

Joint Discriminative and Metric Embedding Learning for Person Re-Identification

Sinan I. Sabri^{1,2}, Zaigham A. Randhawa¹, and Gianfranco Doretto¹

¹ West Virginia University, Morgantown, WV 26506, USA

{sisabri,zar00002,gidoretto}@mix.wvu.edu

² University of Misan, Amarah, Maysan, Iraq

Abstract. Person re-identification is a challenging task because of the high intra-class variance induced by the unrestricted nuisance factors of variations such as pose, illumination, viewpoint, background, and sensor noise. Recent approaches postulate that powerful architectures have the capacity to learn feature representations invariant to nuisance factors, by training them with losses that minimize intra-class variance and maximize inter-class separation, without modeling nuisance factors explicitly. The dominant approaches use either a discriminative loss with margin, like the softmax loss with the additive angular margin, or a metric learning loss, like the triplet loss with batch hard mining of triplets. Since the softmax imposes feature normalization, it limits the gradient flow supervising the feature embedding. We address this by joining the losses and leveraging the triplet loss as a proxy for the missing gradients. We further improve invariance to nuisance factors by adding the discriminative task of predicting attributes. Our extensive evaluation highlights that when only a holistic representation is learned, we consistently outperform the state-of-the-art on the three most challenging datasets. Such representations are easier to deploy in practical systems. Finally, we found that joining the losses removes the requirement for having a margin in the softmax loss while increasing performance.

Keywords: Person Re-Identification · Discriminative Learning · Metric Learning.

1 Introduction

Person re-identification [2,19,26,44] is the task of assigning the same identity to tightly cropped images of people, based solely on their whole body appearance information. The problem is challenging because distinct images of the same person may look very different, since no restrictions are imposed on the nuisance factors of variation, such as pose, illumination, viewpoint, background, and sensor noise, causing a high intra-class variance.

In order to address that challenge, the research landscape has evolved from developing feature-based models [8,21] coupled with metric learning [38], to developing dedicated deep learning architectures [48] trained with classification and verification losses [18], to developing specialized deep learning schemes [47,41]

aiming at extracting more robust feature embeddings by leveraging powerful pretrained backbone architectures like ResNet-50 [11]. In addition to that, some recent works [20,43,33,4,17] used person attributes such as gender, upper and lower body clothing colors, carrying handbag and backpack as a powerful complementary information to improve the performance of person re-identification. These attributes have more discriminative information about person images, and are invariant to nuisance factors, and they could help with coping with intra-class variations.

Among the more recent trends, there has been also the idea of learning feature embeddings directly suitable for re-identification, by improving the ability to control how losses deal with intra-class and inter-class variances, while giving much less importance to the explicit modeling of the nuisance factors of variation. These approaches focus on learning a holistic representation of the image of a person. They are simpler to deploy, and incorporate in a retrieval system. There are two main lines of work. The first one has focussed on improving the triplet loss derived from metric learning [39,13]. The second line of work has focused on improving the softmax loss used for classification, via normalizing weights and representations [23,36,6]. However, we note that restricting the embeddings to live on a hypersphere limits the gradient flow supervising the embedding under training, which could potentially generate a performance gap.

In this work, we improve the learning of a holistic representation in the form of a feature embedding for person re-identification. Inspired by the previous observation, we do so by incorporating the latest findings on the softmax and triplet losses in a revised combination of such losses, which includes also the learning of multiple discriminative tasks, given by the identity classification and the prediction of attributes. The intent is for the triplet loss to help the softmax further decrease intra-class variation, and increase inter-class distance by letting the triplet loss be the proxy for the gradient supervision that the embedding normalization has restricted, and we specify under what conditions this may happen. We also observe that the same strategy used to form the batch of triplets can be used in tandem with the softmax loss to prevent issues due to dataset imbalance, which are common in person re-identification. In addition, we add the discriminative task of learning attributes to further increase robustness against nuisance factors. We perform an extensive evaluation of the proposed combination of losses with and without using person attributes on the latest person re-identification datasets. We found that this approach can achieve competitive performance with the state-of-the-art, and that the combined loss does not require the softmax component to use any margin.

2 Related Work

Person re-identification is a challenging task due to the nuisance factors of variation, such as pose, illumination, viewpoint, background clutter, spatial misalignment, etc. There is a large literature in this area [42], and two predominant

directions for tackling the problem are based on metric learning [2,14,19,26,44], and discriminative feature representation learning [8,21,24].

