Discriminative Transfer Learning Siamese CNN for Person Re-Identification

Yuan Tian, Cairong Zhao*, Kang Chen
Department of Computer Science and Technology
Tongji University
Shanghai, China
zhaocairong@tongji.edu.cn

Yipeng Chen, Zhihua Wei, Duoqian Miao
Key Laboratory of Embedded System and Service Computing, Tongji University
Shanghai, China
dqmiao@tongji.edu.cn

***Abstract*—Person re-identification (Re-ID) has become an increasingly popular computer vision problem. It remains challenging, especially when there are non-overlapping cameras. In this paper, we review the two representative architecture, i.e., identification and verification models. They both have their advancements and limitations. We present a novel method to address the Re-ID problem. First, combine the two models to consist a more effective fusion loss function. Second, we find that CNNs which are pre-trained on large image datasets learn more discriminative knowledge with objective semantic, which can be transferred to subsequent layers to promote accuracy significantly. Experiments on four benchmark datasets show the superiority of our method over the state-of-the-art alternatives.**

Keywords-person re-identification; convolutional neural network; knowledge transfer

I. INTRODUCTION

Person re-identification (Re-ID), which is usually viewed as a challenging problem of matching pedestrian across non-overlapping cameras, has attracted significant attention in the computer vision community recently. Especially with the prosperity of convolutional neural network(CNN) various deep learning frameworks, many new approaches have been proposed to address this task. Person Re-ID has similarities with image retrieval task in many aspects. Given a pedestrian who has been captured by one surveillance camera(query), Re-ID determines whether the identical individual has been observed by another camera. Earlier works focus on learning an effective metric to measure the similarity between images pairs or designing view-insensitive feature descriptors. Recently, due to the availability of large scale person Re-ID datasets and wide spread using of deep learning to obtain discriminative feature. The Re-ID model has made much progress over traditional methods with large margin.

Recently, the CNN has shown potential for learning state-of-the-art feature embed-dings or deep metrics [1], [2], [4], [5], [6], [7], [9]. 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 feature.

The verification model set a pair of image as input and measure the similarity between the pair and determine whether they belong to same identity or not. Several previous works on verification model take the Re-ID as a binary classification task [2], [4], [5]. Given a binary label $p \in \{0,1\}$ indicates the similarity scores quite close so that they are inferred as the same, otherwise the scores vary by large margin for the negative verification. However, there is a problem with verification model which only use weak binary labels [1]. These models do not take full use of the annotated information. Namely they only consider the relationship between the input pair but almost ignore their potential even essential relationship with other image pairs in the dataset.

Identification models view the Re-ID problem as a multi-class recognition task. Contrast with the Verification models, identification models attempt to utilize full information of the re-ID labels, and are employed for learning a more discriminative feature learning [1], [7], [9]. They directly learn the non-linear transformation from an input image to the person ID. Cross-entropy loss is universally used in this kind of task to supervise the training procedure. The learned deep features are employed normalization operation to counter overfitting problem and during the testing procedure, the squared Euclidean distance of the normalized feature vectors are used for measuring the similarity. A theoretical problem with identification is that the training operation is different from the testing procedure. Therefore, the model itself does not provide a reasonable explanation for the effectiveness.

On the other hand, the insufficient label and data may lead to poor performance and even make the problem non-convergent. Then the transfer learning is necessary. Traditional machine learning methods usually has a default underlying assumption that the training and test samples are captured in similar scenarios so that their distributions are assumed to be the same. This assumption doesn't hold in many real visual recognition applications, especially when samples are captured across different cameras in Re-ID problem. By transfer discriminative knowledge from source domain to target domain, the model will achieve consider able performance with few label. The unsupervised transfer learning even can accomplish the multiclass recognition tasks with unlabeled raw data.

