A Discriminatively Learned CNN Embedding for Person Reidentification

ZHEDONG ZHENG and LIANG ZHENG, University of Technology Sydney
YI YANG, University of Technology Sydney and State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences

In this article, we revisit two popular convolutional neural networks in person re-identification (re-ID): verification and identification models. The two models have their respective advantages and limitations due to different loss functions. Here, we shed light on how to combine the two models to learn more discriminative pedestrian descriptors. Specifically, we propose a Siamese network that simultaneously computes the identification loss and verification loss. Given a pair of training images, the network predicts the identities of the two input images and whether they belong to the same identity. Our network learns a discriminative embedding and a similarity measurement at the same time, thus taking full usage of the re-ID annotations. Our method can be easily applied on different pretrained networks. Albeit simple, the learned embedding improves the state-of-the-art performance on two public person re-ID benchmarks. Further, we show that our architecture can also be applied to image retrieval. The code is available at https://github.com/layumi/2016_person_re-ID.

CCS Concepts: • **Computing methodologies → Visual content-based indexing and retrieval**; **Image representations**;

Additional Key Words and Phrases: Person reidentification, convolutional neural networks

ACM Reference format:
Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A Discriminatively Learned CNN Embedding for Person Reidentification. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 1, Article 13 (December 2017), 20 pages.
https://doi.org/10.1145/3159171

1 INTRODUCTION

Person reidentification (re-ID) is usually viewed as an image retrieval problem, which aims to match pedestrians from multiple cameras [25, 41, 53—57]. Given a person-of-interest (query), person re-ID determines whether the person has been observed by another camera. Recent progress in this area has been due to two factors: (1) the availability of the large-scale pedestrian datasets (compared to small datasets, large-scale datasets contain the common visual variance of pedestrian and provide a comprehensive evaluation [18, 52]) and (2) the learned pedestrian descriptor using a convolutional neural network (CNN).

Part of this work was done when Z. Zheng was a visiting student at State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China.
Authors' addresses: Z. Zheng and L. Zheng, University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia; emails: {zdzheng12, liangzheng06}@gmail.com; Y. Yang, University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia; and State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, 4# South Fourth Street, Zhong Guan Cun, Beijing 100190, China; email: yee.i.yang@gmail.com.

https://doi.org/10.1145/3159171

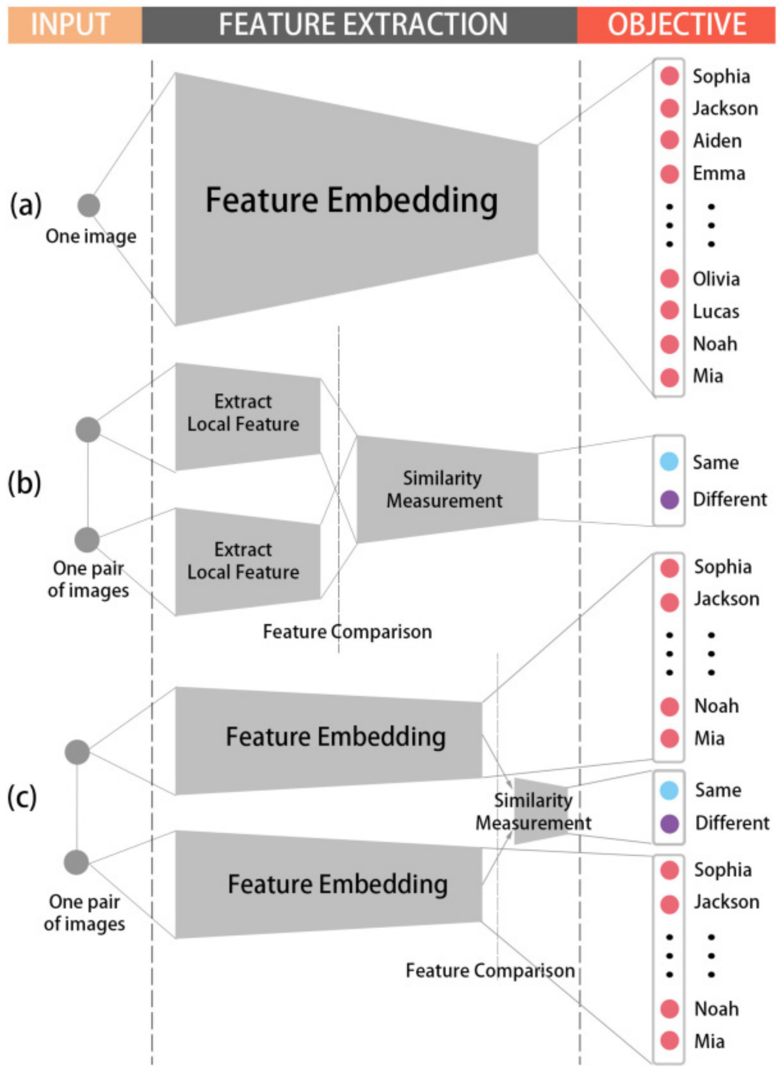

Fig. 1. The difference between identification models, verification models, and our model. Gray blocks represent nonlinear functions by CNN. (a) Identification models treat person re-ID as a multiclass recognition task, which takes one image as input and predicts its identity. (b) Verification models treat person re-ID as a binary class recognition task or a similarity regression task, which take a pair of images as input and determine whether they belong to the same person or not. Here we only show a binary class recognition case. (c) Our model fuses the identification and verification models together. Given one pair of images, we can know who is in the image separately and whether two images depict the same person or not.

Recently, the CNN has shown potential for learning state-of-the-art feature embeddings or deep metrics [5, 18, 38, 43, 44, 47, 56]. As shown in Figure 1, there are two major types of CNN structures: verification models and identification models. The two models are different concerning input, feature extraction, and loss function for training. Our motivation is to combine the strengths of the two models and learn a more discriminative pedestrian embedding.

Table 1. The Advantages and Disadvantages of Verification and Identification Models are Listed. We Assume Sufficient Training Data in All Models. Our Model Takes the Advantages of the Two Models

Method	Strong Label	Similarity Estimation	Re-ID Performance
Verification Models	×	✓	fair
Identification Models	✓	×	good
Our Model	✓	✓	good

Verification models take a pair of images $(x1, x2)$ as input and predict $f(x1, x2) \rightarrow s$; s is a binary label, whether the two inputs belong to the same person or not. If two inputs depict the same person, $s = 1$ and otherwise $s = 0$. Many previous works treat person re-ID as a binary class classification task [1, 18, 43] or a similarity regression task [38, 47]. The verification network forces two images of the same person to be mapped to nearby points in the feature space. If the images are of different people, the points are far apart. However, the major problem in the verification models is that they only use weak re-ID labels (same/different) [56] and do not take all of the annotated information into consideration. Therefore, the verification network lacks the consideration of the relationship between the image pairs and other images in the dataset.

In the attempt to take full advantages of the re-ID labels, identification models take a single image (or a batch of images) as input x and predict the predefined identity label $f(x) \rightarrow t$. The identification models directly learn the nonlinear functions from an input image to the person ID, and the cross-entropy loss is usually used following the final layer [56, 59]. During testing, the feature is extracted from a fully connected (FC) layer and then normalized. The similarity of two images is thus computed by the Euclidean distance between their normalized CNN embeddings. The major drawback of the identification model is that the training objective is different from the testing procedure—that is, it does not account for the similarity measurement between image pairs, which can be problematic during the pedestrian retrieval process.

