularizing way in a step-wise learning scheme to integrate these two pooling methods to explore their combined potential.

To summarize, our proposed method benefits from following advantages:

- We propose a center learning module, which learns the class centers by a two-pronged strategy: 1) minimize the intra-class distances and 2) maximize the inter-class separabilities by reducing the inter-class correlations using orthogonalization.
- We propose a subspace masking mechanism to improve the generalization of the class centers.
- We devise a regularizing way to integrate the average pooling and max pooling to fully unleash their combined power.
- Our method outperforms the state-of-the-art methods of person Re-ID on four datasets. Particularly, our method surpasses the state-

of-the-art method by 7.1% by Rank-1 and 10.8% by mAP on CUHK03 dataset.

2 Related Work

There is a large amount of work on person re-identification. Below, we review the most representative methods that are closely related to our proposed method. Most person Re-ID methods follows two lines of research: feature extraction and metric learning.

Feature extraction. Traditional methods typically devise hand-crafted features which are invariant to viewpoint and occlusion Farenzena et al (2010); Gheissari et al (2006); Gray and Tao (2008); Kviatkovsky et al (2013); Liao et al (2015); Ma et al (2012). With the great success of deep neural networks and their strong representation ability, a lot of recent methods Zheng et al (2016) in person Re-ID are developed based on CNN.

In particular, fine-grained part information has been introduced recently to improve the feature representation. Several works Chen et al (2018b); Kalayeh et al (2018); Saquib Sarfraz et al (2018); Su et al (2017) use advanced pose estimation and semantic segmentation Cao et al (2017); Gong et al (2017); Insafutdinov et al (2016); Newell et al (2016); Xiao et al (2018) tools to predict key points explicitly or locate discriminative local regions implicitly. Apart from using existing pose estimator, attention mechanism becomes popular those days for exploiting discriminative local information. A harmonious attention CNN called HA-CNN Li et al (2018) is devised where soft pixel attention and hard regional attention are jointly learned along with simultaneous optimization of feature representation. In Chang et al. Chang et al (2018), MLFN is proposed where the visual appearance of a person is factorized into latent discriminative factors at multiple semantic levels without manual annotation. In Sun et al. Sun et al (2018), the feature map is split into several horizontal parts upon which supervision is imposed for learning part-level features. HPM Fu et al (2018) directly combines the average and max pooling features in each partition to exploit the global and local information. However, direct fusing the features of the average and max pooling operations on the same feature map cannot fully exploit the merits from both pooling methods. To address this limitation, we propose a regularizing way to integrate these two pooling methods.

Metric learning. Metric learning methods aim to enlarge the inter-class distinction while reducing the intra-class variance, which provides a natural solution for both verification and identification tasks. Representative works on person Re-ID include softmax classification loss, contrastive loss Hadsell et al (2006), triplet loss Hermans et al (2017); Wang et al (2018a), quadruplet loss Chen et al (2017), re-ranking Yu et al (2017); Zhong et al (2017a), etc.

Center loss Jin et al (2017); Wen et al (2016, 2019); Xiao et al (2019) is recently proposed to encourage the intra-class compactness and obtains promising performance for face recognition and person Re-ID. Specifically, it learns a center for each class in a mini-batch and penalizes the Euclidean distances between the deep features and their class centers simultaneously. More recently, based on the softmax loss, several multiplicative angular margin-based methods Liu et al (2017, 2016); Wang et al (2018b,c) have been proposed to enhance the discriminative power of the deep features. Different from the classical center loss Wen et al (2016), we not only minimize the distance between each sample to its corresponding center, but also maximize the separability between samples from different classes by orthogonalization to reduce the inter-class correlations. Unlike Zhang et al (2017) which mounts an instance-level global orthogonal regularization upon the triplet loss to push the negative pairs to be orthogonal (in the feature space), our method performs the orthogonal regularization between different class centers to reduce inter-class correlations. Furthermore, we

introduce a subspace masking mechanism to improve the generalization of class centers in subspaces. Different from Dropout Srivastava et al (2014) and DropBlock Ghiasi et al (2018) which perform dropout operations in feature space, our proposed subspace masking mechanism performs masking in the center embedding space to improve the generalization of the learned class centers. We also explore other sampling strategies different from the Bernoulli distribution typically adopted in Dropout Srivastava et al (2014) and DropBlock Ghiasi et al (2018) to show the effectiveness of the proposed subspace masking mechanism.

3 Method

We aim to optimize feature learning in such a way that the distance between intra-class samples is minimized whilst maximizing the separability between inter-class samples. To this end, we propose to learn centers for each class by encouraging each sample to be close to the corresponding class center while reducing the correlations among class centers. Furthermore, we propose a subspace masking mechanism to improve the generalization of class centers in subspaces.

Figure 1 illustrates the architecture of our model. We employ softmax loss and triplet loss Hermans et al (2017) as the basis loss functions to guide the optimization of the feature learning module (ConvNet $\mathcal{C}$), which has been proven effective in Person Re-Id Sun et al (2017); Zheng et al (2018). Our center learning module is proposed to further enhance the optimization jointly with the basis losses.

3.1 Center Learning

Given a training set comprising N samples (images) $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ and their associated class labels $\mathbf{Y} = \{y_i\}_{i=1}^{N}$ categorized into M classes, we first employ a deep ConvNet $\mathcal{C}$ (ResNet-50 He et al (2016) in our implementation) to extract latent feature embeddings denoted as $\mathbf{V} = \{\mathbf{v}_i \in \mathbb{R}^d\}_{i=1}^{N}$. The obtained features $\mathbf{V}$ are then fed into our proposed center learning module and other two basis losses to steer the optimization of parameters in ConvNet $\mathcal{C}$.

