

# A Discriminatively Learned CNN Embedding for Person Re-identification

Zhedong Zheng, Liang Zheng and Yi Yang

**Abstract**—In this paper, we revisit two popular convolutional neural networks (CNN) in person re-identification (re-ID), *i.e.*, verification and identification models. The two models have their respective advantages and limitations due to different loss functions. In this paper, 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 making full usage of the re-ID annotations.

Our method can be easily applied on different pre-trained networks. Albeit simple, the learned embedding improves the state-of-the-art performance on two public person re-ID benchmarks. Further, we show our architecture can also be applied in image retrieval.

**Index Terms**—Large-scale Person Re-identification, Convolutional Neural Networks.

## I. INTRODUCTION

Person re-identification (re-ID) is usually viewed as an image retrieval problem, which matches pedestrians from different cameras [1]. 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. The datasets contain the general visual variance of pedestrian and provide a comprehensive evaluation [2], [3]. 2) the learned embedding of pedestrian using a convolutional neural network (CNN).

Recently, the convolutional neural network (CNN) has shown potential for learning state-of-the-art feature embeddings or deep metrics [2], [4], [5], [6], [7], [8], [9]. As shown in Fig. 1, there are two major types of CNN structures, *i.e.*, verification models and identification models. The two models are different in terms of 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.

Verification models take a pair of images as input and determine whether they belong to the same person or not. A number of previous works treat person re-ID as a binary-class classification task or a similarity regression task [2], [4], [5], [6]. Given a label  $s \in \{0, 1\}$ , the verification network forces two images of the same person to be mapped to nearby

Zhedong Zheng, Liang Zheng and Yi Yang are with Faculty of Engineering and IT, University of Technology Sydney, NSW, Australia. E-mail: zdzheng12@gmail.com, liangzheng06@gmail.com, yee.i.yang@gmail.com

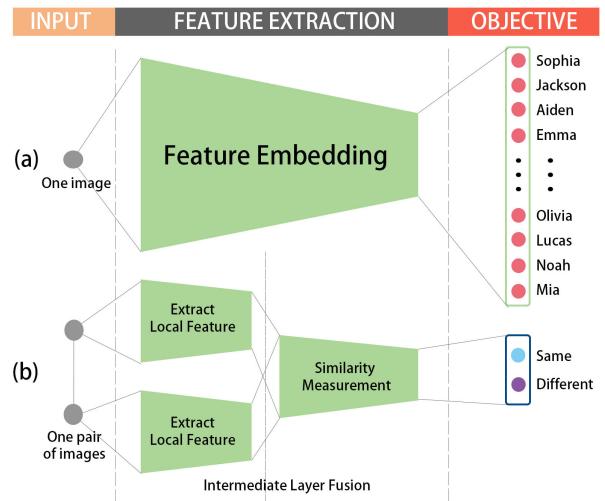


Fig. 1. The difference between the verification and identification models. Green blocks represent non-linear functions by CNN. a) Identification models treat person re-ID as a multi-class recognition task, which take one image as input and predict its identity. b) Verification models treat person re-ID as a two-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 two-class recognition case.

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 [1], and do not take all 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 which treat person re-identification as a multi-class recognition task, are employed for feature learning [1], [7], [8], [9]. They directly learn the non-linear functions from an input image to the person ID and the cross-entropy loss is used following the final layer. During testing, the feature is extracted from a fully connected 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, *i.e.*, it does not account for the similarity measurement between image pairs, which can be problematic during the pedestrian retrieval process.

The above-mentioned observations demonstrate that the two types of models have complementary advantages and

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

TABLE I

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.

limitations as shown in Table I. 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 advantages of the annotated data in terms of pair-wise similarity and image identities. During testing, the final convolutional activations are extracted for Euclidean distance based pedestrian retrieval. To summarize, our contributions are:

- 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 [3] and CUHK03 [2]) and one instance retrieval dataset (Oxford5k [10]).

The paper is organized as follows. We first review some related works in Section II. In Section III, we describe how we combine the two losses and define the CNN structure. The implementation details are provided. In Section IV, we present the experimental results on two large-scale person re-identification datasets and one instance retrieval dataset. We conclude this paper in Section V.

## II. RELATED WORK

In this section we describe previous works relevant to the approach discussed in this paper. They are mainly based on verification models or identification models.

### A. Verification Models

In 1993, Bromley *et al.* [11] 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-identification with a focus on data augmentation and image matching. Yi *et al.* [4] 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.* split the convolutional map into four parts and

fuse the part features with the global features [12]. Li *et al.* [2] 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 regression as binary-class penalized by softmax loss. Later, Ahmed *et al.* [13] improve the verification model by adding a different matching layer that compares the activation of two images in neighboring pixels. Besides, Wu *et al.* [5] use smaller filters and a deeper network to extract features. Varior *et al.* [6] combine CNN with some gate functions, similar to long-short-term memory (LSTM [14]) 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.* [15] use triplet samples for training the network which considers the images from the same people and the different people at the same time.