The observations mentioned earlier demonstrate that the two types of models have complementary advantages and limitations, as shown in Table 1. Motivated by these properties, this work proposes to combine the strengths of the two networks and leverage their complementary nature to improve the discriminative ability of the learned embeddings. The proposed model is a Siamese network that predicts person identities and similarity scores at the same time. Compared to previous networks, we take full advantage of the annotated data regarding pairwise similarity and image identities. During testing, the final convolutional activations are extracted for Euclidean distance–based pedestrian retrieval. Our contributions include the following:

- We propose a Siamese network that has two losses: identification loss and verification loss. This network simultaneously learns a discriminative CNN embedding and a similarity metric, thus improving pedestrian retrieval accuracy.
- We report competitive accuracy compared to the state-of-art methods on two large-scale person re-ID datasets (Market1501 [52] and CUHK03 [18]) and one instance retrieval dataset (Oxford5k [27]).

The article is organized as follows. We first review some related work in Section 2. In Section 3, we describe how we combine the two losses and define the CNN structure. The implementation details are provided. In Section 4, we present the experimental results on two large-scale person re-ID datasets and one instance retrieval dataset. We conclude in Section 5.

2 RELATED WORK

In this section, we first review handcrafted systems and then describe deeply learned systems for person re-ID. The deeply learned systems are mainly based on the verification model and the identification model.

2.1 Handcrafted Systems

Discriminative information plays an important role in multimedia retrieval [8, 46]. Some pioneering works focus on finding discriminative handcrafted features. Local features such as color histogram [50], LBP [24], and LOMO [19] are widely studied. Houle et al. [13] combine the local feature with the global feature to conduct retrieval. Zheng et al. [52] explore the color name descriptor and bag-of-words method on large-scale datasets. Yang et al. [45] use Gaussian of Gaussian feature [23] to conduct semisupervised learning. Chang and Yang [4] analyze the semisupervised features. Recently, Lisanti et al. [20] proposed a kernel canonical correlation analysis to combine multiple features.

In addition to finding invariant features, Zheng et al. [58] formulate the person re-ID as a distance comparison problem. Koestinger et al. [15] propose the KISSME metric learning method based on Mahalanobis distance. Further, Liao et al. [19] extend the Bayesian face and KISSME to learn a discriminative subspace. Zhang et al. [48] propose a discriminative null subspace. Moreover, Zhang et al. [49] learn a specific SVM for each training identity to discriminate between different identities.

2.2 Deeply Learned Systems

Verification models. In 1993, Bromley et al. [3] first used verification models to deep metric learning in signature verification. Verification models usually take a pair of images as input and output a similarity score by calculating the cosine distance between low-dimensional features, which can be penalized by the contrastive loss. Recently, researchers have begun to apply verification models to person re-ID with a focus on data augmentation and image matching. Yi et al. [47] split a pedestrian image into three horizontal parts and train three part-CNNs to extract features. The similarity of two images is computed by the cosine distance of their features. Similarly, Cheng et al. [7] split the convolutional map into four parts and fuse the part features with the global features. Li et al. [18] add a patch matching layer that multiplies the activation of two images in different horizontal stripes. They use it to find similar locations and treat similarity comparison as binary classification penalized by Softmax loss. Later, Ahmed et al. [1] improved the verification model by adding a different matching layer that compares the activation of two images in neighboring pixels. In addition, Wu et al. [43] use smaller filters and a deeper network to extract features. Varior et al. [38] combine CNN with some gate functions, similar to long short-term memory (LSTM) in spirit, which aims to adaptively focus on the similar parts of input image pairs. But it is limited by the computational inefficiency because the query image has to pair with every gallery image to pass through the network. Moreover, Ding et al. [9] use triplet samples for training the network that considers the images from the same people and the different people at the same time.

Identification models. Recent datasets such as CUHK03 [18] and Market1501 [52] provide large-scale training sets, which make it possible to train a deeper classification model without overfitting. Every identity has 9.6 training images on average in CUHK03 [18] and 17.2 images in Market1501 [52]. CNN can learn discriminative embeddings by itself without part matching. Zheng et al. [51, 56] directly use a conventional fine-tuning approach on Market1501 [52], PRW [56], and MARS [51], and outperform many recent results. Additionally, Zheng et al. [59] introduce the GAN model to generate more pedestrian images to regularize the network.

Verification-identification models. In face recognition, the "DeepID networks" train the network with the verification and identification losses [33, 34], which is similar to our network. Sun et al. [33] jointly train face identification and verification. Then more verification supervision is added into the model [34].

Our method is different from their models in the following aspects. First, in face recognition, the training dataset contains 202,599 face images of 10,177 identities [33], whereas the current largest person re-ID training dataset contains 12,936 images of 751 identities [52]. DeepID networks apply contrastive loss to the verification problem, whereas our model uses the cross-entropy loss. We find that the contrastive loss leads to overfitting when the number of images is limited. In the experiment, we show that the proposed method learns more robust person representative and outperforms using contrastive loss. Second, dropout [32] cannot be applied on the embedding before the contrastive loss, which introduces zero values at random locations. However, we can add dropout regularization on the embedding in the proposed model. Third, the DeepID networks are trained from scratch, whereas our model benefits from the networks pretrained on ImageNet [30]. Finally, we evaluate our method on the tasks of person re-ID and instance retrieval, providing more insights into the verification-classification models.

Here we mention a contemporary work to us—that of Geng et al. [10].[1] In this article, we provide more working mechanism on the combination of the two losses.

3 PROPOSED METHOD

3.1 Preview

Figure 2(a) and (b) illustrate the relational graph built by verification and identification models. We were inspired by Song et al. [26] to visualize the relationship. In a sample batch of size $m = 10$, blue edges represent the positive pairs (the same person) and red edges represent the negative pairs (different persons). The dotted edges denote implicit relationships built by the identification loss, and the solid edges denote explicit relationships built by the verification loss.

In verification models, there are several operations between the two inputs. The explicit relationship between data is built by the pairwise comparison, such as part matching [1, 18] or contrastive loss [11]. In identification models, the input is independent of each other. But there is an implicit relationship between the learned embeddings built by the cross-entropy loss. Specifically, the cross-entropy loss can be formulated as $loss = -log(p_{gt})$, where $p_{gt} = W_{gt} f_i$. W is the weight of the linear function. f_m, f_n are the embeddings of the two images x_m, x_n from the same class k. To maximize $W_k f_m$, $W_k f_n$, the network converges when f_m and f_n have similar vector direction with W_k. In Liu et al. [22], similar observation and visualization are shown. Thus, the learned embeddings are eventually close for images within the same class and far away for images in the different classes. The relationship is implicitly built between x_m, x_n and bridged by the weight W_k.

Due to the usage of the weak labels, verification models take limited relationships into consideration. However, classification models do not explicitly consider similarity measurements. Figure 2(c) illustrates how our model works in a batch. We benefit from simultaneously considering the verification and identification losses. The proposed model thus combines the strength of the two models (see Table 1).

