Deep Domain Knowledge Distillation for Person Re-identification

Junjie Yan(✉)

Fudan University, Shanghai, China
jjyan17@fudan.edu.cn

Abstract. Learning generic and robust representations with data from multiple domains is a big challenge in Person ReID. In this paper, we propose an end-to-end framework called Deep Domain Knowledge Distillation (D^2KD) for leaning more generic and robust features with Convolutional Neural Networks (CNNs). Domain-specific knowledge learned by the auxiliary network is transferred to the domain-free subnetwork and guides the optimization of the feature extractor. While person identity information is transferred to the auxiliary network to further accurately identify domain classes. In the test period, just with a single base model as the feature extractor, we improve the Rank-1 and mAP by a clear margin. Experiments on Market-1501, CUHK03 and DukeMTMC-reID demonstrate the effectiveness of our method.

Keywords: Person re-identification · Domain · Knowledge distillation

1 Introduction

Person Re-identification is a cross-camera retrieval task, which aims at retrieving images of a specific pedestrian in a large dataset when given a specific query. The key challenge in this task is the large appearance and background variations, caused by changes in human body poses and camera views as shown in Fig. 1.

Recent years, deep convolutional neural networks have led to a series of breakthroughs for image classification [5,9] and these architectures can be easily transferred into other computer vision tasks. CNNs are also successfully employed in Person ReID with significant performance. For example, several works [3,17,25] employ deep classification model to learn feature representations of images.

In addressing the challenge of camera variations, a previous body of literature chooses to learn stable feature representations that have invariant properties under different cameras. Some networks [21,25] are trained with a pairwise verification loss, which measures the similarity between two images. However, these methods have to be used in a cross-image representation mode. During test time, the query image has to pair with each image in the gallery dataset and passes through the forward network, which is time inefficient and intolerant for large-scale real-world applications. Methods such as [6] directly optimize the distance

I. V. Tetko et al. (Eds.): ICANN 2019, LNCS 11728, pp. 700–713, 2019.
https://doi.org/10.1007/978-3-030-30484-3_55

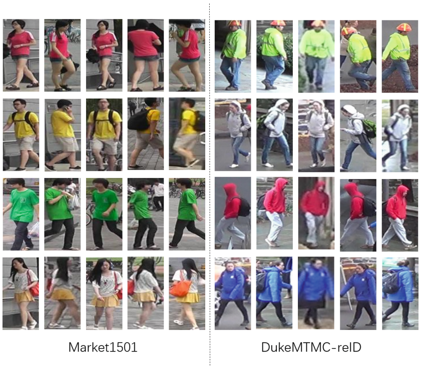

Fig. 1. Example of images from two person re-identification datasets [23,27]. In both datasets, samples in each line have the same identity but are observed from different cameras. Our goal is to extract generic feature representations for each specific identity.

of the images in embedding space and easy to employ in real task, but it does not take advantage of person identity label and waste abundant domain information.

Upon above discussions, this paper focuses on finding the robust feature representations for each person identity among different domains. Based on TriNet [6], we propose a Deep Domain Knowledge Distillation (D^2KD) method for further obtaining more robust features, by taking full advantage of pedestrian labels and wasted abundant domain information.

Inspired by [7], we are trying to find a higher soft bound for the feature extractor to optimize. Based on a traditional end-to-end classification model, we further apply dynamic label smoothing regularization on the training data via knowledge distillation [7], whose typical application is to transfer knowledge from a teacher network to a student network.

We propose an end-to-end framework named Deep Domain Knowledge Distillation which contains two subnetworks, one for extracting domain-free information and the other for domain-specific information. In the training period, we exchange the information between the two networks to generate more discriminative features. While at test time, only the domain-free part is used for further evaluation.

Our experiments improve the Rank-1 and mAP on image based dataset Market-1501, CUHK03 and DukeMTMC-reID by a clear margin.