### B. Identification Models

Recent datasets such as CUHK03 [2] and Market1501 [3] provide large-scale training sets, which make it possible to train a deeper classification model without over-fitting. Every identity has 9.6 training images on average in CUHK03 [2] and has 17.2 images in Market1501 [3]. CNN can learn discriminative embeddings by itself without part-matching. Zheng *et al.* [1], [8], [16] directly use a conventional fine-tuning approach on Market1501 [3], PRW [8] and MARS [16] and outperform many recent results. Wu *et al.* [17] combine CNN embeddings with the hand-crafted features in the FC layer. Besides, Xiao *et al.* [7] jointly train a classification model using multiple datasets and propose a new dropout function to deal with the hundreds of classes. In [9], Xiao *et al.* train a classification model similar to the faster-RCNN [18] method and automatically predict the location of the candidate pedestrian from the whole image, which alleviates the pedestrian detection errors.

### C. Verification-identification Models

In face recognition, the “DeepID networks” train the network with the verification and identification losses [19], [20], [21], which is similar to our network. In [19], Sun *et al.* jointly train face identification and verification. Then more verification supervision is added into the model [20] and a deeper network is used [21].

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 [19] while the current largest person re-id training dataset contains 12,936 images of 751 identities [3]. DeepID networks apply contrastive loss to the verification problem, while our model uses the cross-entropy loss. We find that the contrastive loss leads to over-fitting when the number of images is limited. In the experiment, we show the proposed method learns more robust person representative and outperforms using contrastive loss. Second, dropout [22] cannot be applied on the embedding before the contrastive loss, which introduces zero values at random locations. On the contrary, we can add dropout regularization on the embedding

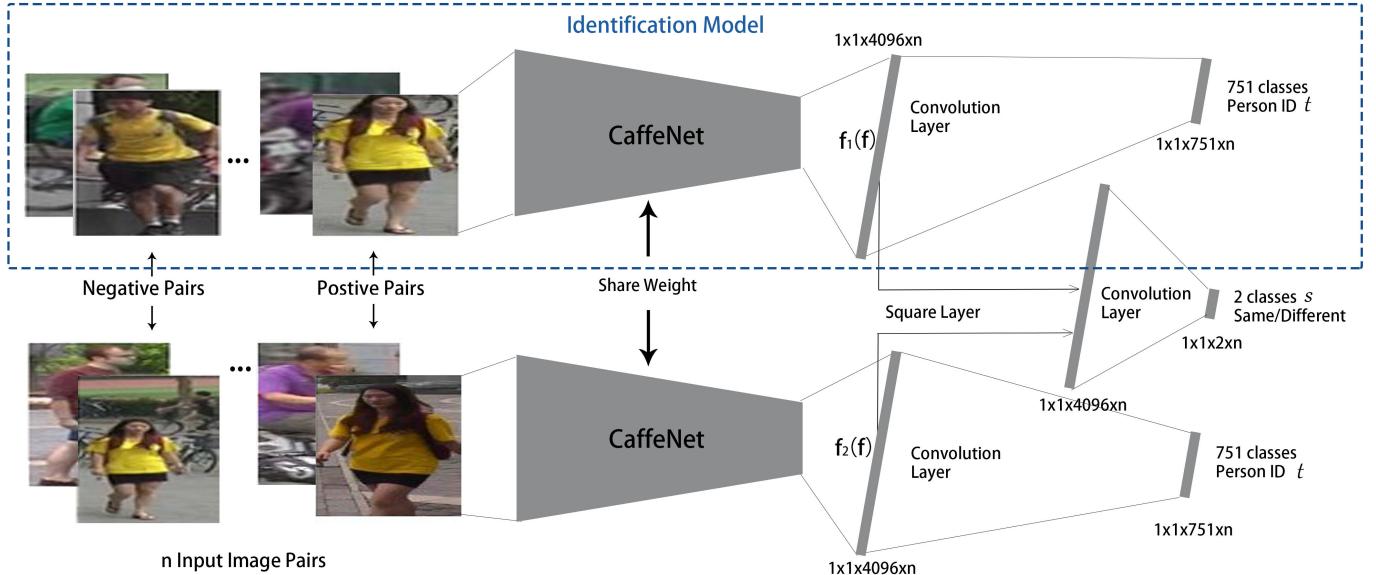


Fig. 3. The proposed model structure. Given  $n$  pairs of images of size  $227 \times 227$ , two identical CaffeNet models are used as the non-linear embedding functions and output 4,096-dim embeddings  $f_1, f_2$ . Then,  $f_1, f_2$  are used to predict the identity  $t$  of the two input images, respectively,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