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

2327-0985/17 $31.00 © 2017 IEEE
DOI 10.1109/ACPR.2017.119

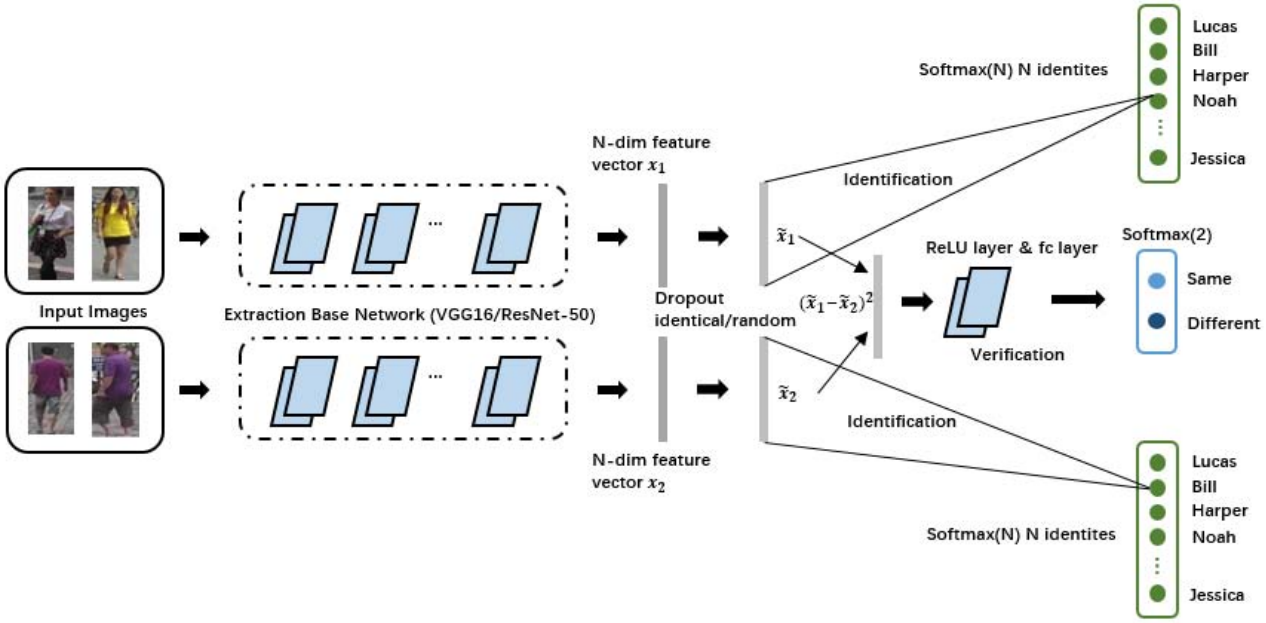

Figure 1. The overall architecture of the proposed discriminative knowledge transferring CNN model

(1) We propose a novel deep Re-ID architecture combines identification and verification loss function.

(2) We use a stepwise train strategy to extract general knowledge from ImageNet[31] and transfer to subsequent layers, where the general knowledge evolves into specialized knowledge with the rise of hierarchy.

(3) Experiments on CUHK01, CUHK03, VIPeR, Market-1501 datasets show the effectiveness and superiority of our model and improve the performance significantly by large margin.

II. PROPOSED METHOD

A. Overall Archeitecture

The overall network architecture of the proposed deep Re-ID model is briefly illustrated in Fig.1. It is basically a convolutional siamese network combines verification and identification losses. It aims to learn a representative local feature of each image of the input pair and then compute the similarity of the pairs or learn a distinguishable feature to classify the image into a specific individual class. The network has two losses: identification losses and a verification loss. The images are inputted into the basenetwork for feature extraction. The last layer output a N-dimension feature descriptor of the image. Then the dropout operation is deployed on both feature. It wa worth to mention that the two dropout are identical. For the identification procedure the feature is used to classify the image into the identity class by softmax classifier respectively. As for the verification procedure, we compute the squared Euclidean distance between images. This operation is elementwise so that the outcome will still be a tensor. Subsequently, the tensor is fed into a rectified linear unit (ReLU) layer and followed by a fully connected (fc) layer which map the feature space to label space and generate the last two-node (same/different) softmax layer to verify whether the image pair belongs to same individual or not. Before further operation, dropout is deployed on the feature vectors. It will randomly freeze part of the elements of x_i to zero. Formally expressed as:

$$\tilde{x}_i = x_i * m^d \tag{1}$$

Here, m^d is the dropout mask and * denotes the elementwise product. Each element of m^d is a random variable sampled following a Bernoulli process, *i.e.*, the *i*th element $m_i^d \backsim Bernoulli\ (p)$ and has a probability of p to be 1.