The main contributions of this paper are summarized as follows:

- We propose an end-to-end framework that helps to learn higher quality camera-invariant property.
- We make full use of domain-specific information and obtain a theoretically higher bound for the classifier to optimize and demonstrate the effectiveness through the experiments.
- In the test period, domain-specific network only plays an auxiliary role. That is to say, there is no additional parameter but a pure base model like ResNet-50.

2 Related Works

2.1 Deep Learning for Person Re-identification

Recent years, feature representations learned by Convolutional Neural Networks have shown great effectiveness in a wide range of computer vision tasks including Person Re-identification. And these methods mainly focus on two categories: learning robust metrics [1,19,21,25] and extracting discriminative features [3,6,16,17,24].

Some methods focus on the similarity between instances. In [1], a pair of cropped pedestrian images passed through a specifically designed CNN with a binary verification loss function for person re-identification. In [21], to formulate the similarity between pairs, images were partitioned into three horizontal parts respectively and calculated the cosine similarity through a siamese CNN [2] model.

Another strategy is directly learning discriminative embeddings which makes full use of the ReID labels. [24] proposed the ID-discriminative embedding (IDE) to train the ReID model in an image classification manner based on imagenet pretrained model. [17] simply partitioned image into several horizontal parts and gave each part a pedestrian identity label supervision individually which helps learn stable part features. And [3] extracted features from multiple sizes to enhance the stability of features.

2.2 Knowledge Distillation

Knowledge distillation [7] is an effective and widely used technique to transfer knowledge from a teacher to a student network. The typical application is to transfer from a cumbersome network to a small network, making the model memory-efficient and fast execution.

3 The Proposed D^2KD Method

In this section, we firstly provide necessary background and notion for person Re-identification. Then we introduce our network architecture in detail and give an insight into our Deep Domain Knowledge Distillation (D^2KD) method.

The overall framework is illustrated in Fig. 2. Our pipeline consists of two subnetworks, one focus on extracting domain-free features and another mainly learns domain-specific features and plays an auxiliary role during the training period. Besides, a Deep Domain Knowledge Distillation (D^2KD) module is implemented to exchange information between two subnetworks.

3.1 Problem Formulation

Suppose the training set contains n labeled images from C persons, we denote the training set as $T = \{x_i, y_i\}_{i=1}^{n}$, where x_i is the i-th image and y_i is a C-dimentional one hot vector which indicates the label of x_i.