Recent works use deep learning to learn robust feature representations [42]. [40] proposes to jointly learn features from multiple domains and then finetuning with domain guided drop out for the specific domain. [1] offers a deep convolutional architecture trained on pairs of images capable of learning features and similarity metric simultaneously. [3] combines the CRF model with DNN to learn more consistent multi-scale similarity metrics for person re-identification. [32] employs partition strategy on convolutional features, and [7] learns embedding of the person image on a hypersphere manifold using a spherical loss. Differently from [7], our proposed model uses a simpler architecture and focusses on combining a margin based softmax loss with a triplet loss to expand feature embedding hypothesis. [45] detects body regions that are discriminative for person re-identification, while [16] learns full body and body parts features through a multi scale context aware network. [15] addresses the limitation of CNNs in representing person images with large variations in body pose and scale by proposing a module to conclude the receptive fields according to the pose and scale of the input person image. [41] explores diverse discriminative visual cues without the assistance of pose estimation and human parsing, and [34] proposes a Fully Attention Block (FAB) plugged into a CNN to overcome the misalignment problem and to localize discriminative local features. Generative adversarial networks (GANs) [10] have been used in person re-identification. [9,27,47] aim at decoupling pose information from image features via adversarial learning.

Person attributes have been used in person re-identification leading to improved robustness against variation of viewpoint, illumination and pose. [20] manually annotated person attributes for the Market-1501 [46] dataset and the DukeMTMC-ReID [49] dataset. It proposes the attribute-person recognition (APR) network. [43] transforms attribute recognition from a high level layer to a mid level layer, and [22] jointly learns appearance and attribute representations via multi-task learning. To learn discriminative person body parts, [33] utilizes person attribute information by integrating attribute features with identity and body part classification. [4] proposes a multi task network to learn identity part-level representation and an attribute global representation. [17] uses person attributes to detect attribute body parts or handle body parts misalignment.

3 Proposed Approach

For the person re-identification task, given a tightly cropped image sample of a person, I , we are seeking to learn a feature embedding $f_{\theta}(I)$, defined by the set of parameters θ , which is as invariant as possible to the nuisance factors of variations. Rather than attempting to model nuisance factors, current deep neural network architectures have shown the promise to cope with their effects, by shifting the focus on designing clever training practices, as well as loss functions.

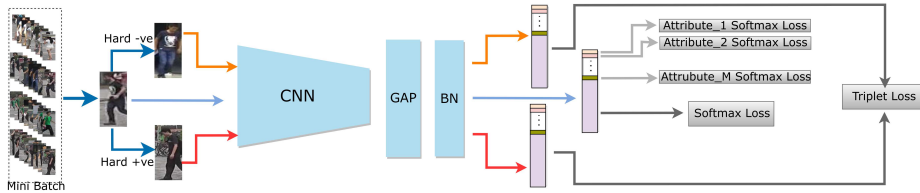


Fig. 1. Architecture. Simple graphical description of the joint optimization of the multi-task re-identification loss, based on a multi-class classification (2) (for the identities), a multi-label classification (3) (for attributes), and a metric learning loss (5).

Here we intend to further explore this holistic-based approach and shed light on additional aspects of this line of work.

3.1 Classification Losses

A successful strategy for learning the embedding f_θ is through the use of a classification loss such as the categorical cross-entropy, which entails adding a softmax layer after the embedding. This leads to the loss

$$\mathcal{L}_S(\theta, W, b) = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{W_{y_i}^\top x_i + b_{y_i}}}{\sum_{j=1}^n e^{W_j^\top x_i + b_j}}, \quad (1)$$

where $x_i = f_\theta(I_i) \in \mathbb{R}^d$ is the embedding of I_i , which has identity y_i . Moreover, $W = [W_1, \dots, W_c] \in \mathbb{R}^{d \times c}$ and $b = [b_1, \dots, b_c]$ are the weights and biases of the softmax layer, while N is the batch size.

Given two images I_i and I_j of the same identity, i.e., $y_i = y_j$, the softmax loss (1) will strive to make the target logit in position y_i be the highest for both images. While this should encourage $f_\theta(I_i)$ and $f_\theta(I_j)$ to be close, in general, there is not an explicit effort to impose $f_\theta(I_i) = f_\theta(I_j)$. This leads to a performance gap, given the large intra-class variability of the person re-identification task due to nuisance factors, which easily cause identity miss-classifications.

Within the context of face recognition, the issue above has been mitigated by taking several steps. First, every logit is produced by comparing the input against ℓ_2 -normalized weights [23,36,35], i.e. $\|W_j\| = 1$. This reduces by one the degrees of freedom by which two different logits could become equal, when activated by images I_i and I_j respectively, each of which depicting the same identity, i.e., $y_i = y_j$. Second, the input of every logit is also ℓ_2 -normalized [36], i.e. $\|x_i\| = 1$, and rescaled to a temperature value s . This further reduces the degrees of freedom by which different logits could become equal, by imposing the embeddings to be defined on the hypersphere of radius s . Also, this suggests using cosine similarity as the metric for comparison between inputs and weights.