Collaborative center learning with softmax loss. A well-learned class center is expected to characterize the samples belonging to this class in the feature space. Intuitively, an optimized center can be calculated as the geometric center of samples belonging to this class in the feature space, which is not feasible since sample features and class centers are optimized dependently on each other. A compromised way Wen et al (2016) is to randomly sample a center position and then iteratively update it using an approximated center position which is calculated as the geometric center of the sample features belonging to this class in each training batch. Hence, sample features and class centers are optimized alternately. A potential drawback of this process is that the class centers are not involved in the optimization by gradient descent of the feature learning (ConvNet C) directly and thus the optimization is inefficient and unstable. To circumvent this issue, we propose to parameterize class centers and optimize them with the ConvNet C jointly.

Specifically, we correspond class centers to parameters $\mathbf{W} \in \mathbb{R}^{d \times M}$ of the linear transformation before the softmax function, which projects feature embeddings from d to M (the number of classes). Each column of $\mathbf{W}$ parameterizes a corresponding class center:

$$\mathbf{c}_i = \mathbf{W}(:, i), \tag{1}$$

where $\mathbf{W}(:, i)$ indicates the i-th column of $\mathbf{W}$. The rationale behind this design is that each column of the transformation matrix $\mathbf{W}$ can be considered as a class embedding to measure the compatibility between this class and the sample feature embeddings by dot product. Thus it is consistent with the intention of our center learning and the class centers $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M) := \mathbf{W}$

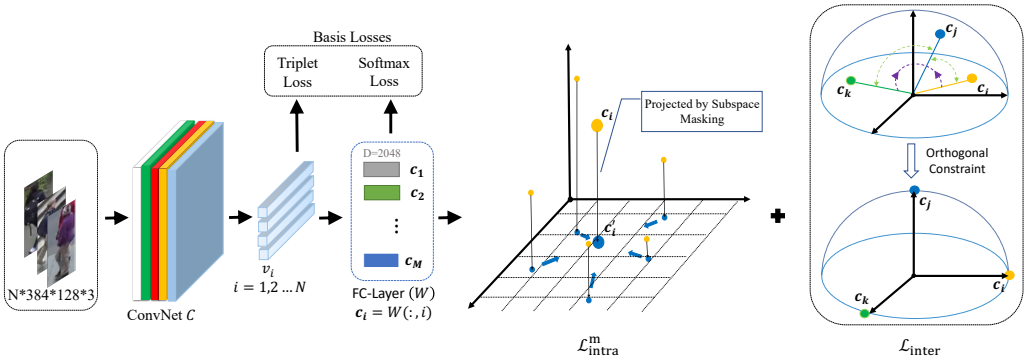

Fig. 1 The architecture of our method. The feature embeddings extracted by ConvNet $\mathcal{C}$ are fed into the basis losses and our center learning module to guide the optimization of the whole model. The center learning module is designed to minimize the intra-class distances ($\mathcal{L}_{\text{intra}}$) and minimize the inter-class correlations by orthogonalization ($\mathcal{L}_{\text{inter}}$). We parameterize the class centers using the linear transformation weights before the softmax loss to perform collaborative learning. We propose a subspace masking mechanism to perform intra-class constraints in subspace ($\mathcal{L}_{\text{intra}}^{m}$) to improve the generalization of class centers.

can be optimized collaboratively by center learning module and softmax loss.

We adopt a two-pronged strategy to guide the optimization of class centers $\mathbf{C}$ in center learning module: minimize intra-class distances and reduce inter-class correlations.

Minimizing intra-class distances. Consider a batch of samples $\{\mathbf{v}\}_{i=1}^{B}$ in a training iteration, we minimize the sum of the Euclidean distance between each sample and its corresponding class center:

$$\mathcal{L}_{\text{intra}} = \sum_{i=1}^{B} \|\mathbf{v}_i - \mathbf{c}_{y_i}\|^2. \tag{2}$$

Reducing inter-class correlations by orthogonalization. We propose to apply orthogonalization to reduce correlations among class centers and thereby increase the separability between samples from different classes. Specifically, we first normalize each class center by L2-norm and then employ a soft orthogonal constraint performed under the standard Frobenius norm in the center learning module:

$$\begin{aligned} \mathbf{c}_i &= \frac{\mathbf{c}_i}{\|\mathbf{c}_i\|}, i = 1, \ldots, M, \\ \mathcal{L}_{\text{inter}} &= \lambda \|\mathbf{C}^\top \mathbf{C} - \mathbf{I}\|_F^2. \end{aligned} \tag{3}$$

Since the optimization of Equation 3 is independent from input samples, it is prone to converge rapidly to a bad local optimum. To make the optimizing process more smooth and synchronize with the optimization of other loss functions, we only apply the orthogonal constraint to the class centers (of samples) involved in the current training batch of each iteration.

Theoretically, a potential flaw of the orthogonal constraint in Equation 3 is that all centers cannot be strictly orthogonal to each other when the number of classes are significantly larger than the dimensions of center embeddings ($M \gg d$). In this case, one feasible solution Bansal et al (2018) is to relax the constraint to minimize the max correlation between any pair of centers, which is equivalent to minimize:

$$\mathcal{L}'_{\text{inter}} = \lambda \|\mathbf{C}^\top \mathbf{C} - \mathbf{I}\|_\infty. \tag{4}$$

In practice, we find that the standard orthogonal loss $\mathcal{L}_{\text{inter}}$ in Equation 3 suffices for the real datasets used in experiments since our aim is to reduce inter-class correlations rather than pursue the strictly orthogonalization between centers.

An alternative way to increase the separability between class centers is to directly maximize the

pairwise Euclidean distance by the Hinge loss:

$$\mathcal{L}_{\text{inter-euclid}} = \sum_{i=1}^{B} \sum_{j=1\&j\neq i}^{B} \max(0, m - \|\mathbf{c}_{y_i} - \mathbf{c}_{y_j}\|). \tag{5}$$

The difference between it and the orthogonalization-based loss in Equation 3 is that $\mathcal{L}_{\text{inter-euclid}}$ performs constraints in the Euclidean space while the orthogonalization operates in the angular space to reduce inter-class correlations. Each has its own merits. Nevertheless, since we adopt the triplet loss as the basis loss which also performs inter-class constraints in Euclidean space, we consider that $\mathcal{L}_{\text{inter-euclid}}$ is not necessary. The follow-up experiments validate our speculation.