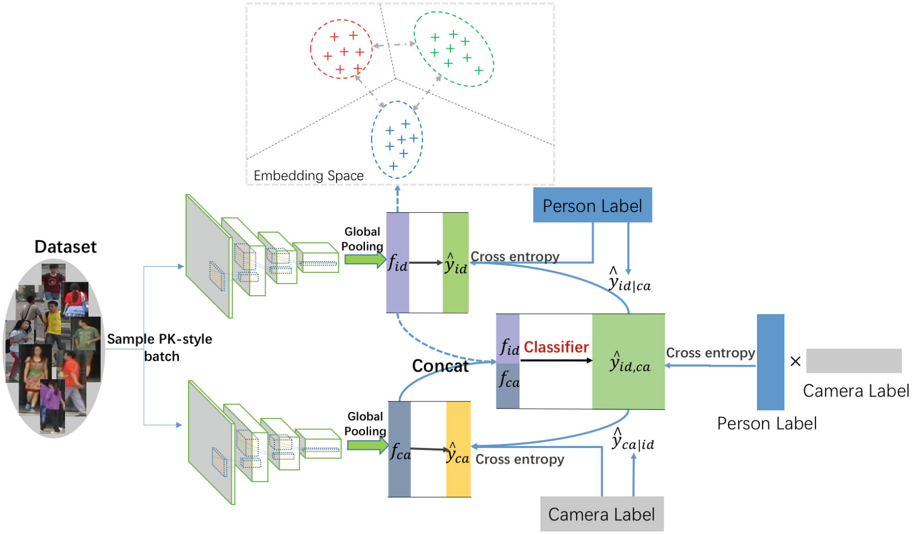

Fig. 2. Structure of our D^2KD framework. The input image passes through the two distinct backbone networks and obtains two 2048-dimentional feature vectors f_{id} and f_{ca} respectively. A following fully-connected layer is added to both feature vectors for the classification task. After that we concatenate f_{id} and f_{ca} together and force the classifier to learn a joint distribution $p(id, ca|x)$. Conditional probabilities $p(id|ca, x)$ and $p(ca|id, x)$ are then used as the supervision of two subnetworks. By the way, we take full advantage of triplet loss and apply it in the feature space to learn stable domain-free features. The dash line means we ignore the gradient from the output in this way at training time.

Based on the training set T, Person Re-identification is to learn a function $f_\theta(x) : \mathbb{R}^F \rightarrow \mathbb{R}^D$ which maps semantically similar points $\mathbb{R}^F$ from data manifold to be closer in the embedding space $\mathbb{R}^D$, where $D \ll F$.

When given a query person image x_q, ReID targets to return images which contain identical person in x_q from a gallery set G. Formally, for a specific query instance x_q, we rank the images in G by the distance $\mathcal{D}(f_\theta(x_q), f_\theta(x_g))$ in an ascending order, where x_g means image sampled from G, $\mathcal{D}(.,.)$ is some kind of distance function like Euclidean distance. Larger distance $\mathcal{D}(.,.)$ is equivalent to lower similarity.

3.2 Domain-Free Network

We formulate the domain-free subnetwork as $f_{id}(x) : \mathbb{R}^F \rightarrow \mathbb{R}^D$ and use the ResNet-50 [5] architecture with parameters pretrained on imagenet. Given an image x with identical person label, we can obtain a D-dimentional features $f_{id}(x)$ by feeding the image to the network.

Suppose there are C pedestrian labels in the dataset, we obtain pedestrian label probability by passing $f_{id}(x)$ through a single fully connected layer. That is to say:

$$\hat{y}_{id}(x) = softmax(W_{id}^T f_{id}(x) + b_{id}) \tag{1}$$

where W_{id} is a $D \times C$ parameter matrix and b_{id} is a C-dimentional bias. In order to learn discriminative and domain-free features, we apply cross entropy on $\hat{y}_{id}$. The cross entropy loss can be formulated as:

$$L_{id} = -\frac{1}{n}\sum_{i=1}^{n} y_{id}^T(x_i) log(\hat{y}_{id}(x_i)) \tag{2}$$

$y_{id}(x_i)$ is the ground truth identity label of the specific instance x_i. Meanwhile, we apply triplet loss in the feature space. For a minibatch B, there are P distinct person identities, and each identity has K specific instances, thus resulting in a batch of $B = PK$ images, we formulate the function as:

$$\begin{aligned} L_{tri\_batch}(\theta; B) = \sum_{a=1}^{B=PK} [m + & \max_{\substack{p=1\ldots B \\ y_p = y_a}} \mathcal{D}(f_a, f_p) \\ & - \min_{\substack{n=1\ldots B \\ y_n \neq y_a}} \mathcal{D}(f_a, f_n)]_+ \end{aligned} \tag{3}$$

That is to say in a minibatch, for an anchor point x_a, we optimize its distance from a positive data point x_p to be lower than a negative data point x_n by at least a margin m in the embedding space.

The triplet loss directly optimizes the distance among instances in feature space. However, only applying the triplet loss does not make use of domain-specific information, resulting in the decrease of the classification ability.