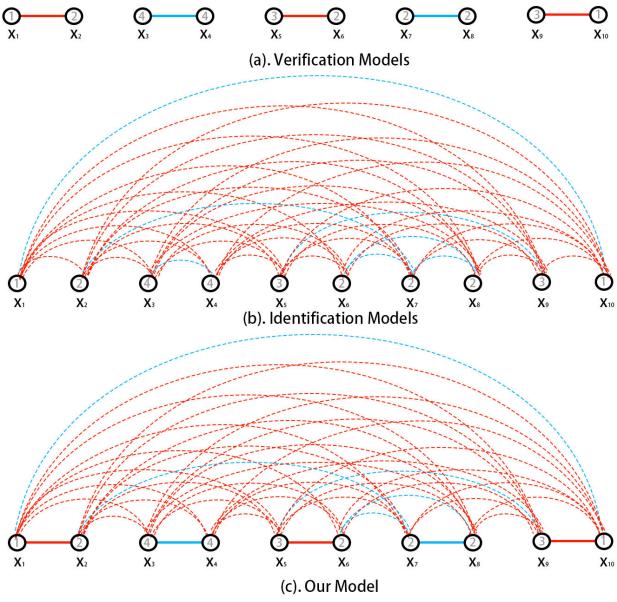


Fig. 2. Illustration for a training batch. The number in the circle is the identity label. Blue and red edges represent whether the image pair depicts the same identity or not. Dotted edges represent implicit relationships and solid edges represent explicit relationships. Our model combine the strengths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

in the proposed model. Third, the DeepID networks are trained from scratch, while our model benefits from the networks pretrained on ImageNet [23]. Finally, we evaluate our method on the tasks of person re-ID and instance retrieval, providing more insights in the verification-classification models.

### III. PROPOSED METHOD

#### A. Preview

Fig. 2 (a) and Fig. 2 (b) illustrate the relational graph built by verification and identification models. In a sample batch of size  $m = 10$ , red edges represent the positive pairs (the same person) and blue 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 pair-wise comparison, such as part matching [2], [13] or contrastive loss [24]. For example, contrastive loss directly calculates the Euclidean distance between two embeddings. In identification models, the input is independent to each other. But there is implicit relationship between the learned embeddings built by the cross-entropy loss. 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 [25], similar observation and visualization are shown. So 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. On the other hand, classification models do not explicitly consider similarity measurements. Fig. 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 I).

### B. Overall Network

Our network is basically a convolutional siamese network that combines the verification and identification losses. Fig. 3 briefly illustrates the architecture of the proposed network. Given an input pair of images resized to  $227 \times 227$ , the proposed network simultaneously predicts the IDs of the two images and the similarity score. The network consists of two ImageNet [23] pre-trained 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 pre-trained CNN model can be CaffeNet [26], VGG16 [27] or ResNet-50 [28], from which we have removed the final fully-connected (FC) layer. The re-ID performance of the three models is comprehensively evaluated in Section IV. Here, we do not provide detailed descriptions of the architecture of the CNN models and only take CaffeNet as an example in the following subsections. 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.

### C. Identification Loss

There are two CaffeNets in our architecture. They share weights and predict the two identity labels of the input image pair simultaneously. In order to fine-tune the network on a new dataset, we replace the final fully-connected layer (1,000-dim) of the pre-trained CNN model with a convolutional layer. The number of the training identities in Market-1501 is 751. So this convolutional layer has 751 kernels of size  $1 \times 1 \times 4096$  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 multi-class recognition approaches, we use the cross-entropy loss for identity prediction, which is

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

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

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

### D. Verification Loss

While some previous works contain a matching function in the intermediate layers [2], [6], [13], 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 [8], [16] and it is more compact than the activations in the intermediate layers. So in our model,

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

TABLE II  
RESULTS ON MARKET1501 [3] BY IDENTIFICATION LOSS AND VERIFICATION LOSS INDIVIDUALLY AND JOINTLY. “I” AND “V” DENOTE THE IDENTIFICATION LOSS AND VERIFICATION LOSS, RESPECTIVELY.

the pedestrian descriptor  $f_1, f_2$  in the identification model are directly supervised by the verification loss. As shown in Fig. 3, we introduce a non-parametric layer called 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  $(\hat{q}_1, \hat{q}_2)$  which represents the predicted probability of the two input images belonging to the same identity.  $\hat{q}_1 + \hat{q}_2 = 1$ . The convolutional layer takes  $f_s$  as input and filters it with 2 kernels of size  $1 \times 1 \times 4096$ . 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

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

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

Here  $f_1, f_2$  are the two tensors of size  $1 \times 1 \times 4096$ .  $s$  is the target class (same/different),  $\theta_S$  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 [19], we do not use the contrastive loss [24]. 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 over-fitting because the number of training of each identity is limited in person re-ID. On the other hand, dropout [22], which introduces zero values at random locations, can not 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 IV, 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 mAP respectively.