3.2 Subspace Masking

We propose a subspace masking mechanism in the center learning module to improve the generalization of the class centers and unleash their full potential. The key idea is to mask some units of center embeddings according to a probability to make them disabled and leave the rest of units activated during training. Thus, it is able to enhance the representation power of the class centers in subspaces. In particular, for each unit of a center embedding we mask it with the probability following the Bernoulli distribution $B(p)$ on the intra-class loss $\mathcal{L}_{\text{intra}}$:

$$\mathcal{L}^{m}_{\text{intra}} = \sum_{i=1}^{B} \sum_{k=1}^{d} B(p) \|\mathbf{v}_i^k - \mathbf{c}_{y_i}^k\|^2, \tag{6}$$

where d is the size of center embeddings (as well as the feature embeddings $\mathbf{v}_i$) and p is the probability of sampling value 1 from Bernoulli distribution. In practice, we handle it as a hyper-parameter and select its value based on a held-out validation set.

The benefits of our subspace masking mechanism are threefold:

- Perspective of center learning: the subspace masking encourages class centers to be physically representative of their corresponding classes in subspaces. Since different subspaces would be randomly selected in different training iterations, the class centers are able to have better generalization in original full space in test time.
- Perspective of feature learning: our subspace masking mechanism also guides the feature learning to be discriminative in subspace. It encourages the model to capture potential discriminative features in local patches.
- Perspective of dropout: By the gradient back-propagation via feature embeddings $\mathbf{v}_i$ in Equation 6 to the ConvNet $\mathcal{C}$, it also has the similar functionality of dropout scheme: train an exponential number of "thinned" networks and aggregate them at test phase.

3.3 Optimization

Given a training set, we optimize the feature learning module ConvNet $\mathcal{C}$ by minimizing our proposed orthogonal center learning losses (L^{m}_{intra} in Equation 6 and $\mathcal{L}_{\text{inter}}$ in Equation 5), jointly with basis losses including softmax loss and triplet loss in an end-to-end manner:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{softmax}} + \alpha_1 \mathcal{L}_{\text{triplet}} + \alpha_2 \mathcal{L}^{m}_{\text{intra}} + \alpha_3 \mathcal{L}_{\text{inter}}, \tag{7}$$

where α_1, α_2 and α_3 are hyper-parameters to balance different losses.

Regularizing feature pooling. Typically, the global average pooling is applied to the last layer of convolutional networks for person Re-ID Sun et al (2018); Wang et al (2018a) to reduce the computation complexity and mitigate the potential overfitting. While the average pooling has been proven to be effective in most cases, a drawback is that it is prone to neutralize the discriminative information which could be captured by max pooling. Actually both average pooling

and max pooling have their own advantages. It would be beneficial to take into account both pooling methods. For instance, a straightforward way Fu et al (2018) is to combine (e.g., add up) the resulting features of two pooling operations on the same feature map and feed the obtained feature to subsequent loss functions. The potential disadvantage of such way is that fusing pooled features by two operations before loss functions may mislead the loss functions during optimization and is hard to learn the desired features that incorporate merits from both average and max poolings.

To circumvent this limitation, we propose a regularizing way to integrate these two pooling methods. Specifically, we employ individual loss functions to learn pooled features for average pooling and max pooling separately. As shown in Figure 2, we split the ConvNet $\mathcal{C}$ into two pathways at the last stage of ResNet-50: one followed with the average pooling and the other followed with the max pooling. Each of them is assigned with an individual triplet loss to learn the correspondingly pooled feature. Meanwhile, two types of pooled features are combined by element-wise averaging operation to be the output feature embeddings of ConvNet $\mathcal{C}$, which are fed into final loss functions presented in Equation 7:

$$\mathbf{v}_i = \frac{\mathbf{v}_i^{AP} + \mathbf{v}_i^{MP}}{2}, \tag{8}$$

where $\mathbf{v}_i^{AP}$ and $\mathbf{v}_i^{MP}$ are the pooled features by average pooling and max pooling respectively for the i-th sample. Refining features by such step-wise supervised learning has been explored before Lee et al (2015); Xie and Tu (2015). Benefited from this step-wise learning scheme, both pooled features are expected to be learned with the desired properties.

4 Experiments

To validate the effectiveness of the proposed method, we conduct experiments on four large person Re-ID benchmarks: Market-1501 Zheng et al (2015), DukeMTMC-ReID Ristani et al (2016); Zheng et al (2017), CUHK03 Li et al (2014) and MSMT17 Wei et al (2018). We first perform ablation studies to investigate the functionality of each component of our method and then compare our method with the state-of-the-art methods on Person Re-ID task.

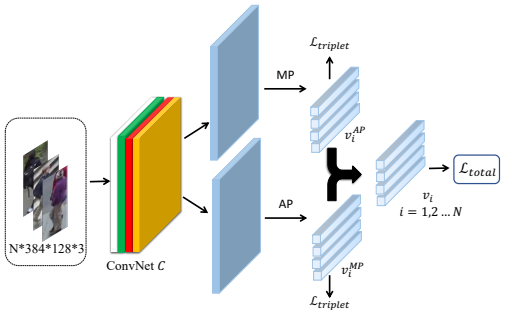

Fig. 2 The illustration of the proposed regularizing way to integrate max pooling and average pooling. We employ individual triplet loss functions to learn pooled features for average pooling and max pooling separately. Meanwhile, the combined pooled features are fed into the final loss functions.

4.1 Datasets and Evaluation Protocol

Market-1501 dataset contains 32668 images captured from six camera views. It includes 12,936 training and 19,732 testing images from 751 and 750 identities respectively.

DukeMTMC-ReID is a subset of the pedestrian tracking dataset DukeMTMC for person Re-ID. It contains 1812 identities captured from 8 cameras, with 16,522 images of 702 persons for training and 19,889 testing images from 1110 persons for testing.