3.3 Domain-Specific Network

Similar to domain-free network, we use another auxiliary ResNet-50 network to learn domain-specific features. We formulate the network as $f_{ca}(x) : \mathbb{R}^F \to \mathbb{R}^D$, and the domain(camera) classification probability can be formulated as:

$$\hat{y}_{ca}(x) = softmax(W_{ca}^T f_{ca}(x) + b_{ca}) \tag{4}$$

where W_{ca}^T is a $D \times M$ matrix and M is the number of domains(cameras), b_{ca} is a M-dimentional bias. And the cross entropy loss can also be applied as:

$$L_{ca} = -\frac{1}{n}\sum_{i=1}^{n} y_{ca}^T(x_i) log(\hat{y}_{ca}(x_i)) \tag{5}$$

where $y_{ca}(x_i)$ is ground truth camera label for x_i in the dataset.

3.4 Distillation Module

In order to make the embeddings more discriminative while applying triplet loss at the same time, we propose Deep Domain Knowledge Distillation (D^2KD) method which aims at finding a higher soft bound for the identity classifier to optimize. By smoothing label in a dynamic manner, our method improves the classification accuracy bound while taking advantage of triplet loss at the same time.

Given an input image x, We concatenate the features $f_{id}(x)$ and $f_{ca}(x)$ together and denote $\phi(x) = [f_{id}(x), f_{ca}(x)]$ which contains person identity and domain information at the same time. We detach the gradient from $\phi(x)$ to $f_{id}(x)$.

Since $\phi(x)$ contains both person identity information and domain specific information, we choose to pass it through a simple classifier (here we use fc-relu-fc-softmax layers), and resize the outputs to a $C \times M$ matrix which represents the joint distribution $P(id, ca|x)$. And the cross entropy loss function for the joint distribution is given by:

$$\begin{aligned} L_{joint}(x) &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C}\sum_{k=1}^{M} y_{j,k}(x_i) log(p(j,k|x_i)) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C}\sum_{k=1}^{M} y_{j,k}(x_i) log(\hat{y}_{j,k}(x_i)) \end{aligned} \tag{6}$$

where $y_{j,k}(x)$ is the ground truth one-hot label for the joint distribution $p(id, ca|x)$, and $y_{j,k}(x) = 1$ when the image instance x belongs to the i-th person identity and j-th domain(camera), otherwise $y_{j,k}(x) = 0$. $\hat{y}_{j,k}(x)$ is a element of the output probability matrix in the j-th row and k-th column $(0 \leq j < C, 0 \leq k < M)$.

Apparently, from the bayes perspective, the posteriori $P(id|ca, x)$ should be a soft target to the prior $\hat{y}_{id}(x) = P(id|x)$, and similarly $P(ca|id, x)$ is a soft higher bound to $\hat{y}_{ca}(x) = P(ca|x)$, this greatly correspond to the knowledge distillation condition. It seems quite simple that the conditional probability can be computed like:

$$\begin{aligned} P(id|ca = k, x) &= \frac{[\hat{y}_{1,k}(x), \hat{y}_{2,k}(x), \ldots, \hat{y}_{C,k}(x)]}{\sum_{j=1}^{C} \hat{y}_{j,k}(x)} \\ P(ca|id = j, x) &= \frac{[\hat{y}_{j,1}(x), \hat{y}_{j,2}(x), \ldots, \hat{y}_{j,M}(x)]}{\sum_{k=1}^{C} \hat{y}_{j,k}(x)} \end{aligned} \tag{7}$$

Upon the conditional probability $P(id|ca, x)$ and $P(ca|id, x)$, we apply the pedestrian identity and the camera label as the supervision for the conditional probability, the loss function can be written as:

$$\begin{aligned} L_{cond} = & -\frac{1}{n}\sum_{i=1}^{n} y_{id}^{T}(x_i)log(P(id|ca = y_{ca}(x_i), x_i)) \\ & -\frac{1}{n}\sum_{i=1}^{n} y_{ca}^{T}(x_i)log(P(ca|id = y_{id}(x_i), x_i)) \end{aligned} \tag{8}$$