### E. Identification vs. 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

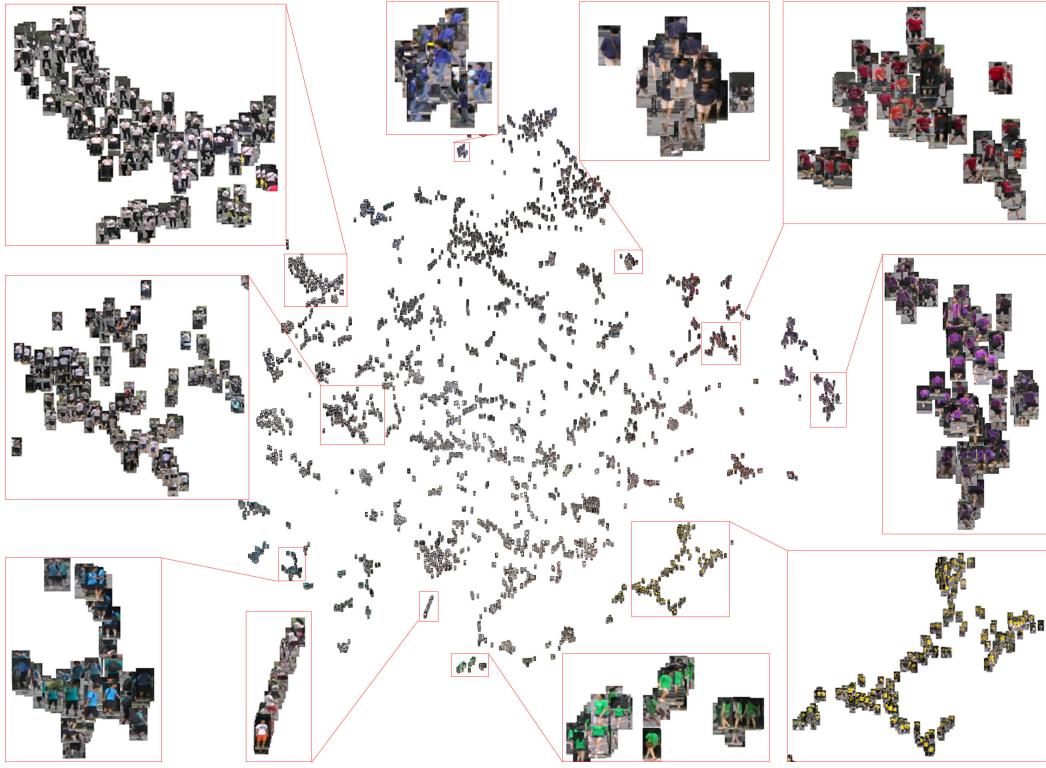


Fig. 4. Barnes-Hut t-SNE visualization [29] of our embedding on a test split (354 identity, 6868 images) of Market1501. Best viewed when zoomed in. We find the color is the major clue for the person re-identification and our learned embedding is robust to some viewpoint variations.

model separately. Following the learning rate setting in Section III-F, 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 shown in Table II, 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 [28] as the pretrained model and try to find the differences between identification loss and verification loss. We select three test images in the Market1501. One image is considered to be well detected and 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  $14 \times 14$ . We visualize the sum of several activation maps. As shown in Fig. 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 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 Fig. 4 we visualize the embedding by plot them to the 2-dimension map. In regard to Fig. 5, we find the network usually has strong attention on the center part of the human (usually clothes) and it also illustrates the color of the clothes is the major clue for the person re-identification.

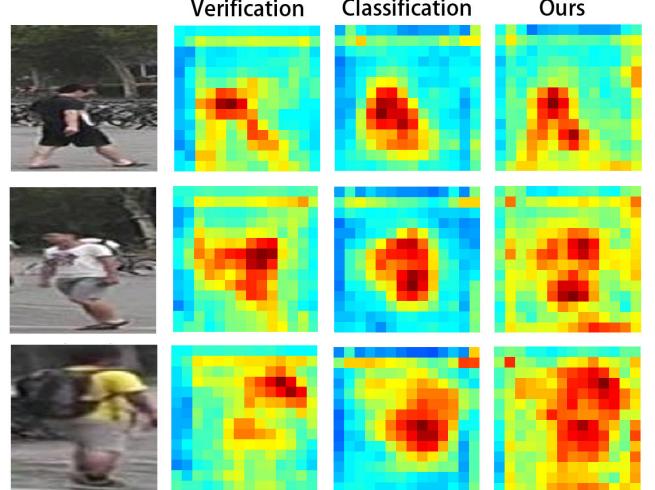


Fig. 5. Visualization of the activation maps in the ResNet-50 [28] model trained by the two losses. The identification and the verification networks exhibit different activation patterns to the pedestrian. The proposed model takes advantages of both networks and the new activation map is almost a union of the two individual maps. Our model activates more neurons.