While input and weight normalizations positively contribute towards reducing intra-class variability, it is possible to further reduce the spread of the embeddings of the samples with same identity. This is done by introducing a margin in

the cosine similarity, $\cos(\alpha)$, of the target logit. Doing so would further pull the embeddings closer to make up for the loss of similarity induced by the margin. There are at least three basic ways to add a cosine similarity margin [23,6,36]. In [6] it is shown that for face recognition the *additive angular* margin m is the most effective, which reduces (1) to

$$\mathcal{L}_{AM}(\theta, W) = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s \cos(\alpha_{y_i} + m)}}{e^{s \cos(\alpha_{y_i} + m)} + \sum_{j \neq y_i} e^{s \cos \alpha_j}}, \quad (2)$$

where we have set $b = 0$ for simplicity, as in [23]. In the experiments we explore the effectiveness of (2) for person re-identification. Its action should be to minimize the intra-class variation, while the denominator attempts to maximize the inter-class discrepancy by distancing the weights on the unit hypersphere.

We further push the training of the embedding to become invariant to the nuisance factors by leveraging the attribute labels. Assuming that a person image I is described by M binary attributes, the original embedding $f_\theta(I)$ is now split into the inputs of two heads, $f_{\theta_{id}}$ and f_{θ_a} , for predicting identity and attributes, respectively. The first input $f_{\theta_{id}}$ will be trained according to (2), which we indicate more specifically as $\mathcal{L}_{AM_{id}}(\theta_{id}, W_{id})$.

The second input f_{θ_a} for attribute prediction is still normalized and the head weights are normalized as well to minimize intra-class variability and maximize the correct prediction of the attributes. However, since every attribute is binary (i.e., present or not), in order to implement the normalization strategy, and leverage the additive angular margin loss [6], we cannot treat this as a multi-label problem where we use the binary cross-entropy loss for every attribute. Instead, we must use a categorical cross-entropy loss for every attribute where the number of categories is 2. Therefore, the corresponding loss for this pool of M classifiers becomes

$$\mathcal{L}_{AM_{attr}}(\theta_a, W_a) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^M \log \frac{e^{s \cos(\alpha_{a_{k,i}} + m)}}{e^{s \cos(\alpha_{a_{k,i}} + m)} + e^{s \cos \bar{\alpha}_{a_{k,i}}}}, \quad (3)$$

where $\bar{\alpha}_{a_{k,i}}$ indicates the input to the cosine corresponding to the option where the attribute is not present. The final multi-task classification loss imposed on the identities and attributes is the sum of $\mathcal{L}_{AM_{id}}(\theta_{id}, W_{id})$ and (3). Also, note that θ_{id} and θ_a actually share a significant amount of weights, since they only differ for the weights in the last layer, leading up to the two embedding components $f_{\theta_{id}}$ and f_{θ_a} . They are indicated in that way to limit notation clutter.

3.2 Metric Learning Loss

By learning a metric embedding we directly train a function f_θ that maps images of the same identity as close as possible, effectively minimizing the intra-class variability of the embeddings, while images of different identities are mapped far away, creating a large inter-class discrepancy. In [39] they developed a margin

based approach for k -nearest neighbor classification, which has then inspired the *triplet loss* formulation of FaceNet [28] as follows

$$\mathcal{L}_T(\theta) = \sum_{(a,p,n) \text{ s.t. } y_a=y_p \neq y_n} [m + D(f_\theta(I_a), f_\theta(I_p)) - D(f_\theta(I_a), f_\theta(I_n))]_+ \quad (4)$$

where $D(\cdot, \cdot)$ denotes a suitable distance, and $[\cdot]_+$ is the hinge function, but other surrogates could be used, such as the softplus function $\ln(1 + \exp(\cdot))$. The triplet loss (4) operates by ensuring that the distance between a positive sample I_p and an anchor I_a , which have same identities, is smaller than the distance between the anchor and a negative sample I_n , which has a different identity, at least by a margin m . When the loss is optimized over a large combination of triplets (I_a, I_p, I_n) , it pulls embeddings of the same identity, while pushing apart those with different identities.

The challenges in using the triplet loss are related to the cubic growth in the number of triplets as the dataset size grows, and in forming meaningful triplets. It turns out that the embedding can quickly learn how to correctly map easy triplets. Conversely, focusing on selecting very hard triplets may not be very useful too, because we would teach the embedding how to map outlier cases, while overlooking how to handle well “average” cases. This is why it is important to efficiently mine moderate positives and negatives [28,30].