After exchanging information between two subnetworks, we hope to distillate useful knowledge from the joint distribution. The distillation loss can be formulated as:

$$\begin{aligned} L_{KD} = \frac{1}{n}\sum_{i=1}^{n} & [H(\hat{y}_{id}(x_i), P(id|ca = y_{ca}(x_i), x_i)) \\ & + H(\hat{y}_{ca}(x_i), P(ca|id = y_{id}(x_i), x_i))] \end{aligned} \tag{9}$$

where $\hat{y}_{id}$ and $\hat{y}_{ca}$ is the probability produced by the two subnetworks respectively. $H(.,.)$ is the binary cross entropy loss function as follows:

$$H(P, Q) = -\sum_{i=1}^{\mathcal{C}}[q_i log(p_i) + (1 - q_i)log(1 - p_i)] \tag{10}$$

Our total loss function is:

$$L_{total} = L_{xent} + \lambda_1 L_{tri} + \lambda_2 L_{KD} \tag{11}$$

where $L_{xent} = L_{id} + L_{ca} + L_{joint} + L_{cond}$ is the total cross entropy loss for classification. λ_1 and λ_2 are loss tradeoffs.

By exchanging the information between two subnetworks, our D^2KD method take the predicted probability $P(id|x, ca)$ as a soft target for the base feature extractor and $P(ca|x, id)$ a supervision for the auxiliary network to obtain higher quality domain- specific features. With the collaboration of two subnetworks, our method can learn more generic and robust features.

3.5 Test Strategy

At test time, we only apply the domain-specific network to extract feature vectors from images. That is to say, the domain specific features just play an auxiliary role in the training period and were not used in the test phase. The network architecture is shown in Fig. 3, images are fed into the domain-free network and obtain a D-dimentional feature representation.

Our D^2KD method offers a soft target for the domain-free network and erases domain information by applying the triplet loss at the same time, making the backbone network generate more discriminative features.

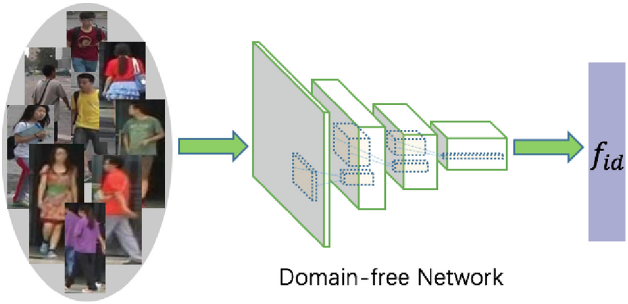

Fig. 3. Test network for our D^2KD method. Only the domain-free subnetwork is used to extract features for further evaluation.

4 Experiment

4.1 Datasets and Settings

Datasets. We conduct experiments mainly on three large scale image-based person re-identification benchmark datasets that contain multiple positive samples for each query in the gallery: including Market-1501 [23], CUHK03 [11] and DukeMTMC-reID [27]. The overview of these datasets is in Table 1.

Market-1501 [23] is the most famous large image-based ReID benchmark dataset. It contains 1,501 identities and 32,668 labeled bounding boxes captured from 6 different view points. The bounding boxes are detected using Deformable Part Model (DPM) [4]. The dataset is split into two parts: the training set contains 12,936 bounding boxes of 751 identities, the rest 19,732 images with 750 identities are included in the test set. In the test period, 3,368 images with 750 identities are used for the query to identify the correct identities on the gallery set. We use the single-query (SQ) evaluation for this dataset.

Table 1. Statistics on three person re-id datasets.

Datasets	Cams	IDs	Identity split		Person bounding box split		
			Training	Test	Training	Gallery	Query
Market-1501 [23]	6	1,501	751	750	12,936	19,732	3,368
CUHK03 [11]	2	1,467	767	700	7,368	5,328	1,400
DukeMTMC-reID [27]	8	1,404	702	702	16,522	17,661	2,228

CUHK03 [11] is constructed by both manual labeling and auto-detection (DPM) [4]. It contains 14,096 images of 1,467 identities and each identity is captured from 2 cameras in the CUHK campus. Each identity has an average of 4.8 images in each camera. The dataset is split into the training set and the test set. The training set contains 7,368 images with 767 identities and the test

set contains the rest 6,728 images with 700 identities. In the test period 1400 queries are given to identity corresponding identities in the gallery dataset.