#### F. Training and Optimization

**Input preparation.** We resize all the training images to  $256 \times 256$ . The mean image computed from all the training images is subtracted from all the images. During training, all the images are randomly cropped to  $227 \times 227$  for CaffeNet [26]

and mirrored horizontally. For ResNet-50 [28] and VGG16 [27], we randomly crop images to  $224 \times 224$ . 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 multiple 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 over-fitting.

**Training.** We use the Matconvnet [30] package for training and testing the embedding with CaffeNet [26], VGG16 [27] and ResNet-50 [28], 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, 48 for VGG16 and ResNet-50. The learning rate is initialized as 0.001 and then set to 0.0001 for the final 5 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 the gradients produced by every objectives 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 [22] before the final convolutional layer.

**Testing.** We adopt an efficient method to extract features as well as the activation in the intermediate layer. Because two CaffeNet share weights, our model has nearly the same memory consumption with the pretrained model. So we extract features by only activating one fine-tuned model. Given a  $227 \times 227$  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 the gallery features to obtain the final ranking result. Note that the cosine distance is equivalent to Euclidean distance when the feature is L2-normalized.

#### IV. EXPERIMENTS

We mainly verify the proposed model on two large-scale datasets Market1501 [3] and CUHK03 [2]. We report the results trained by three network structures. Besides, we also report the result on Market1501+500k dataset [3]. 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, *i.e.*, Oxford Buildings [10]. The performance is comparable to the state of the art.

##### A. Dataset

**Market1501** [3] 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 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 bboxes, which

| Method                | Single Query<br>rank-1 | Query<br>mAP | Multi. Query<br>rank-1 | Query<br>mAP |
|-----------------------|------------------------|--------------|------------------------|--------------|
| BoW + KISSME [3]      | 44.42                  | 20.76        | -                      | -            |
| SL [31]               | 51.90                  | 26.35        | -                      | -            |
| Multiregion CNN [32]  | 45.58                  | 26.11        | 56.59                  | 32.26        |
| DADM [33]             | 39.4                   | 19.6         | 49.0                   | 25.8         |
| CAN [34]              | 48.24                  | 24.43        | -                      | -            |
| DNS [35]              | 55.43                  | 29.87        | 71.56                  | 46.03        |
| Fisher Network [36]   | 48.15                  | 29.94        | -                      | -            |
| S-LSTM [37]           | -                      | -            | 61.6                   | 35.3         |
| Gate Reid [6]         | 65.88                  | 39.55        | 76.04                  | 48.45        |
| CaffeNet-Basel. [26]  | 50.89                  | 26.79        | 59.80                  | 36.50        |
| Ours(CaffeNet)        | 62.14                  | 39.61        | 72.21                  | 49.62        |
| VGG16-Basel. [27]     | 65.02                  | 38.27        | 74.14                  | 52.25        |
| Ours(VGG16)           | 70.16                  | 47.45        | 77.94                  | 57.66        |
| ResNet-50-Basel. [28] | 73.69                  | 51.48        | 81.47                  | 63.95        |
| Ours(ResNet-50)       | <b>79.51</b>           | <b>59.87</b> | <b>85.84</b>           | <b>70.33</b> |

TABLE III  
COMPARISON WITH THE STATE-OF-THE-ART RESULTS ON THE MARKET1501 DATASET. 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.

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-identification. In the realistic setting, the scale of the gallery is usually large. The distractor dataset of Market1501 provides extra 500,000 bboxes, consisting of false alarms on the background as well as the persons not belonging to any of the original 1,501 identities [3]. When testing, we add the 500k images to the original gallery, which makes the retrieval more difficult.

**CUHK03** dataset [2] 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 in 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** buildings [10] consists of 5062 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  $851 \times 921$ .