3.2 Overall Network

Our network is a convolutional Siamese network that combines the verification and identification losses. Figure 3 briefly illustrates the architecture of the proposed network. Given an input pair

[1] The work of Geng et al. [10] was submitted to arXiv on November 16, 2016; this work was submitted to arXiv on November 17, 2016.

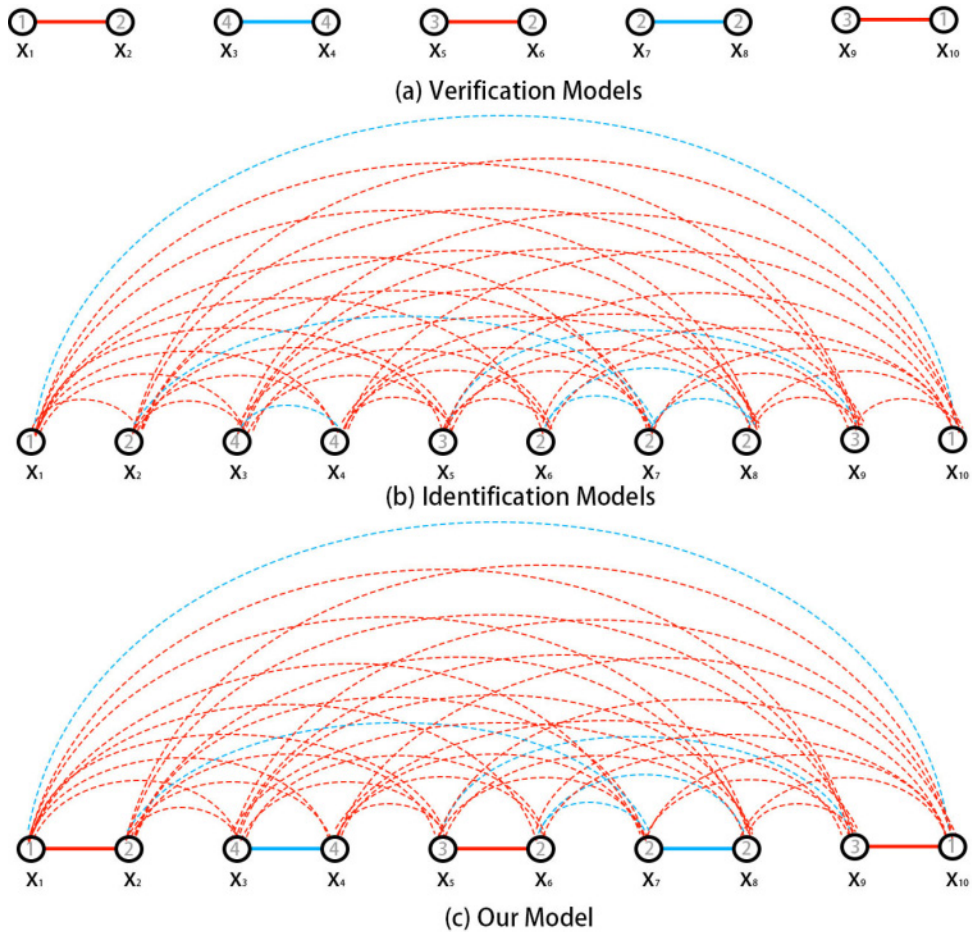

Fig. 2. Illustration for a training batch. The number in the circle is the identity label. Blue and red edges represent whether the image pair depicts the same identity or not. Dotted edges represent implicit relationships, and solid edges represent explicit relationships. Our model combines the strengths of the two models.

of images resized to 227×227, the proposed network simultaneously predicts the IDs of the two images and the similarity score. The network consists of two ImageNet [30] pretrained CNN models, three additional convolutional layers, one square layer, and three losses. It is supervised by the identification label t and the verification label s. The pretrained CNN model can be CaffeNet [17], VGG16 [31], or ResNet-50 [12], from which we have removed the final FC (FC) layer. The re-ID performance of the three models is comprehensively evaluated in Section 4. Here we do not provide detailed descriptions of the architecture of the CNN models and only take CaffeNet as an example in the following sections. The three optimization objectives include two identification losses and one verification loss. We use the final convolutional activations f as the discriminative descriptor for person re-ID, which is directly supervised by three objectives.

3.3 Identification Loss

There are two CaffeNets in our architecture. They share weights and predict the two identity labels of the input image pair simultaneously. To fine tune the network on a new dataset, we replace the

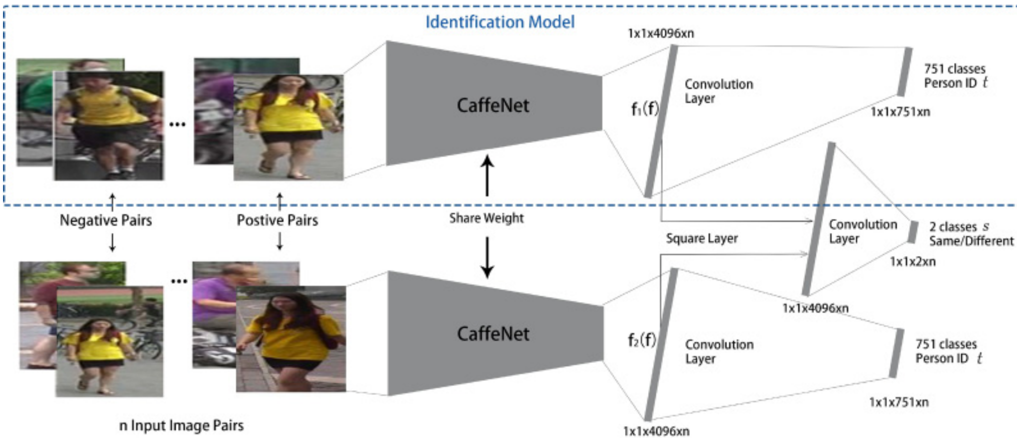

Fig. 3. The proposed model structure. Given n pairs of images of size 227×227, two identical CaffeNet models are used as the nonlinear embedding functions and output 4,096-dim embeddings f_1, f_2. Then f_1, f_2 are used to predict the identity t of the two input images, respectively, and also predict the verification label s jointly. We introduce a nonparametric layer called the *square layer* to compare high-level features f_1, f_2. Finally, the softmax loss is applied on the three objectives.

final FC layer (1,000-dim) of the pretrained CNN model with a convolutional layer. The number of the training identities in Market1501 is 751. Thus, this convolutional layer has 751 kernels of size $1 \times 1 \times 4{,}096$ connected to the output f of CaffeNet, and then we add a softmax unit to normalize the output. The size of the result tensor is $1 \times 1 \times 751$. The rectified linear unit (ReLU) is not added after this convolution. Similar to conventional multiclass recognition approaches, we use the cross-entropy loss for identity prediction, which is

$$\hat{p} = softmax(\theta_I \circ f), \tag{1}$$

$$\text{Identif}(f, t, \theta_I) = \sum_{i=1}^{K} -p_i \log(\hat{p}_i). \tag{2}$$

Here, $\circ$ denotes the convolutional operation. f is a $1 \times 1 \times 4,096$ tensor, t is the target class, and θ_I denotes the parameters of the added convolutional layer. $\hat{p}$ is the predicted probability, and p_i is the target probability. $p_i = 0$ for all i except $p_t = 1$.