DukeMTMC-reID [27] is a subset of the DukeMTMC website in the format of Market-1501 dataset. The DukeMTMC-reID dataset has 34,611 images belonging to 1,404 identities from 8 cameras. Similar to Market-1501, it consists of 16,522 training images from 702 identities, 2,228 query images and 17,661 gallery images from the other 702 identities.

Backbone CNN Model for ReID. To train our model, we resize all images of size $H \times W$ to $1\frac{1}{8}(H \times W)$, of which we take random cropping and random horizontal flipping and get $H \times W$ cropped images. Specifically, we keep the aspect ratio $H \times W$ of all the images to 256×128.

We performed all our experiments using the Pytorch [14] framework. We use the Imagenet pretrained ResNet-50 [5] model from Pytorch official repository for both subnetworks and replace the last 1000-dimensional classification layer with a new fully connected layer on both two networks, with output dimensions equal to the count of identities and camera classes respectively.

The network generates $D = 2048$ dimentional feature vector for each image. In the test period, the D-dimentional Pool-5 feature vector are used for further evaluation and we apply the Euclidean distance to compute the similarity between pairs.

We use the Adam optimizer [8] with the default hyper-parameter ($\epsilon = 10^{-3}, \beta_1 = 0.9, \beta_2 = 0.999$) for all experiments. The distillation temperature T is set to be a constant 1 all over the experiments. The loss tradeoffs λ_1 and λ_2 are set to a constant 1 during all the experiments.

Data Sampling and Batch Generation. In each training epoch, we shuffle all the images in the training set together and map the images with their identities. Then for each specific identity, we sample K different instances. After that, we rerank the dataset with random permutation of person identities. This preprocess is quite convenient when we apply triplet loss in the following training schedule.

Since the triplet loss in Eq. 3 requires slightly different mini-batches. At each iteration, we sequentially sample PK images from the training preprocessed dataset. Under above sampling settings, we train our model on quite a different dataset at every epoch.

Evaluation Metrics. In the test period, for each instance sequentially sampled from the query dataset, we rank the images from the gallery dataset by the distance to the specific instance for further evaluation.

We employ two evaluation metrics to evaluate the performance. The first one is the Cumulated Matching Characteristics (CMC) and the other one is the mean average precision (mAP) while considering ReID as an object retrieval problem, as describe in [23].

Table 2. Comparison of several methods on Market-1501. Rank-1, Rank-5 and mAP are shown. We use ResNet-50 as backbone. '-': No reported result available

Methods	Rank-1	Rank-5	mAP
BoW + KISSME [23]	44.4	63.9	20.8
LOMO + Null Space [22]	55.43	-	29.87
Gated siamese CNN [18]	65.88	-	39.55
CAN [13]	60.3	-	35.9
ResNet 50(I+V) [25]	79.51	90.91	59.87
Latent Parts(Fusion) [10]	80.31	-	57.53
IDE(R)(Re-ranked) [28]	74.85	-	59.87
MultiScale [3]	88.9	-	73.1
TriNet [6]	84.92	94.21	69.14
TriNet [6] (Re-ranked)	86.67	93.38	81.07
AACN [20]	85.90	-	66.87
AACN [20] (Re-ranked)	88.69	-	82.96
PSE [15]	87.7	-	69.0
PSE [15] (Re-ranked)	90.2	-	83.5
D^2KD	**91.09**	97.03	**76.76**
D^2KD(Re-ranked)	**92.73**	96.11	**88.93**

4.2 Performance Evaluation

Evaluation on Market1501. We compare the ReID performance of several existing methods against the proposed D^2KD method on the Market-1501 benchmark. Every epoch we sample $K = 8$ images from each person identity as the training dataset and set the batch size equal to 128. So each iteration contains $P = 128/K = 16$ different identities. By taking above sampling strategy, we are able to make our training dataset spread more uniformly and this can be quite helpful. Since all bounding boxes were given by auto-detection, this dataset represents a more scalable deployment scenario than other datasets with manually labelled bounding boxes.