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

##### B. Person Re-id Evaluation

**Comparison with the CNN baseline.** We train the baseline networks according the conventional fine-tuning method [1], [8]. The baseline networks are pretrained on ImageNet [23]

| Method           | rank-1      | rank-5      | rank-10     | mAP         |
|------------------|-------------|-------------|-------------|-------------|
| KISSME [38]      | 11.7        | 33.3        | 48.0        | -           |
| DeepReID [2]     | 19.9        | 49.3        | 64.7        | -           |
| Bow+HS [3]       | 24.3        | -           | -           | -           |
| LOMO+XQDA [39]   | 46.3        | 78.9        | 88.6        | -           |
| SI-CI [40]       | 52.2        | 84.3        | 94.8        | -           |
| DNS [35]         | 54.7        | 80.1        | 88.3        | -           |
| CaffeNet-Basel.  | 35.8        | 65.3        | 77.96       | 42.6        |
| Ours (CaffeNet)  | 59.8        | 88.3        | 94.2        | 65.8        |
| VGG16-Basel.     | 49.1        | 78.4        | 87.2        | 55.7        |
| Ours (VGG16)     | 71.8        | 93.0        | 97.1        | 76.5        |
| ResNet-50-Basel. | 71.5        | 91.5        | 95.9        | 75.8        |
| Ours (ResNet-50) | <b>83.4</b> | <b>97.1</b> | <b>98.7</b> | <b>86.4</b> |

TABLE IV

COMPARISON WITH THE STATE-OF-THE-ART RESULTS REPORTED ON THE CUHK03 DATASET USING THE SINGLE-SHOT SETTING. THE MAP AND RANK-1 ACCURACY ARE LISTED.

and fine-tuned to predict the person identities. As shown in Tab. III, we obtain 50.89%, 65.02% and 73.69% rank-1 accuracy by CaffeNet [26], VGG16 [27] and ResNet-50 [28], 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 [26], VGG16 [27] and ResNet-50 [28] on Market1501. Similarly, we observe 35.8%, 49.1% and 71.5% baseline rank-1 accuracy on CUHK03 in single-shot setting. As shown in Tab. IV, these baseline 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 to learn more discriminative features.

**Cross-entropy 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 over-fit on the re-ID dataset because no regularization is added to the verification. Cross-entropy loss designed in our model can work with the dropout function and avoid the over-fitting.

**Comparison with the state of the art.** As shown in Table III, we compare our method with other state-of-the-art algorithms in terms of mean average precision (mAP) and rank-1 accuracy on Market1501. We report the single-query as well as 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 [6]. 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 Tab. IV. 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 gallery. The proposed model yields 83.4% rank-1 and 86.4% mAP and

| Method           | rank-1      | rank-5      | rank-10     | mAP         |
|------------------|-------------|-------------|-------------|-------------|
| S-LSTM [37]      | 57.3        | 80.1        | 88.3        | 46.3        |
| Gate-SCNN [6]    | 68.1        | 88.1        | 94.6        | 58.8        |
| CaffeNet-Basel.  | 43.3        | 63.5        | 76.8        | 37.2        |
| Ours (CaffeNet)  | 67.2        | 86.2        | 92.3        | 61.5        |
| VGG16-Basel.     | 58.8        | 80.2        | 87.3        | 51.0        |
| Ours (VGG16)     | 78.8        | 91.8        | 95.4        | 73.9        |
| ResNet-50-Basel. | 77.1        | 89.6        | 93.9        | 73.1        |
| Ours(ResNet-50)  | <b>88.3</b> | <b>95.7</b> | <b>97.8</b> | <b>85.0</b> |

TABLE V  
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE CUHK03 DATASET UNDER THE MULTI-SHOT SETTING. THE MULTI-SHOT SETTING USES THE ALL IMAGES IN THE OTHER CAMERA AS GALLERY. THE MAP AND RANK-1 ACCURACY ARE LISTED.

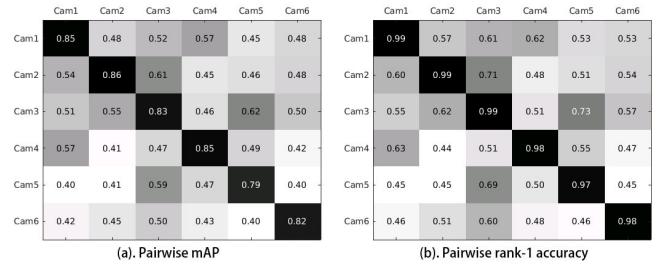


Fig. 6. Re-identification performance between camera pairs on Market1501: (a) mAP and (b) rank-1 accuracy. Cameras on the vertical and horizontal axis correspond to the probe and gallery, respectively. The cross-camera average mAP and average rank-1 accuracy are 48.42% and 54.42%, respectively.

outperforms the state-of-the-art performance.