3.4 Verification Loss

Whereas some previous works contain a matching function in the intermediate layers [1, 18, 38], our work directly compares the high-level features f_1, f_2 for similarity estimation. The high-level feature from the fine-tuned CNN has shown a discriminative ability [51, 56], and it is more compact than the activations in the intermediate layers. Thus, in our model, the pedestrian descriptors f_1, f_2 in the identification model are directly supervised by the verification loss. As shown in Figure 3, we introduce a nonparametric layer called the *square layer* to compare the high-level features. It takes two tensors as inputs and outputs one tensor after subtracting and squaring element-wisely. The square layer is denoted as $f_s = (f_1 - f_2)^2$, where f_1, f_2 are the 4,096-dim embeddings and f_s is the output tensor of the square layer.

We then add a convolutional layer and the softmax output function to embed the resulting tensor f_s to a 2-dim vector $(\hat{q}_1, \hat{q}_2)$ that represents the predicted probability of the two input images

Table 2. Results on Market1501 [52] by Identification Loss and Verification Loss Individually and Jointly

Method	mAP	rank-1
CaffeNet (V)	22.47	41.24
CaffeNet (I)	26.79	50.89
CaffeNet (I+V)	**39.61**	**62.14**
VGG16 (V)	24.29	42.99
VGG16 (I)	38.27	65.02
VGG16 (I+V)	**47.45**	**70.16**
ResNet-50 (V)	44.94	64.58
ResNet-50 (I)	51.48	73.69
ResNet-50 (I+V)	**59.87**	**79.51**

Note: "I" and "V" denote the identification loss and verification loss, respectively.

belonging to the same identity. $\hat{q}_1 + \hat{q}_2 = 1$. The convolutional layer takes f_s as input and filters it with two kernels of size $1 \times 1 \times 4,096$. The ReLU is not added after this convolution. We treat pedestrian verification as a binary classification problem and use the cross-entropy loss that is similar to the one in the identification loss, which is

$$\hat{q} = softmax(\theta_S \circ f_s), \tag{3}$$

$$\text{Verif}(f_1, f_2, s, \theta_S) = \sum_{i=1}^{2} -q_i \log(\hat{q}_i). \tag{4}$$

Here, f_1, f_2 are the two tensors of size $1 \times 1 \times 4,096$. s is the target class (same/different), θ_S denotes the parameters of the added convolutional layer, and $\hat{q}$ is the predicted probability. If the image pair depicts the same person, $q_1 = 1, q_2 = 0$; otherwise, $q_1 = 0, q_2 = 1$.

Departing from Sun et al. [33], we do not use the contrastive loss [11]. On the one hand, the contrastive loss, as a regression loss, forces the same class embeddings to be as close as possible. It may make the model overfitting because the number of training of each identity is limited in person re-ID. On the other hand, dropout [32], which introduces zero values at random locations, cannot be applied on the embedding before the contrastive loss. But the cross-entropy loss in our model can work with dropout to regularize the model. In Section 4, we show that the result using contrastive loss is 4.39% and 6.55% lower than the one using the cross-entropy loss on rank-1 accuracy and mean average precision (mAP), respectively.

3.5 Identification Versus Verification

The proposed network is trained to minimize the three cross-entropy losses jointly. To figure out which objective contributes more, we train the identification model and verification model separately. Following the learning rate setting in Section 3.6, we train the models until convergence. We also train the network with the two losses jointly until two objectives both converge. As the quantitative results are shown in Table 2, the fine-tuned CNN model with two kinds of losses outperforms the one trained individually. This result has been confirmed on the three different network structures.

Further, we visualize the intermediate feature maps that are trained using ResNet-50 [12] as the pretrained model and try to find the differences between identification loss and verification loss. We select three test images in Market1501. One image is considered to be well detected, and

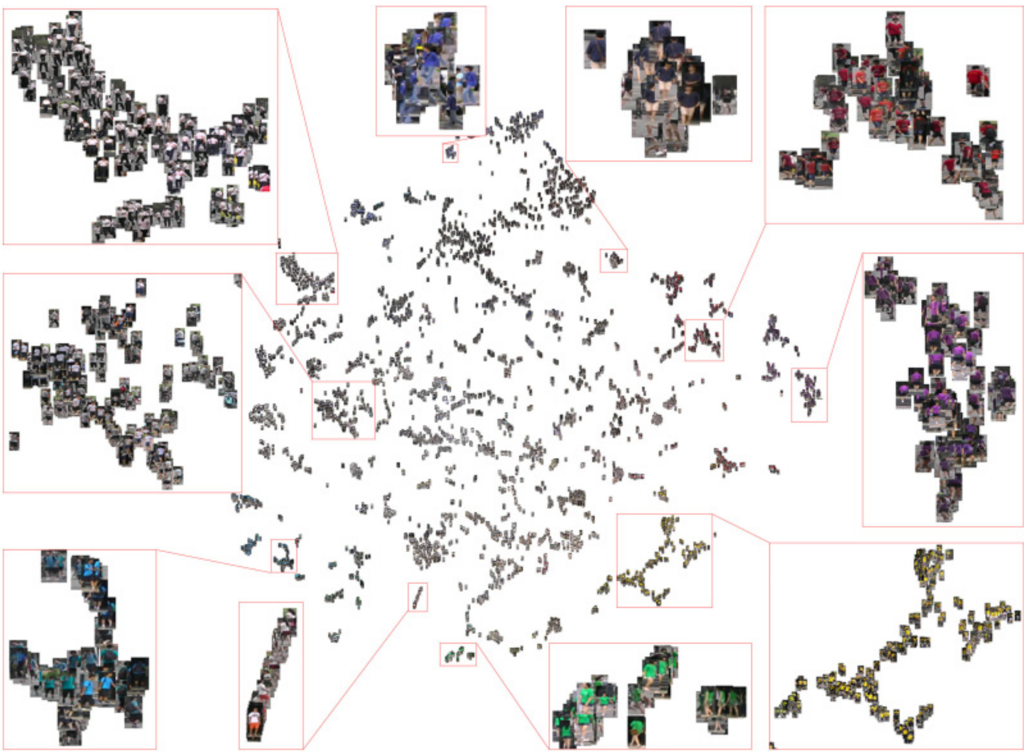

Fig. 4. Barnes-Hut t-SNE visualization [37] of our embedding on a test split (354 identity, 6,868 images) of Market1501. Best viewed when zoomed in. We find that the color is the major clue for the person re-ID, and our learned embedding is robust to some viewpoint variations.

the other two images are not well aligned. Given one image as input, we get its activation in the intermediate layer "res4fx," the size of which is 14 × 14. We visualize the sum of several activation maps. The local patterns (i.e., clothing color and texture) are important clues to person re-ID. This is basically what we want the model to learn. In our work, for both two losses, the network can learn the important local information, which implies the effectiveness of our model. As shown later in Figure 5, the identification and the verification networks exhibit different activation patterns to the pedestrian. We find that if we use only one kind of loss, the network tends to find one discriminative part. The proposed model takes the advantages of both networks, so the new activation map is mostly a union of the two individual maps. This also illustrates the complementary nature of the two baseline networks. The proposed model makes more neurons activated.