As described in [13], it turns out that there is an effective way to address both of the issues above. Triplets can be formed out of selecting P identities, and K samples per identity, with a total of PK samples in a batch. Since we are operating only within a batch, these hard selections will not be outliers, but mostly non-trivial moderate cases. In addition, this approach avoids the overhead induced by mining moderate cases from the full dataset processed by the latest update of f_θ . This procedure, named *batch hard* [13], changes (4) into

$$\mathcal{L}_{BH}(\theta) = \sum_{i=1}^P \sum_{a=1}^K [m + \max_p D(f_\theta(I_a^i), f_\theta(I_p^i)) - \min_{j,n,j \neq i} D(f_\theta(I_a^i), f_\theta(I_n^j))]_+ \cdot \quad (5)$$

3.3 Joint Classification and Metric Loss

Besides evaluating the additive angular margin softmax loss (2) and the batch hard triplet loss (5) on the most recent re-identification datasets, we plan to study their contribution into a joint loss

$$\mathcal{L}_{AMBH}(\theta, W) = \mathcal{L}_{AM}(\theta, W) + \gamma \mathcal{L}_{BH}(\theta) \cdot \quad (6)$$

where γ is a hyperparameter balancing the relative strengths of the losses.

There are a couple reasons that motivate the exploration of the joint loss (6). The first one comes from observing that a major drawback of the loss (2) is that the gradient $\nabla_\theta \mathcal{L}_A$ is proportional to the gradient $\nabla_\theta \tilde{f}_\theta$, where $\tilde{f}_\theta \doteq f_\theta / \|f_\theta\|$ because of the ℓ_2 normalization of the softmax inputs. Since \tilde{f}_θ lives on the unit hypersphere, the gradient $\nabla_\theta \tilde{f}_\theta$ will always be tangent to it. Therefore, no

gradients perpendicular to the hypersphere will be back-propagated to supervise f_θ for reducing the intra-class variability of the embedding, while maximizing the inter-class discrepancy. This issue suggests that adding a regularizing term to the loss (2), which allows orthogonal gradients to flow back could increase the hypothesis space exploration of the embedding f_θ , and better become invariant to nuisance factors of variation.

By adding (5) to (2) as in (6), we are addressing the issue highlighted above. Indeed, the intent of (5) and (2) is the same, but in (5) we do not have the requirement for f_θ to be ℓ_2 normalized. Hence, by picking a distance $D(\cdot, \cdot)$ that does not normalize the embedding, (6) enables the gradient to flow in all directions. In [6] they attempted merging (2) with (4) without success, but there they used a distance with a normalized embedding, which we advocate not to use in this case. In our experiments, we picked $D(\cdot, \cdot)$ to be the Euclidean distance.

The second reason for using (6) comes from the composition of a batch, which has certain requirements because of the batch hard mining. We note that the same batch made of $N = PK$ samples can actually be used for the loss (2). More importantly, this approach may prevent issues related to imbalanced data. Since datasets for person re-identification may have identities with a lot more samples than others, sampling a constant number of identities, from which we sample a constant number of images, imposes the embedding to be trained uniformly across the identities, rather than being under/over trained on some of them. In all of our experiments we sample the batches as it is done for the batch hard mining, regardless of the loss that we use.

Moreover, we note that since the loss (5) exercises a set of push-pull forces, it might be that when used as in (6), the effect of the additive angular margin in (2) could become less relevant. Indeed, this is one of our conclusions.

In addition to (6), we also replace the classification loss with the multi-task loss including the identity component $\mathcal{L}_{AM_{id}}(\theta_{id}, W_{id})$, and the attribute component (3). This leads to the full re-identification training model, given by

$$\mathcal{L}_{AMBH_{Attr}}(\theta, W) = \mathcal{L}_{AM}(\theta_{id}, W_{id}) + \lambda \mathcal{L}_{AM_{Attr}}(\theta_a, W_a) + \gamma \mathcal{L}_{BH}(\theta_{id}), \quad (7)$$

where λ and γ are hyperparameters that stryke a balance between the identity, the attributes, and the metric learning terms.

3.4 Network Architecture

As in most of the recent literature on person re-identification, we use a pretrained ResNet-50 [11] as backbone network. We simply discard the fully connected layer, and we change the stride of the last convolutional stage from 2 to 1. We then add a global average pooling (GAP) layer and a batch normalization layer (BN). The dimensionality of the embedding features is 2048. At this point weight normalization and ℓ_2 feature normalization is applied before entering the additive angular margin softmax loss (3), while no normalization is needed for the batch hard triplet loss component (5). Figure 1 is a simple exemplification of the architecture. During testing, unless otherwise specified, we perform all the

Table 1. Comparison with the state-of-the-art methods on Market-1501 and DukeMTMC-reID. The best and second best are shown in red and blue respectively.