CUHK03 contains 14,097 images from 1,467 identities. Both manually cropped and automatically detected pedestrian images are provided. We follow the recently proposed protocol Yu et al (2017), in which 767 identities are used for training and 700 identities for testing. Our evaluation

Methods		Market-1501		CUHK03		DukeMTMC-ReID	
		mAP	R1	mAP	R1	mAP	R1
Dropout		82.0	92.9	65.6	67.8	73.5	85.4
DropBlock		80.4	92.7	64.8	68.5	72.5	85.2
Subspace Masking	Bernoulli distribution	82.4	93.3	66.4	68.3	**74.4**	85.4
	Weighed sampling	82.5	93.0	66.5	**68.9**	74.2	**85.6**
	Hard-unit sampling	**82.6**	**93.7**	**66.8**	68.6	74.0	**85.6**

Table 1 Comparison of our subspace masking mechanism (using 3 different sampling strategies respectively) with Dropout and DropBlock on three datasets in terms of Rank-1 (R1) and mAP.

is based on the detected label images, which is close to real scenes.

MSMT2017 is currently the largest and most challenging public dataset for person Re-ID. It contains 4101 identities and 126441 bounding boxes, where 32,621 bounding boxes from 1041 identities are used for training and 93,820 bounding boxes of 3060 identities for testing. The raw videos are captured by 15-camera network in both indoor and outdoor scenes, and present large lighting variations.

For performance evaluation, two standard Re-ID evaluation metrics are employed: Cumulative Match Characteristic (CMC) Gray et al (2007) and mean Average Precision (mAP) Zheng et al (2015). For the CMC, we report the Rank-1, Rank-5, Rank-10 accuracies. All results are reported under single-query setting.

$\mathcal{L}^{m}_{\text{intra}}$	gor	$\mathcal{L}_{\text{inter}}$	Market-1501		CUHK03		DukeMTMC-ReID	
			mAP	R1	mAP	R1	mAP	R1
√			82.4	93.3	66.4	68.3	74.4	85.4
√	√		81.4	92.5	65.2	67.4	73.9	85.6
√		√	**83.3**	**93.5**	**67.5**	**70.4**	**74.6**	**86.4**

Table 2 Comparison of our proposed $\mathcal{L}_{\text{inter}}$ with GOR Zhang et al (2017) on three datasets in terms of Rank-1 (R1) and mAP.

4.2 Implementation Details

We adopt Resnet-50 He et al (2016) pretrained on ImageNet Deng et al (2009) as our feature learning module ConvNet $\mathcal{C}$. Following Sun et al. Sun et al (2018), we remove the spatial downsampling operation of the last stage in ConvNet $\mathcal{C}$ to preserve more fine-grained information. The learned feature embeddings of ConvNet $\mathcal{C}$ further go through a Batch Normalization Ioffe and Szegedy (2015) followed by the LeakyReLU before fed into loss functions. The input images are preprocessed by resizing them to 384×128 and horizontal flipping, normalization and random erasing Zhong et al (2017b) are used for data augmentation.

We randomly select 16 persons with 4 images for each person for each batch during training, resulting in a batch size of 64. To make the training at the early stage more stable, we utilize the gradual warming up strategy Goyal et al (2017). Adam Kingma and Ba (2014) is employed with the weight decay of 1e-4 for gradient descent optimization. The training process lasts for 400 epoches and the learning rate starts from 0.001 and decreases by 0.1 at $\{80, 180, 300\}$ epochs. The hyper-parameters α_1, α_2 and α_3 are validated on a held-out validation set. To evaluate our model, we provide a customized *baseline* which also applies ConvNet $\mathcal{C}$ for feature extraction while using softmax loss and triplet loss (basis losses in our model) to guide the optimization.

4.3 Ablation Study

We first perform quantitative evaluation to investigate the effect of each component of our center learning module. To this end, we conduct ablation

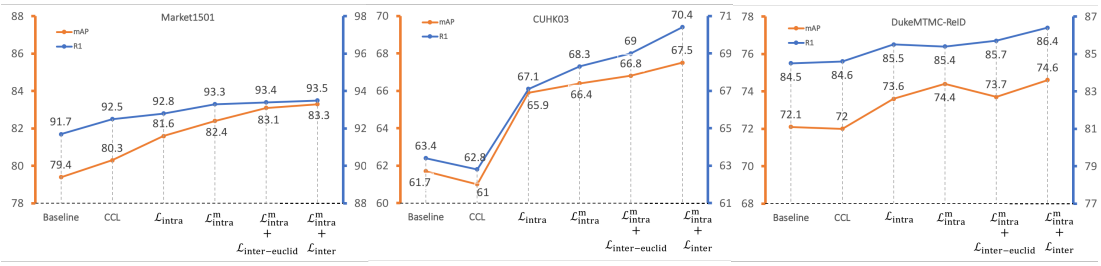

Fig. 3 Performance for different combinations of loss functions on Market-1501, DukeMTMC-ReID and CUHK03 in terms of Rank-1 (R1) and mAP. We incrementally augment the loss function with proposed $\mathcal{L}_{\text{intra}}$, $\mathcal{L}^{m}_{\text{intra}}$ and $\mathcal{L}_{\text{inter}}$. Besides, we also evaluate the classical center loss(CCL) and $\mathcal{L}_{\text{inter-euclid}}$ in Equation 5. Note that the proposed regularized pooling method is not applied in this set of experiments.