Moreover, as shown in Figure 4, we visualize the embedding by plotting them to the 2D map. Regarding Figure 5, we find that the network usually has strong attention on the center part of the human (usually clothes), and it also illustrates that the color of the clothes is the major clue for the person re-ID.

3.6 Training and Optimization

Input preparation. We resize all training images to 256 × 256. The mean image computed from all training images is subtracted from all of the images. During training, all of the images are randomly cropped to 227 × 227 for CaffeNet [17] and mirrored horizontally. For ResNet-50 [12]

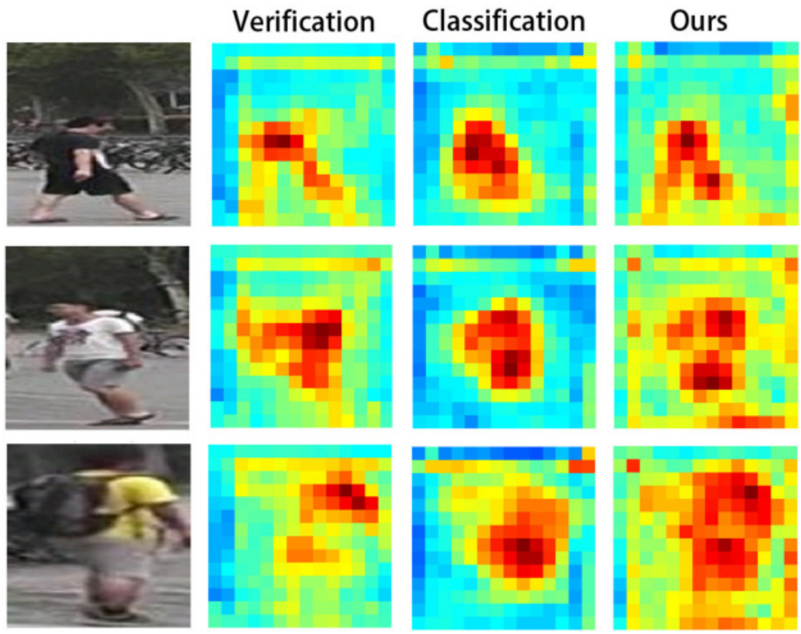

Fig. 5. Visualization of the activation maps in the ResNet-50 [12] model trained by the two losses. The identification and the verification networks exhibit different activation patterns to the pedestrian. The proposed model takes the advantages of both networks, and the new activation map is almost a union of the two individual maps. Our model activates more neurons.

and VGG16 [31], we randomly crop images to 224×224. We shuffle the dataset and use a random order of the images. Then we sample another image from the same/different class to compose a positive/negative pair. The initial ratio between negative pairs and positive pairs is 1:1 to alleviate the prediction bias, and we multiply it by a factor of 1.01 every epoch until it reaches 1:4 since the number of positive pairs is so limited that the network risks overfitting.

Training. We use the MatConvNet [40] package for training and testing the embedding with CaffeNet [17], VGG16 [31], and ResNet-50 [12], respectively. The maximum number of training epochs is set to 75 for ResNet-50, 65 for VGG16net, and 155 for CaffeNet. The batch size (in image pairs) is set to 128 for CaffeNet, and 48 for VGG16 and ResNet-50. The learning rate is initialized as 0.001 and then set to 0.0001 for the final five epochs. We adopt the mini-batch stochastic gradient descent (SGD) to update the parameters of the network. There are three objectives in our network. Therefore, we first compute all gradients produced by every objective respectively and add the weighted gradients together to update the network. We assign a weight of 1 to the gradient produced by the verification loss and 0.5 for the two gradients produced by two identification losses. Moreover, we insert the dropout function [32] before the final convolutional layer.

Testing. We adopt an efficient method to extract features as well as the activation of the intermediate layer. Because two CaffeNets share weights, our model has nearly the same memory consumption with the pretrained model. Thus, we extract features by only activating one fine-tuned model. Given a 227×227 image, we feed forward the image to one CaffeNet in our network and obtain a 4,096-dim pedestrian descriptor f. Once the descriptors for the gallery sets are obtained, they are stored offline. Given a query image, its descriptor is extracted online. We sort the cosine distance between the query and all gallery features to obtain the final ranking result. Note that the cosine distance is equivalent to Euclidean distance when the feature is L2 normalized.

4 EXPERIMENTS

We mainly verify the proposed model on two large-scale datasets: Market1501 [52] and CUHK03 [18]. We report the results trained by three network structures. In addition, we also report the result on a Market1501+500k dataset [52]. Meanwhile, the proposed architecture is also applied on the image retrieval task. We modify our model and test it on a popular image retrieval dataset: Oxford Buildings [27]. The performance is comparable to the state of the art. The code is available at https://github.com/layumi/2016_person_re-ID.

4.1 Dataset

Market1501. Market1501 [52] contains 32,668 annotated bounding boxes of 1,501 identities. Images of each identity are captured by at most six cameras. According to the dataset setting, the training set contains 12,936 cropped images of 751 identities, and the testing set contains 19,732 cropped images of 750 identities and distractors. They are directly detected by the deformable part model (DPM) instead of using hand-drawn bounding boxes, which is closer to the realistic setting. For each query, we aim to retrieve the ground-truth images from the 19,732 candidate images.

The searching pool (gallery) is important to person re-ID. In the realistic setting, the scale of the gallery is usually large. The distractor dataset of Market1501 provides an extra 500,000 bounding boxes, consisting of false alarms on the background as well as the persons not belonging to any of the original 1,501 identities [52]. When testing, we add the 500k images to the original gallery, which makes the retrieval more difficult.

CUHK03. The CUKH03 dataset [18] contains 14,097 cropped images of 1,467 identities collected in the CUHK campus. Each identity is observed by two camera views and has 4.8 images on average for each view. The author provides two kinds of bounding boxes. We evaluate our model on the bounding boxes detected by DPM, which is closer to the realistic setting. Following the setting of the dataset, the dataset is partitioned into a training set of 1,367 persons and a testing set of 100 persons. The experiment is repeated with 20 random splits. Both the single-shot and multiple-shot results will be reported.

Oxford5k. Oxford Buildings [27] consists of 5,062 images collected from the Internet and corresponding to particular Oxford landmarks. Some images have complex structures and may contain other buildings. The images corresponding to 11 Oxford landmarks are manually annotated, and a set of 55 queries for 11 different landmarks are provided. This benchmark contains many high-resolution images, and the mean image size of this dataset is 851×921.

We use the rank-1 accuracy and mAP for performance evaluation on Market1501 (+100k) and CUHK03, whereas on Oxford, we use mAP.