Method	Backbone	Market-1501		DukeMTMC-reID	
		Rank-1	mAP	Rank-1	mAP
FD-GAN [9]	ResNet	90.5	77.7	80.0	65.4
Part-aligned [31]	GoogleNet	91.7	79.6	84.4	69.3
SGGNN [29]	ResNet	92.3	82.8	81.1	68.2
PCN+PCP [4]	ResNet	92.8	78.8	85.7	71.2
Mancs [34]	ResNet	93.1	82.3	84.9	71.8
APDR [17]	ResNet	93.1	80.1	84.3	69.7
DeepCRF [3]	ResNet	93.5	81.6	84.9	69.5
PCB [32]	ResNet	93.8	81.6	83.3	69.2
AA-Net [33]	ResNet	93.9	83.4	87.7	74.3
IA-Net [15]	ResNet	94.4	83.1	87.1	73.4
SphereReID [7]	ResNet	94.4	83.6	83.9	68.5
CAMA [41]	ResNet	94.7	84.5	85.8	72.9
DG-Net [47]	ResNet	94.8	86.0	86.6	74.8
AM0BH (Ours)	ResNet	94.6 \pm 0.21	85.9 \pm 0.28	89.2 \pm 0.40	76.7 \pm 0.26
AM0BH_{Attr} (Ours)	ResNet	94.9 \pm 0.13	86.3 \pm 0.10	89.3 \pm 0.19	77.4 \pm 0.14

experiments with the ℓ_2 normalized embedding $f_\theta/\|f_\theta\|$, and re-identification is done via cosine similarity. Specifically θ is actually θ_{id} , when the network has been trained with the full model (7).

4 Experiments

We evaluate our model on three person re-identification datasets. Every evaluation was repeated 10 times. We report the average performance metrics with their standard deviations for the following datasets.

Market-1501: contains 32668 images of 1501 identities captured by six cameras [46].

DukeMTMC-reID: contains 36441 images of 1812 identities captured by eight high resolution cameras [49].

MSMT17: is the most recent and challenging person re-identification dataset. It contains 126441 images of 4101 identities captured by 15 cameras [37].

We use the data provided in [20] as attribute labels for Market-1501 and DukeMTMC. These attributes are manually annotated at the identity level. There are 27 attributes for Market-1501 and 23 attributes for DukeMTMC. Some examples of attributes include: gender, hair length, carrying backpack, carrying handbag, wearing hat, different upper body and lower body clothing colors, length and type of lower body clothing, shoe type and shoes color.

4.1 Implementation Details

We implemented our approach with PyTorch [25], and for the backbone network we use ResNet-50 [12] pre-trained on ImageNet [5]. The batch size is 32 where

Table 2. Comparison with the-state-of-the-art methods on the MSMT17 dataset. The best and second best are shown in red and blue respectively.

Method	Backbone	Rank-1	Rank-5	Rank-10	mAP
PCB [32]	ResNet	68.2	81.2	85.5	40.4
IA-Net [15]	ResNet	75.5	85.5	88.7	46.8
DG-Net [47]	ResNet	77.2	87.4	90.5	52.3
AM0BH (Ours)	ResNet	78.1 \pm 0.40	88.3 \pm 0.13	91.2 \pm 0.19	53.4 \pm 0.32

$P = 4$ and $K = 8$. For the Market-1501 dataset the input image size is 256×128 while it is 288×144 for DukeMTMC-reID and the MSMT17 datasets.

For data augmentation, training images are randomly flipped and erased. The model is trained using the Adam optimizer with default hyper parameters for 150 epochs. The learning rate is linearly increased from 10^{-5} to 10^{-3} for the first 20 epochs to help the network bootstrap. Then the learning rate is set to 10^{-3} after the first 20 epochs, and it decreases to 10^{-4} and 10^{-5} after epochs 90, and 130, respectively.

There are two settings for our proposed approach. The first one, AM0BH, uses the joint loss (6) with no attributes. γ is 0.43, 0.5 and 0.4 for Market-1501, DukeMTMC-reID and MSMT17, respectively. The second setting, AM0BH_{Attr}, leverages the full model (7). The 2048 embedding features are divided into f_{θ_a} and $f_{\theta_{id}}$. f_{θ_a} has size $M \times Q$, where M is the total number of attributes and Q is the size of the input features to each of the attribute classifiers. Q is set to 16. The rest of the embedding features, $f_{\theta_{id}}$, is the input to the identity classifier and triplet loss. In (7) we set λ to 0.25 and γ to 0.54 for Market-1501. While for DukeMTMC-reID, we set λ to 0.2 and γ to 0.33.

4.2 Comparison with State-of-the-Art Methods

The performance is evaluated by CMC (Cumulative Matching Characteristic) and mAP (Mean Average Precision) after computing the matching score between the probe image and gallery images. We discard the score if the probe image and gallery image are from the same view.