B. Verification Loss

Our model directly compares the feature after the dropout operation. While some previous architectures adopt intermediate matching measurement. $\tilde{x}_1,\tilde{x}_2$ represents the feature vector after dropout. We compute the squared Euclidean distance between them $(\tilde{x}_1-\tilde{x}_2)^2$,this operation is elementwise and we got a novel fused feature vector for verification. Most previous deep Re-ID work view verification process as a binary classification problem, here we follow this idea and ensemble a softmax classifier which has two nodes represent same/different respectively. We use the cross-entropy loss function which is widely adopted in multiclass image recognition tasks:

$$p^s = softmax_{verification}\left((\tilde{x}_1-\tilde{x}_2)^2 \circ \theta_{verif}\right) \tag{2}$$

$$L_{verification}\left(\theta_{verif}, s\right) = \sum_{i=1}^{2} p_i^s log\left(\frac{1}{p_i}\right) \tag{3}$$

Here, s is the target class (same/different), $\circ$ denotes a convolutional computation, p_i^s is the similarity score, and the transformation is parametrized by θ_{verif}. If the image pair is predicted to be the same person. Then $p_1 = 1, p_2 = 0$ and $p_1 = 0, p_2 = 1$ otherwise.

C. Identification Loss

The two basenetwork for feature extraction are siamese.

TABLE I. THE CMC PERFORMANCE OF THE PROPOSED METHOD IN CONTRAST WITH STATE-OF-THE-ART METHODS ON THREE DATASETS

Method	VIPeR			CUHK01(p=100)			CUHK03		
	r=1	*r=5*	*r=10*	*r=1*	*r=5*	*r=10*	*r=1*	*r=5*	*r=10*
ITML[19]	-	-	-	17.10	42.31	55.07	5.53	18.89	29.96
eSDC[21]	26.31	46.61	58.86	22.84	43.89	57.67	8.76	24.07	38.28
KISSME[20]	19.60	48.00	62.20	29.40	57.67	62.43	14.17	48.54	52.57
DML[22]	34.40	62.15	75.89	-	-	-	-	-	-
kLFDA[27]	32.33	65.78	79.72	42.76	69.01	79.63	48.20	59.34	66.38
IDLA[28]	34.81	63.32	74.79	65.00	89.50	93.00	54.74	86.50	94.00
FPNN[13]	-	-	-	27.87	64.00	77.00	20.65	51.00	67.00
DeepRanking[26]	38.37	69.22	81.22	70.94	92.30	96.90	-	-	-
MetriceEnsembles[30]	45.90	77.50	88.90	-	-	-	62.10	89.10	94.30
DeepRDC[8]	40.50	60.80	70.40	-	-	-	-	-	-
DeepLDA[29]	44.11	72.59	81.66	67.12	89.45	91.68	62.23	89.95	92.73
SIRCIR[25]	35.76	67.00	82.50	72.50	91.00	95.50	52.17	85.00	92.00
NullReid[24]	42.28	71.46	82.94	-	-	-	58.90	85.60	92.45
GOG[23]	49.70	79.70	88.70	-	-	-	67.30	91.00	96.00
Gated S-CNN[6]	37.80	66.90	77.40	-	-	-	68.10	88.10	94.60
DGD[7]	35.40	62.30	69.30	-	-	-	80.50	94.90	97.10
VGG16-Baseline[17]	32.37	60.03	72.75	53.43	75.39	85.27	58.43	75.35	84.95
Ours(VGG16)	46.89	76.49	85.84	70.13	88.42	92.12	71.56	90.65	94.70
ResNet-50-Baseline[18]	47.30	78.34	87.98	70.76	89.73	94.92	72.94	91.26	95.03
Ours(ResNet-50)	**51.30**	**82.32**	**90.24**	**74.10**	**93.53**	**98.12**	**80.80**	**95.03**	**98.06**

They share weights and predict the labels of the pedestrian. Analogously, we use cross-entropy loss function to supervise the identification procedure:

$$q^r = softmax_{identification}(\tilde{x}_i \circ \theta_{idntif}) \tag{4}$$

$$L_{identification}(\theta_{identif}, r) = \sum_{i=1}^{N} q_i^r log\left(\frac{1}{q_i}\right) \tag{5}$$

Here, r is the target class, q_i^r is the predict probability, and the transformation is parametrized by $\theta_{identif}$. For all images $q_i = 0$ except the ith image is classified successfully whilst $q_i = 1$.