Table 2 shows the superiority of our D^2KD model over all the competitions. The left column lists several methods these years, Rank-1, Rank-5 and mAP of these methods are given at the same time. Without any additional parameters, our model's performance is substantially better. Our method outperforms TriNet [6] by a clear margin, improving Rank-1 by 6.17% and mAP by 7.62 %. And after re-ranking [28], the difference changes to 6.06% and 7.86%, which means re-ranking methods still can significantly enhance our model's performance. This indicates the robustness of features generated by our method by transferring domain-specific knowledge to guide the optimization of the domain-free feature extractor.

Table 3. Comparison of several methods on CUHK03. Rank-1 accuracy (%) and mAP (%) are shown. We apply the new evaluation protocal on the CUHK03 proposed in [28]. We use ResNet-50 as backbone.

Methods	CUHK03		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
BoW + KISSME [23]	6.4	6.4	25.1	12.2
LOMO + XQDA [12]	12.8	11.5	30.8	17.0
IDE [24]	21.3	19.7	65.2	45.0
PAN [26]	36.3	34.0	71.6	51.5
MultiScale [3]	40.7	37.0	79.2	60.6
SVDNet [16]	41.5	37.2	76.7	56.8
TriNet [6]	50.5	46.5	72.4	53.5
D^2KD	**60.9**	**56.3**	**80.5**	**64.1**

Evaluation on CUHK03 and DukeMTMC-reID. We also evaluate the ReId performance of our method compared to several existing methods on CUHK03 and DukeMTMC-reID. Unlike Market-1501, CUHK03 provides both manually labeled and auto detected bounding boxes of the same identity. We only apply our experiments on the detected bounding box since it's enough to test our model's ability. In DukeMTMC-reID the person bounding boxes of images are manually cropped in a labour-intensive manner.

Table 3 shows the competition between our model and others. From the statistics we can see that our model outperforms other methods by a clear margin. This further validates that our model can maintain more information via applying a knowledge distillation strategy.

4.3 Further Evaluation and Discussion

Experiments on all three above mentioned benchmark datasets show that our method is superior to several state-of-the-art methods these years.

***D^2KD* Improve the Baseline Performance.** In order to further prove the effectiveness of our methods. We compare our D^2KD method with two baselines. Both based on TriNet and choose ResNet-50 as the backbone network, one is optimized with only the triplet loss, while another is optimized with both the triplet loss and the classification loss. Figure 4 illustrates the CMC curve for both above mentioned baselines and our D^2KD methods. Rank-1 to Rank-20 accuracy is shown in the Fig. 4.

***D^2KD* Outperforms Traditional Label Smoothing.** Label smoothing (LS) is a method which assigns less confidence on the ground-truth label and assigns small weights to the other classes.

Formally, in the classification task, we assume $y(x)$ is the ground truth one-hot label for instance x. We assign $y_i(x) = 1$ when x belongs to the i-th person

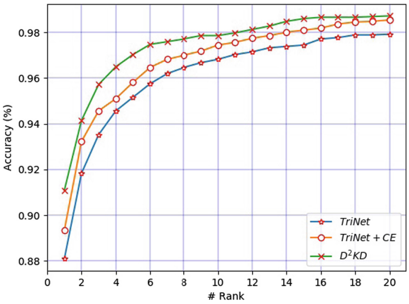

Fig. 4. CMC curve for pure TriNet, TriNet with classification loss and our D^2KD method. All the three methods output a 2048-dimentional feature vector for evaluation. We use ResNet-50 as backbone and single-query setting. The tag "TriNet" means pure ResNet-50 trained with triplet loss. "TriNet+CE" is the baseline when we apply a classification loss on TriNet.

identity otherwise $y_i(x) = 0$. In label smoothing, $y_i(x) = 1 - \epsilon$ for ground-truth label and otherwise $y_i(x) = \epsilon/(C-1)$, where C is total class number, ϵ is usually set to 0.1.