To show the performance of our proposed approach, we compare it with the state-of-the-art methods on three person re-identification dataset. However, we are not able to implement AM0BH_{Attr} on MSMT17 since there is no attributes annotation available. Table 1 shows the results on Market-1501 and DukeMTMC-reID. For Market-1501, AM0BH outperforms most of the state-of-the-art and the performance is close to the best, i.e., DG-Net [47] in terms of rank-1 and mAP. While for DukeMTMC-reID, AM0BH outperforms the state-of-the-art by achieving 89.2% on rank-1 and 76.7% mAP. Table 2 shows results on MSMT17. The proposed AM0BH outperforms DG-Net [47] by a gap of 0.9%, 0.9%, 0.7% and 1.1% for rank-1, rank-5, rank-10 and mAP respectively. By using attributes, AM0BH_{Attr} further improves the performance over AM0BH by 0.3% and 0.4% for rank-1 and mAP for Market-1501, and by 0.1% and 0.7% for rank-1 and mAP for DukeMTMC-reID.

Table 3. Ablation study. Shows the effect of different loss combinations. Losses: a) AM0 - softmax loss (2) when margin is set to 0; b) AM - softmax loss (2) when margin is set to 0.5; c) BH - batch hard triplet loss (5); d) AM0BH1 - softmax loss when margin is set to 0 combined with batch hard triplet loss with feature normalization; e) AMBH - softmax loss when margin is set to 0.5 combined with batch hard triplet loss; f) AM0BH - softmax loss when margin is set to 0.0 combined with batch hard triplet loss; g) AM0BHsp - softmax loss when margin is set to 0.0 combined with batch hard triplet loss with softplus function instead of hinge loss.

Loss	Market-1501				DukeMTMC-reID				MSMT17			
	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP
AM0	92.86 \pm 0.34	97.57 \pm 0.12	98.47 \pm 0.04	83.67 \pm 0.17	87.74 \pm 0.29	94.06 \pm 0.17	95.62 \pm 0.35	74.60 \pm 0.30	77.50 \pm 0.25	87.80 \pm 0.30	90.78 \pm 0.18	51.82 \pm 0.19
AM	94.16 \pm 0.23	98.04 \pm 0.21	98.92 \pm 0.08	84.54 \pm 0.22	88.31 \pm 0.18	94.34 \pm 0.48	95.78 \pm 0.24	75.51 \pm 0.16	78.20 \pm 0.28	88.15 \pm 0.42	91.15 \pm 0.35	53.00 \pm 0.49
BH	84.74 \pm 0.23	94.64 \pm 0.36	96.74 \pm 0.13	67.40 \pm 0.46	81.60 \pm 0.49	91.02 \pm 0.38	93.46 \pm 0.18	65.16 \pm 0.31	56.34 \pm 0.92	73.26 \pm 0.99	79.30 \pm 0.70	30.86 \pm 0.67
AM0BH1	94.28 \pm 0.13	97.90 \pm 0.20	98.80 \pm 0.07	84.52 \pm 0.16	88.02 \pm 0.18	93.96 \pm 0.27	95.48 \pm 0.26	74.56 \pm 0.32	77.46 \pm 0.13	87.64 \pm 0.15	90.60 \pm 0.19	51.86 \pm 0.22
AMBH	93.29 \pm 0.40	97.80 \pm 0.11	98.70 \pm 0.14	84.00 \pm 0.09	88.18 \pm 0.46	94.62 \pm 0.17	96.19 \pm 0.18	76.71 \pm 0.21	77.93 \pm 0.47	88.07 \pm 0.47	91.07 \pm 0.38	52.70 \pm 0.53
AM0BH(Ours)	94.64 \pm 0.21	98.22 \pm 0.16	99.02 \pm 0.11	85.90 \pm 0.28	89.20 \pm 0.40	94.72 \pm 0.33	96.26 \pm 0.17	76.68 \pm 0.26	78.14 \pm 0.40	88.34 \pm 0.13	91.24 \pm 0.19	53.44 \pm 0.32
AM0BHsp(Ours)	94.42 \pm 0.15	98.22 \pm 0.19	99.02 \pm 0.13	85.76 \pm 0.19	88.79 \pm 0.31	94.85 \pm 0.22	96.32 \pm 0.19	77.42 \pm 0.3	78.26 \pm 0.27	88.38 \pm 0.11	91.20 \pm 0.14	53.36 \pm 0.40

4.3 Ablation Study

In the ablation study we compare different loss combinations on all three datasets used in Section 4.2. First, we examine how the identity classification loss (2), and the metric loss (5) behave independently. The summary results of this experiment are included in Table 3. Then we examine different combinations.

Identity classification loss. We start by examining the additive angular loss applied for identity classification (2). When the margin is set to 0 (row AM0 in Table 3), this case is equivalent to the loss described in [7]. Then, we use the loss (2) when the margin is set to 0.5, as in [6] (row AM in the Table 3). We observe a significant improvement of all metrics, which proves that the additive angular margin has a positive effect when the softmax loss is used alone.

Metric learning loss. We continue by examining batch hard triplet loss (5). This is the row BH in Table 3. It can be seen that this loss alone underperforms the identity classification losses in rows AM, and AM0.