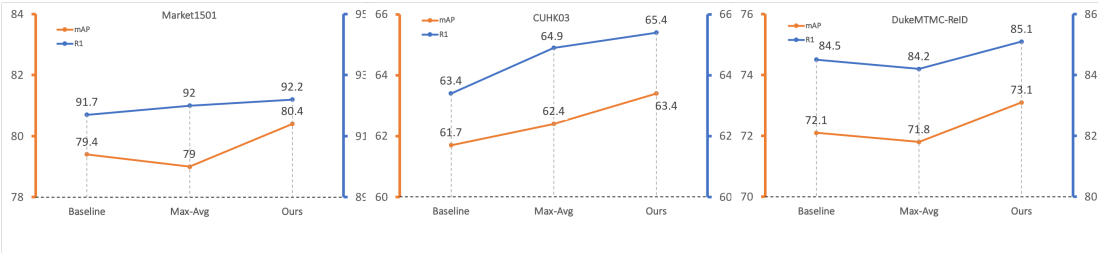

Fig. 4 Performance for different ways of pooling methods on Market-1501, DukeMTMC-ReID and CUHK03 in terms of Rank-1 (R1) and mAP. Herein, *Baseline* utilizes max pooling and *Max-Avg* corresponds to the pooling methods used in HPM Fu et al (2018). Note that we only employ the basis losses (softmax loss and triplet loss) for optimization in this set of experiments.

experiments which begin with the customized *baseline* (using only basis losses) and then incrementally augments loss functions with the proposed $\mathcal{L}_{\text{intra}}$, $\mathcal{L}^{m}_{\text{intra}}$ and $\mathcal{L}_{\text{inter}}$. Besides, we also evaluate the classical center loss(CCL) Wen et al (2016) and $\mathcal{L}_{\text{inter-euclid}}$ in Equation 5 (which maximizes inter-class distance in Euclidean space) for comparison. Figure 3 presents the experimental results on Market-1501, DukeMTMC-ReID and CUHK03.

Effect of $\mathcal{L}_{\text{intra}}$. Compared to the *baseline* using only basis losses, our proposed $\mathcal{L}_{\text{intra}}$ improves the performance by a large margin, especially on CUHK03. It shows the robustness and effectiveness of $\mathcal{L}_{\text{intra}}$. In contrast, the classical center loss (CCL) only boosts the performance upon the baseline on Market-1501. Thus, it validates that parameterizing the class centers with weights of the linear transformation before softmax loss is beneficial for collaborative training between the center learning and softmax loss.

Effect of subspace masking ($\mathcal{L}^{m}_{\text{intra}}$). Figure 3 shows that employing subspace masking $\mathcal{L}^{m}_{\text{intra}}$ outperforms $\mathcal{L}_{\text{intra}}$ on all three datasets, which indicates that optimizing class centers and feature learning in subspace via intra-class constraints is indeed able to further improve the performance.

Typically we sample the masking units following the Bernoulli distribution. To further investigate the effect of different sampling strategies, we also explore two more sampling protocols: **Weighted sampling** which samples the unmasked units according to the probability proportional to the euclidean intra-class distance of the corresponding units and **Hard-unit sampling** which directly selects the units with large euclidean intra-class distance (corresponding to hard units). We compare between these three different sampling strategies as well as Dropout Srivastava et al (2014) and

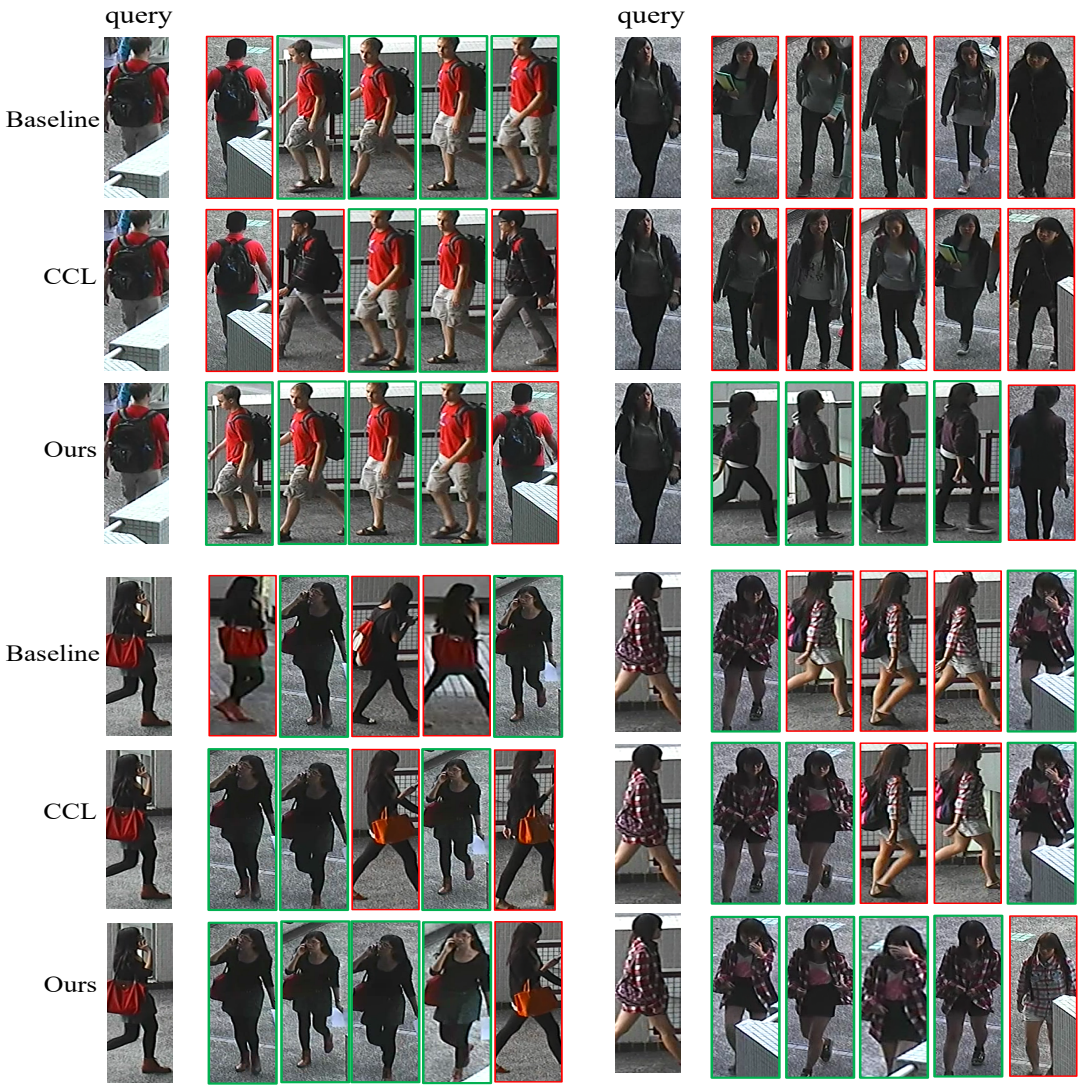

Fig. 5 Four groups of challenging samples of CUHK03 test data. For each group, we list rank-5 retrieved images based on query. Green bounding boxes indicate correct results and red ones correspond to false results.