D. Fusion Loss and Stepwise Training

We formulate the final fusion loss function as:

$$L_{fusion}(\theta, r, s) = \lambda L_{verification} + L_{identification} \tag{6}$$

Here λ is a coefficient to balance the weight of the two elements of the fusion loss function. It is determined as 3 through the cross validation experiment which is visualized in fig. 2. In this work, we use two large scale Re-ID datasets and two smaller ones. Follow the order: ImageNet → Large scale datasets → general scale datasets → small scale dataset. We adopt a stepwise training strategy to improve the performance of our model in small volume circumstance due to the insufficient data may cause overfitting and generate the reduction in accuracy. It is worth mentioning the training order abide by the knowledge transferring principle, the further behind layer in spatial sequence shows a higher semantic relevant characteristic during the fine-tuning process.

III. EXPERIMENT

A. Datasets

1) CUHK03: CUHK03 consists of 13164 images from 1360 identities. It provides two settings, one is annotated by human (labeled) and the other one is annotated by detectors (detected). We adopt the detected setting since it is closer to practical scenarios. We do 20 random splits, wherein 1160 identities are for training, 100 identities are for testing. The evaluation is in single shot.

2) Market1501: Market1501 contains 32,668 detected person bounding boxes of 1,501 identities from 6 cameras. We use the training and test splits provided in under both the single-query (SQ) and multi-query (MQ) evaluation settings.

3) CUHK01: CUHK01 contains 971 individuals captured from two camera views, and each identity from each view has two images. There are two settings; the first is the single-shot setting, that is, one image for each individual in each camera view is randomly selected for both training and testing, and 485 identities are used for training and the other 486 for testing. Under the other setting only 100 identities are used for testing with the rest 871 for training.

4) VIPeR: VIPeR contains 632 identities with two camera views. Each identity from each view has one image.

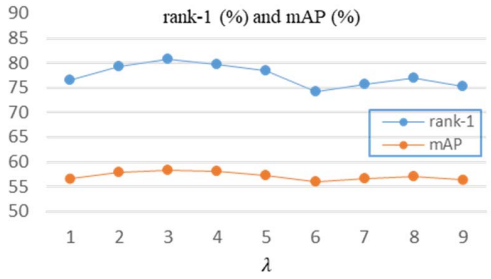

Figure 2. Rank-1 and mAP accuracy validation on Market-1501[3]

TABLE II. THE CMC PERFORMANCE ON MARKET-1501[3]

Method	Single Query		Multi Query	
	rank-1	*mAP*	*rank-1*	*mAP*
Gated S-CNN[6]	65.88	39.55	76.04	48.45
NullReid[24]	55.43	29.87	71.56	46.03
DADreid[32]	39.40	19.60	49.00	25.80
VGG16-Baseline[17]	63.78	36.54	72.45	51.03
Ours(VGG16)	68.73	46.43	75.26	55,84
ResNet-50-Baseline[18]	72.34	50.34	79.43	57.32
Ours(ResNet-50)	78.03	58.34	82.97	64.32

Half of the identities are used for training, and the other half are for testing. The evaluation is also based on 10 random splits, in single shot.

B. Evaluation Metrics

We use the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP). The average precision (AP) of each query is computed from its precision-recall data. Then mAP is the mean value of average precisions across all queries. It reflects the recall while CMC reflects the precision.

C. Implementation Details

We use the Caffe framework in the implementation of our method. This section will introduce some details about input data preparation, training settings, data augmentation and loss function selection.

1) Input data preparation: We shuffle the dataset and make a random order of the images. To compose the negative pair for training, we sample another image from a different class. In order to reduce prediction bias, the initial number of positive and negative pairs is same. For fear of the network risks over-fitting since the positive pairs are so limited, we multiple the ratio between positive and negative pairs by a factor of 1.01 every epoch during the training until it attains 3:1.