Identity classification and metric learning losses. We analyze four cases: the combination of AM0 and BH, the combination of AM and BH, the combination of AM0 and BH with feature normalization for BH, and the combination of AM0 and BH with softplus instead of hinge loss.

Combination of AM0 and BH. The combination of the softmax loss (2) with margin set to 0 and batch hard triplet loss (5) is presented in the row AM0BH of Table 3. We observe improvement on almost all metrics compared to AM0 or BH individually, which signifies that they complement each other.

Combination of AM and BH. The combination of the additive angular margin softmax loss (2) with margin set to 0.5 and the batch hard triplet loss (5) is presented in the row AMBH of Table 3. We observe that it does not provide improvement compared to AM0BH. This means that the angular margin becomes less relevant, since the batch hard triplet loss exercises a set of push-pull forces that is likely comparable to the effect of the margin, and might even generate conflicts when this is present, leading in this case, to results closer to AM alone.

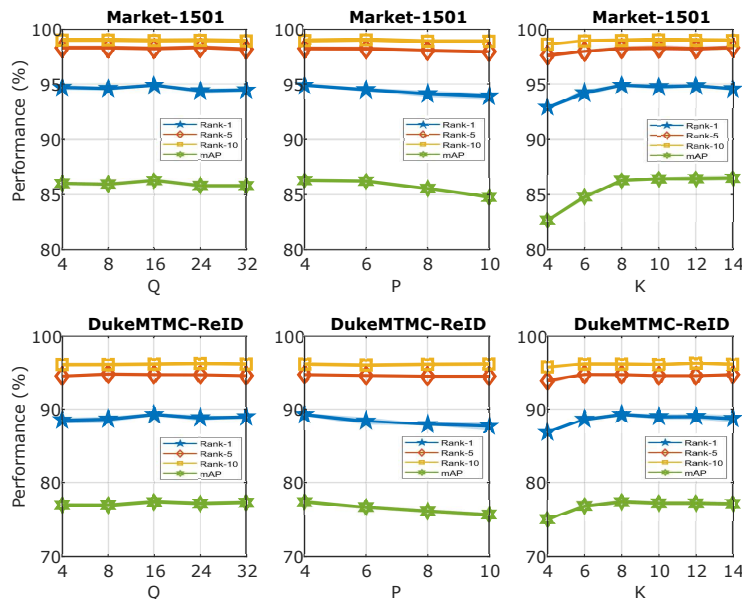


Fig. 2. Ablation study. Shows the effect of Q , P and K on the performance of $AM0BH_{Attr}$. Q is the size of embedding features fed to each attribute classifier, P is the number of identities in each mini-batch (P is fixed to 4) and K is the number of samples per identity in each mini-batch (K is fixed to 8).

Combination of AM0 and BH with normalized features. The combination of the softmax loss (2) with margin set to 0, and the batch hard triplet loss (5) with normalized features is presented in row AM0BH1 of Table 3. We note a decreased performance, when compared with row AM0BH. As described in Section 3.3, this might be due to the feature normalization prior to the triplet loss, which forces no gradients perpendicular to the hypersphere to be back-propagated to supervise f_θ . Removing that constraint would allow the orthogonal gradients flow that could increase the hypothesis space exploration of the embedding f_θ .

Combination of AM0 and BH with softplus. The combination of the softmax loss (2) with margin set to 0 and batch hard triplet loss (5) with softplus function instead of the hinge loss is presented in row AM0BHsp of Table 3. It shows overall similar performance to AM0BH, but slightly higher mAP, while slightly lower rank1-10 metrics. We speculate that hinge loss concentrates only on the triplets within the margin, ignoring the tail of the distribution, which is beneficial for rank1 - rank10 metrics, while with softplus the whole distribution of triples is accounted in the loss, which is beneficial for the mAP metric.

Identity and attribute classification with metric learning losses. Here we present the ablation study that supports the addition of the attribute classification task to further improve the robustness against nuisance factors, as

previously suggested. We study the influence of Q and batch size on the performance of $AM0BH_{Attr}$. Figure 2 shows the effect of the size, Q , of the embedding features fed to each attribute classifier. The best performance is achieved when Q is 16. Figure 2 also shows the effect of the batch size by examining different values of P and K respectively on the performance.

5 Conclusions

We have further studied the learning of a feature embedding for person re-identification via a joint optimization of a discriminative and a metric learning loss to minimize the intra-class variation and maximize the inter-class separation. Our approach was motivated by observing untapped limitations imposed by a margin based softmax loss onto the gradient flow that supervises the training of the embedding. We have verified that adding a triplet loss as regularizer serves as proxy for the missing gradient directions, and enables learning a better embedding. Moreover, we have shown that adding a discriminative semantic task like predicting attributes, further strengthens the robustness of the representation. We have verified that on the three most challenging datasets by setting new state-of-the-art performance for the case of holistic representations for person re-identification that do not leverage explicit modeling of nuisance factors (e.g., pose). Moreover, we found that the joint loss achieves its best performance when we do not require a margin in the softmax portion, showing the importance of the contribution added by the triplet component, when it is used to expand the directions of the gradient flow.