DropBlock Ghiasi et al (2018) in Table 1. We observe that our subspace masking with any of 3 sampling strategies consistently outperforms Dropout and DropBlock which indicates its advantages over other two methods. Besides, there is not much performance difference between 3 sampling strategies, thus our subspace masking mechanism is not sensitive to the selection of sampling strategy.

Effect of $\mathcal{L}_{\text{inter}}$ based on orthogonalization. $\mathcal{L}_{\text{inter}}$ is expected to reduce the inter-class correlation by orthogonalization. Adding $\mathcal{L}_{\text{inter}}$ to loss functions achieves additional performance gain compared to using *baseline* and $\mathcal{L}^{m}_{\text{intra}}$. Another interesting observation is that adding $\mathcal{L}_{\text{inter-euclid}}$

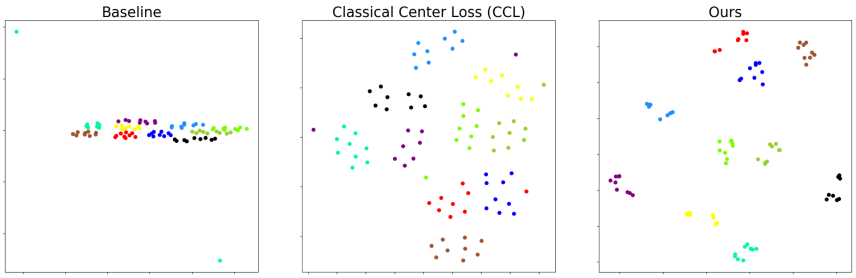

Fig. 6 t-SNE maps of CUHK03 test data from 10 randomly selected classes, constructed by the ConvNet $\mathcal{C}$ supervised by respectively the basis losses (baseline), classical center loss (CCL) and our proposed loss (ours). The points with different colors refer to different classes.

makes little contribution to the performance. We surmise that this is because we adopt the triplet loss as the basis loss which also performs inter-class constraints in Euclidean space, hence $\mathcal{L}_{\text{inter-euclid}}$ is not necessary anymore.

Furthermore, We conduct experiments to compare our method with GOR Zhang et al (2017) which performs orthogonal regularization for negative pairs in triplet loss. The results presented in Table 2 shows that it is helpless for the overall performance and worse than the performance of our proposed $\mathcal{L}_{\text{inter}}$.

Effect of Regularizing feature pooling. Next we perform ablation study to investigate our proposed regularized pooling method, which aims to explore the full potential of both average pooling and max pooling. We compare with the pooling method (denoted as Max-Avg) used in HPM Fu et al (2018) which simply adds up the resulting features of two pooling operations on the same feature map. Figure 4 presents the experimental results. We observe that our proposed pooling method achieves remarkable performance gain over *baseline* and outperforms HPM by a large margin. It demonstrates the effectiveness of our pooling method, which employs a step-wise learning scheme to assign an individual triplet loss for both max and average poolings.

Random Erasing	Method	R1	R5	R10	mAP
No	MLFN Chang et al (2018)	52.8	–	–	47.8
	HA-CNN Li et al (2018)	41.7	–	–	38.6
	PCB+RPP Sun et al (2018)	63.7	80.6	86.9	57.5
	RB Ro et al (2019)	52.9	–	–	47.4
	HPM Fu et al (2018)	63.9	79.7	86.1	57.5
	ours	**71.0**	**85.1**	**90**	**68.3**
Yes	DaRe Wang et al (2018e)	61.6	–	–	58.1
	Mancs Wang et al (2018a)	65.5	–	–	60.5
	ours	**80.5**	**91.9**	**95.1**	**77.9**

Table 3 Comparison of our method with state-of-the-arts on CUHK03 in terms of Rank-1 (R1), Rank-5 (R5), Rank-10 (R10) and mAP. Note that we only compare with the methods which follow the newly proposed protocol Yu et al (2017).

Random Erasing	Method	R1	R5	R10	mAP
No	SVDNet Sun et al (2017)	82.3	92.3	95.2	62.1
	BraidNet-CS+SRL Wang et al (2018d)	83.7	–	–	69.8
	MLFN Chang et al (2018)	90	–	–	74.3
	HA-CNN Li et al (2018)	91.2	–	–	75.7
	SPReID$_{\text{combined-ft}}$ Kalayeh et al (2018)	92.54	97.15	98.1	81.34
	PABR Suh et al (2018)	91.7	96.9	98.1	79.6
	PCB+RPP Sun et al (2018)	93.8	**97.5**	98.5	81.6
	RB Ro et al (2019)	91.2	–	–	77.0
	HPM Fu et al (2018)	94.2	**97.5**	98.5	82.7
	ours	**94.3**	**97.5**	**98.7**	**83.6**
Yes	DaRe Wang et al (2018e)	88.5	–	–	74.2
	GSRW Shen et al (2018)	92.7	96.9	98.1	82.5
	CRF-GCL Chen et al (2018a)	93.5	97.7	–	81.6
	Mancs Wang et al (2018a)	93.1	–	–	82.3
	ours	**94.6**	**98.3**	**99.0**	**87.4**

Table 4 Comparison of our method with state-of-the-arts on Market-1501 in terms of Rank-1 (R1), Rank-5 (R5), Rank-10 (R10) and mAP.

4.4 Qualitative Evaluation

We conduct experiments on CUHK03 to show the ability of our proposed method to compact sam-

ples within each class as well as separate samples from different classes. To this end, we apply t-SNE Maaten and Hinton (2008) on feature embeddings output by ConvNet $\mathcal{C}$, and visualize the t-SNE maps learned by the baseline (softmax + triplet loss), classical center loss (CCL) and our proposed loss in Figure 6. It is obvious that CCL improves the baseline, and our proposed loss significantly enhances the compactness within the same class and the dissociation of different classes over baseline and CCL.
Besides, we present four groups of challenging examples of CUHK03 test set in Figure 5 to show that our method is more powerful than CCL and baseline.