As shown in Tab. V, we also report the results in the multi-shot setting, which uses all the images from the other camera as gallery and the number of the gallery images is about 500. We think this setting is much closer to image retrieval and alleviate the unstable effect caused by the random searching pool under single-shot settings. Fig. 7 presents some re-ID samples on 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 on the third row, we find it is a reasonable prediction since the man with 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 multi-shot setting.

**Results between camera pairs.** CUHK03 [2] only contains two camera views. So this experiment is evaluated on Market1501 [3] since it contains six different cameras. We provide the re-identification results between all camera pairs in Fig. 6. Although camera-6 is a  $720 \times 576$  low-resolution camera and captures distinct background with the other HD cameras, the re-ID accuracy between camera 6 and the others is relatively high. We also compute the cross-camera average mAP and average rank-1 accuracy: 48.42% and 54.42% respectively. Comparing to the previous reported results, *i.e.*, 10.51% and 13.72% in [3], our method largely improves the performance and observes a smaller standard deviation between cameras. It suggests that the discriminatively learned embedding works under different viewpoints.

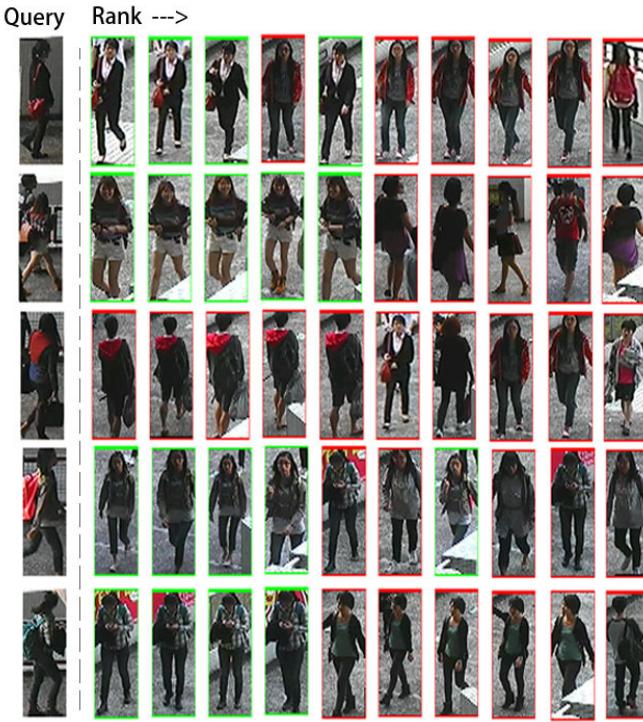


Fig. 7. Pedestrian retrieval samples on CUHK03 dataset [2] in the multi-shot setting. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right.

Further, Fig. 4 shows the Barnes-Hut t-SNE visualization [29] on the learned embeddings of our model. By the clustering algorithm, the persons wearing the similar-color clothes are quite clustered together and are apart from other persons. The learned pedestrian descriptor pay more attention to the color and it is robust to some illusion and viewpoint variations. In realistic setting, we think 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 Tab. VI. As the searching pool gets larger, the accuracy drops. With the gallery size of  $500,000 + 19,732$ , we still achieve 68.26% rank1 accuracy and 45.24% mAP. A relative drop 24.4% from 59.87% to 45.24% on mAP is observed, compared to a relative drop 37.88% from 13.94% to 8.66% in our previous work [3]. Besides, we also compare our result with the performance of the ResNet Baseline. As shown in Fig. 8, it is interesting that the re-ID precision of our model decreases more quickly comparing 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. In fact, the 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 500 k images. In the future, we will look into this interesting problem and design more robust

| Method        | Gallery size | 19,732 | 119,732 | 219,732 | 519,732 |
|---------------|--------------|--------|---------|---------|---------|
| ResNet Basel. | 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   |

TABLE VI  
IMPACT OF DATA SIZE ON MARKET1501+500K DATASET. AS THE DATASET GETS LARGER, THE ACCURACY DROPS.

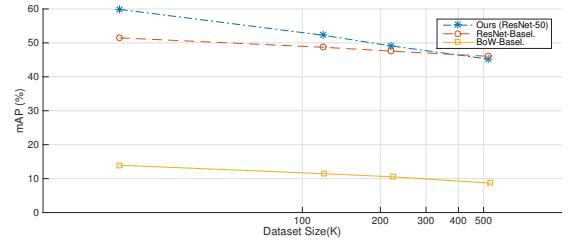


Fig. 8. Impact of data size on Market1501+500K dataset. As the dataset gets larger, the accuracy drops.

descriptors for the transfer dataset.