4.2 Person Re-ID Evaluation

Comparison with the CNN baseline. We train the baseline networks according to the conventional fine-tuning method [56]. The baseline networks are pretrained on ImageNet [30] and fine tuned to predict the person identities. As shown in Table 3, we obtain 50.89%, 65.02%, and 73.69% rank-1 accuracy by CaffeNet [17], VGG16 [31], and ResNet-50 [12], respectively, on Market1501. Note that using the baseline alone exceeds many previous works. Our model further improves these baselines on Market1501. The improvement can be observed on three network architectures. To be specific, we obtain 11.25%, 5.14%, and 5.82% improvement, respectively, using CaffeNet [17], VGG16 [31], and ResNet-50 [12] on Market1501. Similarly, we observe 35.8%, 49.1%, and 71.5% baseline rank-1 accuracy on CUHK03 in single-shot setting. As show in Table 4, these baseline

Table 3. Comparison With the State-of-the-Art Results on the Market1501 Dataset

Method	Single Query		Multiple Query	
	Rank-1	mAP	Rank-1	mAP
BoW + KISSME [52]	44.42	20.76	—	—
Multiregion CNN [36]	45.58	26.11	56.59	32.26
CAN [21]	48.24	24.43	—	—
Fisher Network [42]	48.15	29.94	—	—
SL [6]	51.90	26.35	—	—
S-LSTM [39]	—	—	61.6	35.3
DNS [48]	55.43	29.87	71.56	46.03
Gate Reid [38]	65.88	39.55	76.04	48.45
CaffeNet Baseine [17]	50.89	26.79	59.80	36.50
Ours (CaffeNet)	62.14	39.61	72.21	49.62
VGG16 Baseline [31]	65.02	38.27	74.14	52.25
Ours (VGG16)	70.16	47.45	77.94	57.66
ResNet-50 Baseline [12]	73.69	51.48	81.47	63.95
Ours (ResNet-50)	**79.51**	**59.87**	**85.84**	**70.33**

Note: We also provide the results of the fine-tuned CNN baseline. The mAP and rank-1 precision are listed. SQ and MQ denote single query and multiply queries, respectively.

Table 4. Comparison With the State-of-the-Art Results Reported on the CUHK03 Dataset Using the Single-Shot Setting

Method	Rank-1	Rank-5	Rank-10	mAP
KISSME [16]	11.7	33.3	48.0	—
DeepReID [18]	19.9	49.3	64.7	—
BoW+HS [52]	24.3	—	—	—
LOMO+XQDA [19]	46.3	78.9	88.6	—
DNS [48]	54.7	80.1	88.3	—
CaffeNet Baseline	35.8	65.3	77.96	42.6
Ours (CaffeNet)	59.8	88.3	94.2	65.8
VGG16 Baseline	49.1	78.4	87.2	55.7
Ours (VGG16)	71.8	93.0	97.1	76.5
ResNet-50 Baseline	71.5	91.5	95.9	75.8
Ours (ResNet-50)	**83.4**	**97.1**	**98.7**	**86.4**

Note: The mAP and rank-1 accuracy are listed.

results exceed some previous works as well. We further get 14.0%, 22.7%, and 11.9% improvement on the baseline by our method.

These results show that our method can work with different networks and improve their results. It indicates that the proposed model helps the network learn more discriminative features.

Cross entropy loss vs. contrastive loss. We replace the cross-entropy loss with the contrastive loss as used in DeepID network. However, we find a 4.39% and 6.55% drop in rank-1 and mAP. The ResNet-50 model using the contrastive loss has 75.12% rank-1 accuracy and 53.32% mAP. We speculate that the contrastive loss tends to overfit on the re-ID dataset because no

Table 5. Comparison With the State-of-the-Art Methods on the CUHK03 Dataset Under the Multishot Setting

Method	Rank-1	Rank-5	Rank-10	mAP
S-LSTM [39]	57.3	80.1	88.3	46.3
Gate-SCNN [38]	68.1	88.1	94.6	58.8
CaffeNet Baseline	43.3	63.5	76.8	37.2
Ours (CaffeNet)	67.2	86.2	92.3	61.5
VGG16 Baseline	58.8	80.2	87.3	51.0
Ours (VGG16)	78.8	91.8	95.4	73.9
ResNet-50 Baseline	77.1	89.6	93.9	73.1
Ours (ResNet-50)	**88.3**	**95.7**	**97.8**	**85.0**

Note: The multishot setting uses all images in the other camera as the gallery. The mAP and rank-1 accuracy are listed.

regularization is added to the verification. Cross-entropy loss designed in our model can work with the dropout function and avoid the overfitting.

Comparison with the state of the art. As shown in Table 3, we compare our method to other state-of-the-art algorithms in terms of mAP and rank-1 accuracy on Market1501. We report both single- and multiple-query evaluation results. Our model (CaffeNet) achieves 62.14% rank-1 accuracy and 39.61% mAP, which is comparable to the state-of-the-art 65.88% rank-1 accuracy and 39.55% mAP [38]. Our model using ResNet-50 produces the best performance—79.51% in rank-1 accuracy and 59.87% in mAP—which outperforms other state-of-the-art algorithms.

For CUHK03, we evaluate our method in the single-shot setting as shown in Table 4. There is only one right image in the searching pool. In the evaluation, we randomly select 100 images from 100 identities under the other camera as the gallery. The proposed model yields 83.4% rank-1 and 86.4% mAP and outperforms the state-of-the-art performance.

As shown in Table 5, we also report the results in the multishot setting, which uses all images from the other camera as the gallery, and the number of the gallery images is about 500. We believe that this setting is much closer to image retrieval and alleviates the unstable effect caused by the random searching pool under single-shot settings. Figure 7 presents some re-ID samples on the CUHK03 dataset. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right. Most ground-truth candidate images are correctly retrieved. Although the model retrieves some incorrect candidates in the third row, we find that it is a reasonable prediction since the man with the red hat and blue coat is similar to the query. The proposed model yields 88.3% rank-1 and 85.0% mAP and also outperforms the state-of-the-art performance in the multishot setting.

Results between camera pairs. CUHK03 [18] only contains two camera views. Thus, this experiment is evaluated on Market1501 [52] since it contains six different cameras. We provide the re-ID results between all camera pairs in Figure 6. Cross-camera variations lay difficulties in finding the queried pedestrian. In our work, the lowest cross-camera result obtained by our method still achieves about 45% rank-1 accuracy, which demonstrates the robustness of the proposed method. Note that Cam-6 is a 720×576 low-resolution camera. Although low resolution usually compromises the cross-camera re-ID accuracy, we still achieve relatively high results between Cam-6 and the other cameras. We also compute the cross-camera average mAP and average rank-1 accuracy: 48.42% and 54.42%, respectively. Compared to the results reported previously (i.e., 10.51% and 13.72% in Zheng et al. [52]), our method largely improves the performance and observes a

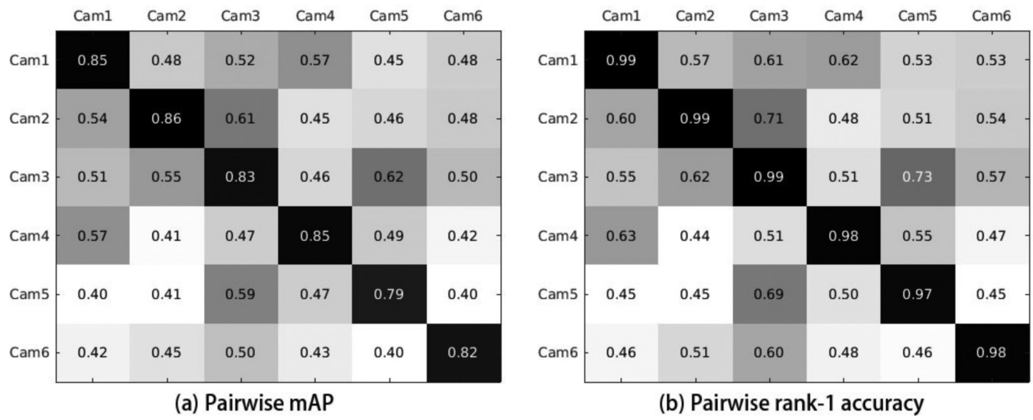

Fig. 6. Re-ID performance between camera pairs on Market1501: mAP (a) and rank-1 accuracy (b). Cameras on the vertical and horizontal axis correspond to the probe and gallery, respectively.

smaller standard deviation between cameras. It suggests that the discriminatively learned embedding works under different viewpoints.