4.5 Comparison with State-of-the-arts

We conduct experiments on Market-1501, DukeMTMC-ReID, CUHK03 and MSMT2017, and compare with the state-of-the-art person Re-ID methods including the harmonious attention HA-CNN Li et al (2018), the multi-task attentional network with curriculum sampling Mancs Wang et al (2018a), the part-aligned bilinear representations PABR Suh et al (2018), the horizontal pyramid matching apporach HPM Fu et al (2018), and other methods Chang et al (2018); Kalayeh et al (2018); Ro et al (2019); Shen et al (2018); Sun et al (2017, 2018); Wang et al (2018d,e). All four popular evaluation metrics including Rank-1, Rank-5, Rank-10 and mAP are reported. Since random erasing is a fairly effective way of data augmentation which typically leads to a large performance gain. Hence we conduct experiments in two settings: with or without random erasing.

Evaluation on CUHK03. We first conduct the experiments on CUHK03 Li et al (2014) with auto-detected perdestrian bounding boxes to compare our method to the state-of-the-art methods for person Re-ID. The comparison results are shown in Table 3. Our method performs best on all four metrics and surpasses other state-of-the-arts on significantly. In particular, our method outperforms the second best model HPM by 7.1% on Rank-1 and 10.8% on mAP, which illustrates the substantial superiority of our proposed method over other methods.

Method	Gallary size							
	19732		119732		219732		519732	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Zheng *et al* Zheng et al (2018)	79.5	59.9	73.8	52.3	71.5	49.1	68.3	45.2
APR Lin et al (2017)	84	62.8	79.9	56.5	78.2	53.6	75.4	49.8
TriNet Hermans et al (2017)	84.9	69.1	79.7	61.9	77.9	58.7	74.7	53.6
PABR Suh et al (2018)	91.7	79.6	88.3	74.2	86.6	71.5	84.1	67.2
ours	**94.3**	**83.6**	**91.2**	**78.3**	**89.6**	**76**	**88**	**72.3**

Table 5 Comparision of our method with state-of-the-arts on the Market-1501+500k. Experiments are performed on four different sizes of gallery sets. Larger gallery sets has more distractors and thus is more challenging.

Evaluation on Market-1501. Table 4 reports the comparison results on Market-1501 Zheng et al (2015). Our method achieves the best performance among all metrics in both two settings (with or without random erasing), which indicates the superiority of our method. Note that HPM utilizes both original images and flipped images to extract features and combines them in test phase, which is not used by other methods.

Furthermore, we perform experimental comparison over an expanded dataset with additional 500K distractors. Table 5 reports Rank-1 accuracy and mAP over four with different sizes of gallery sets containing $19,732$, $119,732$, $219,732$, and $519,732$ images respectively. Our method consistently outperforms other methods by a large margin across different gallery sets, which implies the robustness of our method.

Evaluation on DukeMTMC-ReID. Table 6 lists the experimental results of our method and the state-of-the-arts on DukeMTMC-ReID Ristani et al (2016); Zheng et al (2017) dataset. Our method performs best on rank-5, rank-10 and mAP and ranks the second place on Rank-1 in the setting without random erasing. HPM achieves best on Rank-1 a5d performs slightly better than ours. In the setting with random erasing, our model substantially outperforms other models.

Random Erasing	Method	R1	R5	R10	mAP
No	SVDNet Sun et al (2017)	76.7	86.4	89.9	56.8
	BraidNet-CS+SRL Wang et al (2018d)	76.44	–	–	59.49
	MLFN Chang et al (2018)	81.0	–	–	62.8
	HA-CNN Li et al (2018)	80.5	–	–	63.8
	$\text{SPReID}_{\text{combined-ft}}$ Kalayeh et al (2018)	84.43	91.88	93.72	70.97
	PABR Suh et al (2018)	84.4	92.2	93.8	69.3
	PCB+RPP Sun et al (2018)	83.3	90.5	92.5	69.2
	RB Ro et al (2019)	82.4	–	–	66.6
	HPM Fu et al (2018)	**86.6**	93	95.1	74.3
	ours	86.4	**93.6**	**95.5**	**74.6**
Yes	DaRe Wang et al (2018e)	79.1	–	–	63.0
	GSRW Shen et al (2018)	80.7	88.5	90.8	66.4
	CRF-GCL Chen et al (2018a)	84.9	92.3	–	69.5
	Mancs Wang et al (2018a)	84.9	–	–	71.8
	ours	**87.7**	**94.1**	**96.1**	**79.0**

Table 6 Comparison of our method with state-of-the-arts on DukeMTMC-ReID in terms of Rank-1 (R1), Rank-5 (R5), Rank-10 (R10) and mAP.

Evaluation on MSMT2017. MSMT17 Wei et al (2018) is currently the largest and most challenging public dataset for person Re-ID. Since it is newly released, hence there is not many baseline models for comparison. We provide in Table 7 the results of our method and the baselines reported by MSMT17 Wei et al (2018). Our method beats the baselines by a significant margin. Particularly, compared to the GLAD Wei et al (2018) which performs the second place, our method gains 15.2% and 17.7% on Rank-1 and mAP respectively. This observation validates the scalability and robustness of our method in large-scale scenes.

Method	R1	R5	R10	mAP
GoogleNet Wei et al (2018)	47.6	65.0	–	23
PDC Wei et al (2018)	58.0	73.6	–	29.7
GLAD Wei et al (2018)	61.4	76.8	–	34
ours	**76.8**	**86.8**	**90.1**	**51.7**
ours*	78.8	88.8	91.6	57.0

Table 7 Performance of our method and other baseline models on MSMT17 in terms of Rank-1 (R1), Rank-5 (R5), Rank-10 (R10) and mAP. We also provide the results (the line denoted as ours*) in the setting with random erasing.