### C. Instance Retrieval

We apply the identification-verification model to the generic image retrieval task. Oxford5k [10] is a testing dataset containing buildings in the Oxford University. We train the network on another scene dataset proposed in [41], which comprises of a number of buildings without overlapping with the 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. In order to obtain more information from the high-resolution building images, we modify the final pooling layer of our model to a MAC layer [42], which outputs the maximum value over the whole activation map. This layer helps us to handle large images without 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  $320 \times 320$  from  $362 \times 362$  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 VII, many previous works are based on CaffeNet or VGG16. For fair comparison, we report the baseline results and the results of our model based on these two network structures, respectively. Our model which uses CaffeNet as pretrained model outperforms the state of the art. Meanwhile, the model using VGG16 is comparable to the state-of-the-arts methods. The proposed method show a 6.0% and 6.6% improvement over the baseline networks CaffeNet and VGG16, respectively. We visualize some retrieval results in Fig. 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 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 our model is robust to the scale variations.

| Method             | CaffeNet mAP | VGG16 mAP   |
|--------------------|--------------|-------------|
| mVoc/BoW [43]      | 48.8         | -           |
| CroW [44]          | -            | 68.2        |
| Neural codes [45]  | 55.7         | -           |
| R-MAC [42]         | 56.1         | 66.9        |
| R-MAC-Hard [41]    | 62.5         | 77.0        |
| MAC-Hard(V) [41]   | 62.2         | <b>79.7</b> |
| Finetuned-Baseline | 60.2         | 69.8        |
| Ours               | <b>66.2</b>  | 76.4        |

TABLE VII

COMPARISON OF STATE-OF-THE-ART RESULTS ON THE OXFORD5K DATASET. THE MAP IS LISTED. RESULTS REPORTED WITH THE USE OF ALEXNET [26] OR VGGNET [27] ARE MARKED BY (A) OR (V) RESPECTIVELY.

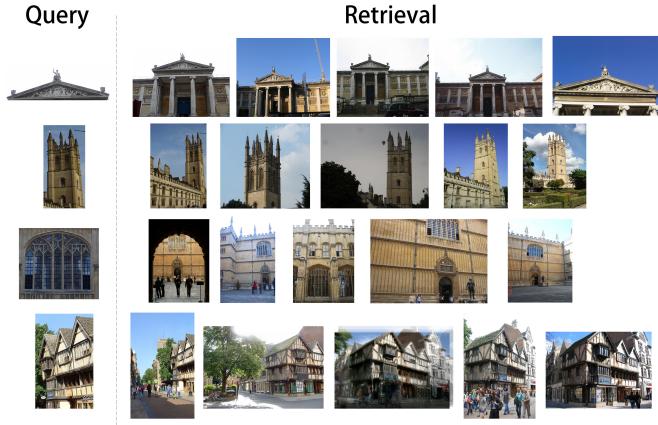


Fig. 9. Example retrieval results on Oxford5k dataset [10] using the proposed embedding. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right. The query images are usually from the part of the architectures.

## V. CONCLUSION

In this work, we propose a siamese network that simultaneously considers the identification loss and the 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 potential ability to apply on the generic instance retrieval task.

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

## REFERENCES

- [1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.
- [2] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.
- [3] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.
- [4] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 34–39.
- [5] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.
- [6] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 791–808.
- [7] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1249–1258.
- [8] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.
- [9] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," *arXiv preprint arXiv:1604.01850*, 2016.
- [10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.
- [11] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.
- [12] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.
- [13] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.
- [14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [15] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.
- [16] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 868–884.
- [17] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8.
- [18] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.
- [19] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.
- [20] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900.
- [21] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.
- [22] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [24] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 1735–1742.
- [25] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 507–516.
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

- [29] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *Journal of machine learning research*, vol. 15, no. 1, pp. 3221–3245, 2014.
- [30] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," in *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.
- [31] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1268–1277.
- [32] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multiregion bilinear convolutional neural networks for person re-identification," *arXiv preprint arXiv:1512.05300*, 2015.
- [33] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," *arXiv preprint arXiv:1605.03259*, 2016.
- [34] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *arXiv preprint arXiv:1606.04404*, 2016.
- [35] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," *arXiv preprint arXiv:1603.02139*, 2016.
- [36] L. Wu, C. Shen, and A. v. d. Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *arXiv preprint arXiv:1606.01595*, 2016.
- [37] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 135–153.
- [38] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2288–2295.
- [39] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.
- [40] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1288–1296.
- [41] F. Radenović, G. Tolias, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," *arXiv preprint arXiv:1604.02426*, 2016.
- [42] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.
- [43] F. Radenović, H. Jégou, and O. Chum, "Multiple measurements and joint dimensionality reduction for large scale image search with short vectors," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 587–590.
- [44] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *European Conference on Computer Vision*. Springer, 2016, pp. 685–701.
- [45] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European Conference on Computer Vision*. Springer, 2014, pp. 584–599.