Further, Figure 4 shows the Barnes-Hut t-SNE visualization [37] on the learned embeddings of our model. By the clustering algorithm, the persons wearing the similar-color clothes are quit clustered together and are apart from other persons. The learned pedestrian descriptor pays more attention to the color, and it is robust to some illusion and viewpoint variations. In the realistic setting, we believe that color provides the most important information to figure out the person.

Large-scale experiments. The Market1501 dataset also provides an additional distractor set with 500k images to enlarge the gallery. In general, more candidate images may confuse the image retrieval. The re-ID performance of our model (ResNet) on the large-scale dataset is presented in Table 6. As the searching pool gets larger, the accuracy drops. With the gallery size of $500,000 + 19,732$, we still achieve 68.26% rank-1 accuracy and 45.24% mAP. A relative smaller drop of $24.4\%(1 - 45.24\%/59.87\%)$ on mAP is observed, compared to a drop of $37.88\%(1 - 8.66\%/13.94\%)$ in our previous work [52]. In addition, we compare our result to the performance of the ResNet baseline. As shown in Figure 8, it is interesting that the re-ID precision of our model decreases more quickly compared to the baseline model. We speculate that the Market1501 training set is relatively small in covering the pedestrian variations encountered in a much larger test set. We note that the extra 500k dataset was collected in a different time (the same location) with the Market1501 dataset, so the transfer effect is large enough that the learned embedding is inferior to the baseline on the scale of 500k images. Our method has higher learning ability on a dataset in which the training and testing sets are randomly drawn from the same distribution, whereas the baseline has lower accuracy on the dataset itself while having higher generalization ability on a different dataset. Our result also reflects important future issues in the community of person re-ID that the scalability/generalization ability is critical in a person re-ID system. Through this work, we call upon attention from the community for this problem. In the future, we will look into this interesting problem and design more robust descriptors for the large testing dataset.

4.3 Instance Retrieval

We apply the identification-verification model to the generic image retrieval task. Oxford5k [27] is a testing dataset containing buildings in the Oxford University. We train the network on another scene dataset proposed in Radenović et al. [29], which comprises some buildings without

Fig. 7. Pedestrian retrieval samples on the CUHK03 dataset [18] in the multishot setting. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right. The correct matches are in the green rectangles, and the false matching images are in the red rectangles.

Table 6. Impact of Data Size on the Market1501+500K Dataset

Method	Gallery Size	19,732	119,732	219,732	519,732
ResNet Baseline	Rank-1	73.69	72.15	71.55	70.67
	mAP	51.48	48.72	47.57	46.05
Ours (ResNet)	Rank-1	79.51	73.78	71.50	68.26
	mAP	59.87	52.28	49.11	45.24

Note: As the dataset gets larger, the accuracy drops.

overlapping with Oxford5k. Similarly, the model is trained to not only tell which building the image depicts but also determine whether the two input images are from the same architecture. The training data is high resolution. To obtain more information from the high-resolution building images, we modify the final pooling layer of our model to a MAC layer [35], which outputs the maximum value over the whole activation map. This layer helps us handle large images without

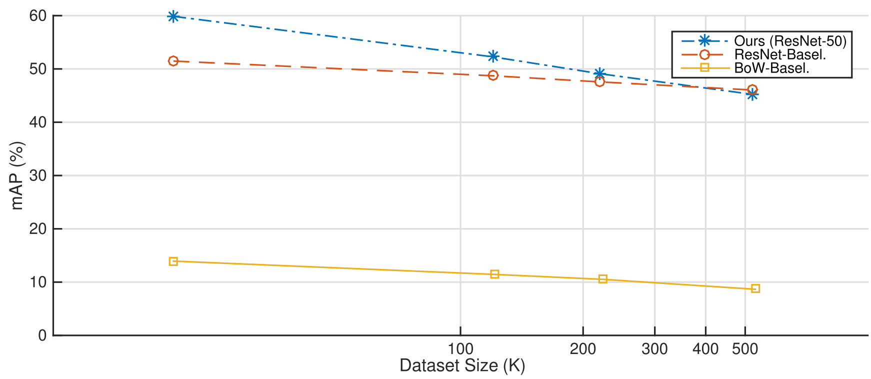

Fig. 8. Impact of data size on the Market1501+500 K dataset. As the dataset gets larger, the accuracy drops.

Table 7. Comparison of State-of-the-Art Results on the Oxford5k Dataset

Method	CaffeNet mAP	VGG16 mAP
mVoc/BoW [28]	48.8	—
CroW [14]	—	68.2
Neural Codes [2]	55.7	—
R-MAC [35]	56.1	66.9
R-MAC-Hard [29]	62.5	77.0
MAC-Hard [29]	62.2	**79.7**
Fine-Tuned Baseline	60.2	69.8
Ours	**66.2**	76.4

Note: The mAP is listed.

resizing them to a fixed size and output a fixed-dimension feature to retrieve the images. During training, the input image is randomly cropped to 320×320 from 362×362 and mirrored horizontally. During testing, we keep the original size of the images that are not cropped or resized and extract the feature.

In Table 7, many previous works are based on CaffeNet or VGG16. For a fair comparison, we report the baseline results and the results of our model based on these two network structures, respectively. The state-of-the-art method [29] uses an extra 3D model to conduct hard sampling, whereas our method does not. Our model based on VGG16 is not higher than that of Radenović et al. [29], but it is still competitive (76.4% vs. 79.7%). Our model, which uses CaffeNet as a pre-trained model, outperforms Radenović et al. [29]. We mainly seek to investigate the effect of the combination of the two losses. The primary concern of our work is to prove that the verif+identif model is superior to the fine-tuned baseline (76.4% vs. 69.8%). The proposed method shows a 6.0% and 6.6% improvement over the baselines based on CaffeNet and VGG16, respectively. We visualize some retrieval results in Figure 9. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right. The main difficulty in the image retrieval is the various object sizes in the image. In the first row, we use the roof (part of the building) to retrieve the images, and the top five images are correct candidate images. The other retrieval samples also show that our model is robust to the scale variations.

Fig. 9. Example retrieval results on the Oxford5k dataset [27] using the proposed embedding. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right. The query images are usually from the part of the architectures.

5 CONCLUSION

In this work, we propose a Siamese network that simultaneously considers identification loss and verification loss. The proposed model learns a discriminative embedding and a similarity measurement at the same time. It outperforms the state of the art on two popular person re-ID benchmarks and shows the potential ability to apply it on the generic instance retrieval task.