5 Conclusion

In this work, we have presented a novel orthogonal center learning module to learn the class centers with subspace masking for person re-identification. We formulate its learning objective by minimizing the intra-class distances and reducing the inter-class correlations via orthogonalization. Then, a subspace masking mechanism is introduced to further improve the generalization of the learned class centers. Besides, we propose a regularized way to combine the average pooling and max pooling to fully unleash their combined power. Our model surpasses the state-of-the-art work on the challenging person Re-ID datasets including Market-1501, DukeMTMC-ReID, CUHK03 and MSMT17.

References

Arjovsky M, Shah A, Bengio Y (2016) Unitary evolution recurrent neural networks. In: ICML

Bansal N, Chen X, Wang Z (2018) Can we gain more from orthogonality regularizations in training deep networks? In: NeurIPS

Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR

Chang X, Hospedales TM, Xiang T (2018) Multi-level factorisation net for person re-identification. In: CVPR

Chen D, Xu D, Li H, Sebe N, Wang X (2018a) Group consistent similarity learning via deep crf for person re-identification. In: CVPR

Chen D, Zhang S, Ouyang W, Yang J, Tai Y (2018b) Person search via a mask-guided two-stream cnn model. In: ECCV

Chen W, Chen X, Zhang J, Huang K (2017) Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: CVPR

Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: CVPR

Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z, Huang T (2018) Horizontal pyramid matching for person re-identification. arXiv:180405275

Gheissari N, Sebastian T, Hartley R (2006) Person reidentification using spatiotemporal appearance. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA

Ghiasi G, Lin TY, Le QV (2018) Dropblock: A regularization method for convolutional networks. In: Advances in Neural Information Processing Systems, pp 10,727–10,737

Gong K, Liang X, Zhang D, Shen X, Lin L (2017) Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR

Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, Tulloch A, Jia Y, He K (2017) Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv:170602677

Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV

Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol 3, pp 1–7

Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: CVPR

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR

Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:170307737

Huang L, Liu X, Lang B, Yu AW, Wang Y, Li B (2018) Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In: AAAI

Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B (2016) Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:150203167

Jin H, Wang X, Liao S, Li SZ (2017) Deep person re-identification with improved embedding and efficient training. In: IJCB

Kalayeh MM, Basaran E, Gökmen M, Kamasak ME, Shah M (2018) Human semantic parsing for person re-identification. In: CVPR

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:14126980

Kviatkovsky I, Adam A, Rivlin E (2013) Color invariants for person reidentification. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(7):1622–1634

Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z (2015) Deeply-supervised nets. In: Artificial Intelligence and Statistics

Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR

Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: CVPR

Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: CVPR

Lin Y, Zheng L, Zheng Z, Wu Y, Yang Y (2017) Improving person re-identification by attribute and identity learning. arXiv:170307220

Liu W, Wen Y, Yu Z, Yang M (2016) Large-margin softmax loss for convolutional neural networks. In: ICML

Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: Deep hypersphere embedding for face recognition. In: CVPR

Ma B, Su Y, Jurie F (2012) Local descriptors encoded by fisher vectors for person re-identification. In: ECCV

Maaten Lvd, Hinton G (2008) Visualizing data using t-sne. Journal of machine learning research 9(Nov):2579–2605

Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: ECCV

Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV

Ro Y, Choi J, Jo DU, Heo B, Lim J, Choi JY (2019) Backbone can not be trained at once: Rolling back to pre-trained network for person re-identification. AAAI

Saquib Sarfraz M, Schumann A, Eberle A, Stiefelhagen R (2018) A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: CVPR

Shen Y, Li H, Xiao T, Yi S, Chen D, Wang X (2018) Deep group-shuffling random walk for person re-identification. In: CVPR

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1):1929–1958

Su C, Li J, Zhang S, Xing J, Gao W, Tian Q (2017) Pose-driven deep convolutional model for person re-identification. In: ICCV

Suh Y, Wang J, Tang S, Mei T, Mu Lee K (2018) Part-aligned bilinear representations for person re-identification. In: ECCV

Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. In: ICCV

Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV

Vorontsov E, Trabelsi C, Kadoury S, Pal C (2017) On orthogonality and learning recurrent networks with long term dependencies. In: ICML

Wang C, Zhang Q, Huang C, Liu W, Wang X (2018a) Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: ECCV

Wang F, Cheng J, Liu W, Liu H (2018b) Additive margin softmax for face verification. IEEE Signal Processing Letters 25(7):926–930

Wang H, Wang Y, Zhou Z, Ji X, Gong D, Zhou J, Li Z, Liu W (2018c) Cosface: Large margin cosine loss for deep face recognition. In: CVPR

Wang Y, Chen Z, Wu F, Wang G (2018d) Person re-identification with cascaded pairwise convolutions. In: CVPR

Wang Y, Wang L, You Y, Zou X, Chen V, Li S, Huang G, Hariharan B, Weinberger KQ (2018e) Resource aware person re-identification across multiple resolutions. In: CVPR

Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: CVPR

Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: ECCV

Wen Y, Zhang K, Li Z, Qiao Y (2019) A comprehensive study on center loss for deep face recognition. International Journal of Computer Vision pp 1–16

Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. In: ECCV

Xiao J, Xie Y, Tillo T, Huang K, Wei Y, Feng J (2019) Ian: the individual aggregation network for person search. Pattern Recognition 87:332–340

Xie D, Xiong J, Pu S (2017) All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In: CVPR

Xie S, Tu Z (2015) Holistically-nested edge detection. In: ICCV

Yu R, Zhou Z, Bai S, Bai X (2017) Divide and fuse: A re-ranking approach for person re-identification. arXiv:170804169

Zhang X, Yu FX, Kumar S, Chang SF (2017) Learning spread-out local feature descriptors. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4595–4603

Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: ICCV

Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: Past, present and future. arXiv:161002984

Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV

Zheng Z, Zheng L, Yang Y (2018) A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14(1):13

Zhong Z, Zheng L, Cao D, Li S (2017a) Re-ranking person re-identification with k-reciprocal encoding. In: CVPR

Zhong Z, Zheng L, Kang G, Li S, Yang Y (2017b) Random erasing data augmentation. arXiv:170804896