Acknowledgements This material is based upon work supported in part by the National Science Foundation under Grant No. 1920920.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: IEEE CVPR. pp. 3908–3916 (2015)
2. Bak, S., Carr, P.: One-shot metric learning for person re-identification. In: IEEE CVPR. pp. 2990–2999 (2017)
3. Chen, D., Xu, D., Li, H., Sebe, N., Wang, X.: Group consistent similarity learning via deep crf for person re-identification. In: IEEE CVPR. pp. 8649–8658 (2018)
4. Chikontwe, P., Lee, H.J.: Deep multi-task network for learning person identity and attributes. IEEE Access **6**, 60801–60811 (2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE CVPR. pp. 248–255. Ieee (2009)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. arXiv:1801.07698 (Jan 2018)
7. Fan, X., Jiang, W., Luo, H., Fei, M.: Spheredid: Deep hypersphere manifold embedding for person re-identification. J. of Vis. Comm. and Image Rep. **60**, 51–58 (2019)

8. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE CVPR. pp. 2360–2367. IEEE (2010)
9. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: NeurIPS. pp. 1222–1233 (2018)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778. IEEE (Jun 2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778 (2016)
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person Re-Identification. arXiv:1703.07737 (Mar 2017)
14. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: ECCV. pp. 780–793. Springer (2012)
15. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: IEEE CVPR. pp. 9317–9326 (2019)
16. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: IEEE CVPR. pp. 384–393 (2017)
17. Li, S., Yu, H., Hu, R.: Attributes-aided part detection and refinement for person re-identification. *Pattern Rec.* **97**, 107016 (2020)
18. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In: IJCAI. pp. 2194–2200. IJCAI’17, AAAI Press (Aug 2017)
19. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE CVPR. pp. 2197–2206 (2015)
20. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., Yang, Y.: Improving person re-identification by attribute and identity learning. *Pattern Rec.* **95**, 151–161 (2019)
21. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: What features are important? In: ECCV. pp. 391–401. Springer (2012)
22. Liu, J., Zha, Z.J., Xie, H., Xiong, Z., Zhang, Y.: Ca3net: Contextual-attentional attribute-appearance network for person re-identification. In: ACM Multimedia. pp. 737–745 (2018)
23. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SphereFace: Deep hypersphere embedding for face recognition. arXiv (Apr 2017)
24. Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification (2012)
25. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
26. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: IEEE CVPR. pp. 3318–3325 (2013)
27. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. In: ECCV. pp. 650–667 (2018)
28. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE CVPR. pp. 815–823 (2015)

29. Shen, Y., Li, H., Yi, S., Chen, D., Wang, X.: Person re-identification with deep similarity-guided graph neural network. In: ECCV. pp. 486–504 (2018)
30. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: IEEE CVPR. pp. 4004–4012 (2016)
31. Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K.: Part-aligned bilinear representations for person re-identification. In: ECCV. pp. 402–419 (2018)
32. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV. pp. 480–496 (2018)
33. Tay, C.P., Roy, S., Yap, K.H.: Aanet: Attribute attention network for person re-identifications. In: IEEE CVPR. pp. 7134–7143 (2019)
34. Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: ECCV. pp. 365–381 (2018)
35. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: NormFace: L2 hypersphere embedding for face verification. arXiv:1704.06369 (Apr 2017)
36. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: CosFace: Large margin cosine loss for deep face recognition. arXiv (Jan 2018)
37. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: IEEE CVPR. pp. 79–88 (2018)
38. Wei-Shi Zheng, Shaogang Gong, Tao Xiang: Person re-identification by probabilistic relative distance comparison. In: IEEE CVPR. vol. 0, pp. 649–656 (Jun 2011)
39. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. *JMLR* **10**(2) (2009)
40. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: IEEE CVPR. pp. 1249–1258 (2016)
41. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: IEEE CVPR. pp. 1389–1398 (2019)
42. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person Re-Identification: A survey and outlook. *IEEE TPAMI* **44**(6), 2872–2893 (Jun 2022)
43. Zhang, G., Xu, J.: Person re-identification by mid-level attribute and part-based identity learning. In: ACCV. pp. 220–231. PMLR (2018)
44. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: IEEE CVPR. pp. 1239–1248 (2016)
45. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: IEEE ICCV. pp. 3219–3228 (2017)
46. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE ICCV. pp. 1116–1124 (2015)
47. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: IEEE CVPR. pp. 2138–2147 (2019)
48. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person re-identification. arXiv:1611.05666 (Nov 2016)
49. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: IEEE ICCV. pp. 3754–3762 (2017)