Future work includes exploring more novel applications of the proposed method, such as car recognition and fine-grain classification. In addition, we will investigate how to learn a robust descriptor to further improve the performance of the person re-ID on a large-scale testing set.

ACKNOWLEDGMENTS

We appreciate the support of Data to Decisions Cooperative Research Centre (http://www.d2dcrc.com.au), a Google Faculty Research Award, and NVIDIA Corporation for the donation of a TITAN X (Pascal) GPU.

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K. Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 3908–3916.
[2] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural codes for image retrieval. In *Proceedings of the European Conference on Computer Vision*. 584–599.
[3] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a Siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence* 7, 04, 669–688.

[4] Xiaojun Chang and Yi Yang. 2017. Semisupervised feature analysis by mining correlations among multiple tasks. *IEEE Transactions on Neural Networks and Learning Systems* 28, 10, 2294–2305.
[5] Xiaojun Chang, Yao-Liang Yu, Yi Yang, and Eric P. Xing. 2017. Semantic pooling for complex event analysis in untrimmed videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 8, 1617–1632.
[6] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. 2016. Similarity learning with spatial constraints for person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 1268–1277.
[7] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 1335–1344.
[8] Cheng Deng, Xu Tang, Junchi Yan, Wei Liu, and Xinbo Gao. 2016. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE Transactions on Multimedia* 18, 2, 208–218.
[9] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition* 48, 10, 2993–3003.
[10] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. 2016. Deep transfer learning for person re-identification. arXiv:1611.05244.
[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Vol. 2. IEEE, Los Alamitos, CA, 1735–1742.
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 770–778.
[13] Michael E. Houle, Xiguo Ma, Vincent Oria, and Jichao Sun. 2017. Query expansion for content-based similarity search using local and global features. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3, 25.
[14] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. 2016. Cross-dimensional weighting for aggregated deep convolutional features. In *Proceedings of the European Conference on Computer Vision*. 685–701.
[15] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 2288–2295.
[16] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 2288–2295.
[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems*. 1097–1105.
[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 152–159.
[19] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 2197–2206.
[20] Giuseppe Lisanti, Svebor Karaman, and Iacopo Masi. 2017. Multichannel-kernel canonical correlation analysis for cross-view person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 2, 13.
[21] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. 2016. End-to-end comparative attention networks for person re-identification. arXiv:1606.04404.
[22] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*. 507–516.
[23] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. 2016. Hierarchical Gaussian descriptor for person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 1363–1372.
[24] Alexis Mignon and Frédéric Jurie. 2012. PCCA: A new approach for distance learning from sparse pairwise constraints. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 2666–2672.
[25] Prabhu Natarajan, Pradeep K. Atrey, and Mohan Kankanhalli. 2015. Multi-camera coordination and control in surveillance systems: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11, 4, 57.
[26] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 4004–4012.
[27] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 1–8.

[28] Filip Radenović, Hervé Jégou, and Ondrej Chum. 2015. Multiple measurements and joint dimensionality reduction for large scale image search with short vectors. In *Proceedings of the International Conference on Multimedia Retrieval.* ACM, New York, NY, 587–590.

[29] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. arXiv:1604.02426.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3, 211–252.

[31] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

[32] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1, 1929–1958.

[33] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. In *Proceedings of the International Conference on Neural Information Processing Systems.* 1988–1996.

[34] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2015. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the Conference on Computer Vision and Pattern Recognition.* 2892–2900.

[35] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2015. Particular object retrieval with integral max-pooling of CNN activations. arXiv:1511.05879.

[36] Evgeniya Ustinova, Yaroslav Ganin, and Victor Lempitsky. 2015. Multiregion bilinear convolutional neural networks for person re-identification. arXiv:1512.05300.

[37] Laurens Van Der Maaten. 2014. Accelerating t-SNE using tree-based algorithms.*Journal of Machine Learning Research* 15, 1, 3221–3245.

[38] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. 2016. Gated Siamese convolutional neural network architecture for human re-identification. In *Proceedings of the European Conference on Computer Vision.* 791–808.

[39] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. 2016. A Siamese long short-term memory architecture for human re-identification. In *Proceedings of the European Conference on Computer Vision.* 135–153.

[40] A. Vedaldi and K. Lenc. 2015. MatConvNet—convolutional neural networks for MATLAB. In *Proceedings of the ACM International Conference on Multimedia.*

[41] Zheng Wang, Ruimin Hu, Chao Liang, Yi Yu, Junjun Jiang, Mang Ye, Jun Chen, and Qingming Leng. 2016. Zero-shot person re-identification via cross-view consistency. *IEEE Transactions on Multimedia* 18, 2, 260–272.

[42] Lin Wu, Chunhua Shen, and Anton van den Hengel. 2016. Deep linear discriminant analysis on Fisher networks: A hybrid architecture for person re-identification. arXiv:1606.01595.

[43] Lin Wu, Chunhua Shen, and Anton van den Hengel. 2016. PersonNet: Person re-identification with deep convolutional neural networks. arXiv:1601.07255.

[44] Yan Yan, Feiping Nie, Wen Li, Chenqiang Gao, Yi Yang, and Dong Xu. 2016. Image classification by cross-media active learning with privileged information. *IEEE Transactions on Multimedia* 18, 12, 2494–2502.

[45] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3, Article No. 27.

[46] Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang. 2009. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the ACM International Conference on Multimedia.* ACM, New York, NY, 175–184.

[47] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. Deep metric learning for person re-identification. In *Proceedings of the Conference on Pattern Recognition.* IEEE, Los Alamitos, CA, 34–39.

[48] Li Zhang, Tao Xiang, and Shaogang Gong. 2016. Learning a discriminative null space for person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition.* 1239–1248.

[49] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan. 2016. Sample-specific SVM learning for person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition.* 1278–1287.

[50] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Person re-identification by salience matching. In *Proceedings of the International Conference on Computer Vision.* 2528–2535.

[51] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. 2016. MARS: A video benchmark for large-scale person re-identification. In *Proceedings of the European Conference on Computer Vision.* 868–884.

[52] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the International Conference on Computer Vision.* 1116–1124.

[53] Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian. 2015. Fast image retrieval: Query pruning and early termination. *IEEE Transactions on Multimedia* 17, 5, 648–659.

[54] Liang Zheng, Shengjin Wang, and Qi Tian. 2014. Coupled binary embedding for large-scale image retrieval. *IEEE Transactions on Image Processing* 23, 8, 3368–3380.
[55] Liang Zheng, Shengjin Wang, Jingdong Wang, and Qi Tian. 2016. Accurate image search with multi-scale contextual evidences. *International Journal of Computer Vision* 120, 1, 1–13.
[56] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. 2016. Person re-identification: Past, present and future. arXiv:1610.02984.
[57] Liang Zheng, Yi Yang, and Qi Tian. 2017. SIFT meets CNN: A decade survey of instance retrieval. arXiv:1608.01807. DOI: http://dx.doi.org/10.1109/tpami.2017.2709749
[58] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2013. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 3, 653–668.
[59] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision.*

Received July 2017; revised October 2017; accepted October 2017