2) Training settings: We adopt the mini-batch stochastic gradient descent (SGD) to update the parameters of the network. The batch size is set to 48 for VGG-16 and 64 for ResNet-50. The maximum training epoch is set to 75. The learning rate is initialized as 0.001 and then set to 0.0001 for the last 10 epochs. For the network updating, we compute all the gradients produced by every objective respectively and add them up. A weight of 3 to the gradient produced by the verification loss and 1 for the gradients produced by two identification losses are assigned in the training. The coefficient λ is set to 3 through validation experiment as fig.2. illustrated

TABLE III. VALIDATION RESULTS ON CUHK03[2] DATASET

Network	*ImageNet Pre-train?*	*Rank-1*	*mAP*
VGG16 (V)	YES	41.32	22.46
	NO	40.98	22.33
VGG16 (I)	YES	64.32	37.03
	NO	43.23	22.58
VGG16 (V+I)	YES	69.42	45.85
	NO	51.76	32.13
ResNet-50 (V)	YES	65.20	43.54
	NO	65.77	43.76
ResNet-50 (I)	YES	72.13	50.93
	NO	53.24	31.39
ResNet-50 (V+I)	YES	**80.80**	**58.30**
	NO	64.56	45.30

The second column indicates that whether the model is pre-trained on ImageNet[31] or train from scratch and using verification loss or identification loss respectively or jointly. "V" denotes verification loss and "I" denotes identification loss.

3) Data augmentation: To counter over-fitting since the data volume may not adequate, we also perform data augmentation on the Re-ID datasets as in most deep Re-ID works. We resize all the training images to 256×256 and randomly crop the images to 224×224 generate 5 augmented images around the image center by performing random 2D transformation for each training image. For VIPeR dataset, we also deploy the horizontal reflection.

4) Loss selection: In order to figure out the implicit relationship between loss function and their contribution to the Re-ID task. We also solely trained the model with either identification loss or verification loss. Results are shown in table 3. we could find that the identification loss plays a decisive role in the knowledge transferring from ImageNet[31] to Re-ID database. The evaluation index drop by a large margin without either of the loss function, suggest that fusion loss is effective and necessary explicitly.

D. Comparison with VGG16 and ResNet-50 Baseline

To get more data evidence to verify superiority of our method. We also compare the model with VGG-16[17] and ResNet-50[18] basenetwork. For each model we have pre-train and train from scratch two train strategies. The results are shown in table 3. Prove that the knowledge extract from ImageNet dataset is effective and has been successfully fed into subsequent CNN layers for Re-ID oriented tasks. With the semantic hierarchy rises, the general knowledge derivate into specialized knowledge for pedestrian re-identification.

IV. CONCLUSION

In this work, we propose a novel Siamese-subnetwork deep transfer learning model which considers both verification and identification losses to address the challenging person Re-ID problem. The result of the experiments validated the effectiveness of the knowledge transfer from large image datasets to small ones. Our

method outperforms state-of-the-art on five popular person Re-ID datasets and shows the advantage of the application of fusion loss.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the China National Natural Science Foundation under Grant No. 61673299, 61203247, No. 61573259, 61573255. It was also supported by the Fundamental Research Funds for the Central Universities (Grant No. 0800219327). It was also partially supported by the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) under Grant No. MJUKF201721.

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.

[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.

[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.

[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. 2016.

[7] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In CVPR, 2016.

[8] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition, 48(10):2993–3003, 2015.

[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.

[11] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a siamese time delay neural network" International Journal of Pattern Recognition and Artificial Intelligence, vol. 7, 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 CVPR 2016.

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

[14] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In Scandinavian conference on Image analysis 2011.

[15] H. Hu, Z. Lin, J. Feng, and J. Zhou. Smooth representation clustering. In CVPR 2014

[16] C. Huang, C. C. Loy, and X. Tang. Unsupervised learning of discriminative attributes and visual representations. In CVPR, 2016.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[18] 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.

[19] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In ICML, 2007.

[20] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H.Bischof. Large scale metric learning from equivalence constraints. In CVPR, 2012.

[21] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In CVPR, 2013.

[22] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In ICPR, 2014.

[23] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical Gaussian descriptor for person re-identification. In CVPR, 2016.

[24] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In CVPR, 2016.

[25] 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 CVPR, 2016.

[26] S.Z. Chen, C.C. Guo, and J.-H. Lai. Deep ranking for person re-identification via joint representation learning. TIP, 25(5):2353–2367, 2016.

[27] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In ECCV, 2014.

[28] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In CVPR, 2015.

[29] L. Wu, C. Shen, and A. van den Hengel. Deep linear dis-criminant analysis on fisher networks: A hybrid architecture for person re-identification. Pattern Recognition, 2016.

[30] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In CVPR, 2015.

[31] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248–255. IEEE, 2009.