Table 4. Comparison of our D^2KD method to label smoothing (LS)

Methods	Rank-1
TriNet	88.0
TriNet + CE	89.3
TriNet + LS	89.8
D^2KD	91.1

The results are shown in Table 4. Compared to label smoothing (LS), our D^2KD method smooths the label in a dynamic manner. Our method outperforms traditional label smoothing.

From the results, we have the following observations: (1) TriNet with cross entropy loss can achieve higher performance. We can infer that applying the classification task can help to generate more discriminative features. (2) Our Deep Domain Knowledge Distillation (D^2KD) method outperforms both two baselines and label smoothing (LS). This indicates that our D^2KD method which smooths the label in a dynamic manner can lead to a better performance in the person re-identification task.

5 Conclusion

In this paper, we propose a method named Deep Domain Knowledge Distillation (D^2KD) by learning to provide a higher soft bound for the subnetwork to optimize and it is easy to implement. By transferring domain-specific knowledge to the domain-free network, the base model is capable of extracting more discriminative features. Extensive comparative evaluations on three person re-identification benchmark datasets were conducted to validate the advantages of the proposed D^2KD method over a wide range of models on three benchmark datasets. We hope that in the future work we can find a more satisfied higher soft bound for the base model to optimize.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015). https://doi.org/10.1109/CVPR.2015.7299016
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Advances in Neural Information Processing Systems, pp. 737–744 (1994)
3. Chen, Y., Zhu, X., Gong, S., et al.: Person re-identification by deep learning multi-scale representations (2018). https://doi.org/10.1109/ICCVW.2017.304
4. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
6. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint: arXiv:1703.07737 (2017)
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint: arXiv:1503.02531 (2015)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint: arXiv:1412.6980 (2014)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012). https://doi.org/10.1145/3065386
10. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 384–393 (2017)
11. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014). https://doi.org/10.1109/CVPR.2014.27
12. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015). https://doi.org/10.1109/CVPR.2015.7298832
13. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Trans. Image Process. **26**(7), 3492–3506 (2017). https://doi.org/10.1109/TIP.2017.2700762. ISSN: 1057-7149
14. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)

15. Sarfraz, M.S., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. arXiv preprint: arXiv:1711.10378 (2017). https://doi.org/10.1109/CVPR.2018.00051
16. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. arXiv preprint **1**(6) (2017). https://doi.org/10.1109/ICCV.2017.410
17. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling. arXiv preprint: arXiv:1711.09349 (2017)
18. Varior, R.R., Haloi, M., Wang, G.: Gated Siamese convolutional neural network architecture for human re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part VIII. LNCS, vol. 9912, pp. 791–808. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_48
19. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A Siamese long short-term memory architecture for human re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part VII. LNCS, vol. 9911, pp. 135–153. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_9
20. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. arXiv preprint: arXiv:1805.03344 (2018). https://doi.org/10.1109/CVPR.2018.00226
21. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 34–39. IEEE (2014)
22. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1239–1248 (2016). https://doi.org/10.1109/CVPR.2016.139
23. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015). https://doi.org/10.1109/ICCV.2015.133
24. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. arXiv preprint: arXiv:1610.02984 (2016)
25. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person reidentification. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **14**(1), 13 (2017). https://doi.org/10.1145/3159171. ISSN: 1551-6857
26. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. arXiv preprint: arXiv:1707.00408. https://doi.org/10.1109/TCSVT.2018.2873599 (2017). ISSN: 1051-8215
27. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-dentification baseline in vitro. https://doi.org/10.1109/ICCV.2017.405. arXiv preprint: arXiv:1701.07717 (2017)
28. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3652–3661. IEEE (2017). https://doi.org/10.1109/CVPR.2017.389