What Makes You Look Like You: Learning an Inherent Feature Representation for Person Re-Identification

Wen-Li Wei
Institute of Information Science
Academia Sinica, Taiwan
lilijinjin@google.com

Jen-Chun Lin
Department of Electrical Engineering
Yuan Ze University, Taiwan
Institute of Information Science
Academia Sinica, Taiwan
jenchunlin@gmail.com

Yen-Yu Lin
Research Center for IT Innovation
Academia Sinica, Taiwan
yylin@citi.sinica.edu.tw

Hong-Yuan Mark Liao
Institute of Information Science
Academia Sinica, Taiwan
liao@iis.sinica.edu.tw

Abstract

In this work, we address person re-identification (ReID) by learning an inherent feature representation (inherent code) that is unique to each individual. This task is difficult because the appearance of a person may vary dramatically due to diverse factors, such as illuminations, viewpoints, and human pose changes. To tackle this issue, we propose new learning objectives to learn the inherent code for each person based on deep learning. Specifically, the proposed deep-net model is trained by jointly optimizing the multiple objectives that pulls the instances of the same person closer while pushing the instances belonging to different persons far from each other. Owing to such complementary designs, the deep-net model yields a robust code for each individual and hence better solve person ReID. Promising experimental results demonstrate the robustness and effectiveness of our proposed method.

1. Introduction

Person re-identification (ReID) is an active topic in computer vision and pattern recognition. It aims at determining whether a query and a gallery image correspond to the same person by measuring their visual similarity. As a crucial component for video content understanding, ReID is essential to a broad range of applications to intelligent surveillance, such as searching criminal suspects or missing children from a large surveillance camera network. However, identifying persons from their visual appearance is non-trivial owing to the large intra-person visual variations caused by different illuminations, viewpoints, and poses [14, 4, 5, 21, 17, 24, 26, 25, 19, 9, 22].

With the success of deep learning, deep-net models such as convolutional neural networks (CNN) have been applied to person ReID and achieved significant performance improvement [14, 4, 5, 21, 17, 24, 26, 25, 19, 9, 12, 8, 23, 6, 3, 1, 20, 15]. The core behind these methods is to learn robust feature representations for persons to be identified. One of the key components for this task is the triplet loss or its variants for learning a CNN-based embedding function [14, 5, 21, 25, 9, 8, 6, 3]. The triplet loss optimizes the embedding space such that data points of the same identity are closer to each other than those of different identities. However, this loss cannot guarantee that data points of the same identity, *i.e.* anchor and positive examples, are close to each other when unfavorable intra-person variations are present. As a consequence, instances of a person may form a large cluster with large intra-class distances in the learned feature space [5]. This issue currently hinders the advances in building an accurate person ReID system, even if some studies have tried to alleviate this issue by developing the variants of the triplet loss [5, 25, 8, 3].

To address this issue, we propose new deep-learning-based objectives to derive an inherent feature representation (inherent code) for each individual, which emphasizes both intra- and inter-class (identity) distances. Specifically, three loss functions with complementary properties are designed for deep-net model learning. They are 1) the *classification loss* for deriving the feature representation in a discriminative manner, 2) the *compact loss* for enforcing different instances of a person to have similar feature representations, and 3) the *scatter loss* for pushing the feature representations belonging to different identities far from each other. By jointly optimizing the three objectives, the resultant deep-net model can learn a unique and identifiable feature representation for each individual to better solve person ReID. In summary, the main contributions of this work lie in

- We present new learning objectives, the compact loss and the scatter loss, to help supervise the learning of deep-net models and produce the inherent feature representation of each individual.
- The proposed objectives are flexible in the sense that they can be integrated into different deep-net models and be trained in an end-to-end fashion.
- The proposed method is evaluated quantitatively and qualitatively. The experimental results demonstrate that the deep-net model derived with the proposed objectives performs favorably against those based on the triplet loss.

The remainder of this paper is organized as follows. Section 2 reviews the previous work on person ReID. The technical details of the proposed objectives are described in Section 3. The experimental results are presented in Section 4. Finally, Section 5 concludes this work.

2. Related work

Person ReID has been studied extensively in the literature of computer vision and pattern recognition, *e.g.* [14, 4, 5, 21, 17, 24, 26, 25, 19, 9, 12, 8, 23, 6, 3, 1, 20]. Inheriting CNN for joint feature extraction and nonlinear classifier learning, CNN-based methods have currently dominated this field. They typically learn identifiable feature representations in an end-to-end fashion through various metric learning losses such as the contrastive loss [23], triplet loss [6], improved triplet loss [5], quadruplet loss [3], and hard triplet loss [8]. Among these metrics, the most popular and successful ones may be the triplet loss and its variants. Ding *et al.* [6] use the triplet training examples and the triplet loss to learn the feature representation for person ReID. Liu *et al.* [14] propose the visual attention model that is a triplet recurrent neural network which takes several glimpses of triplet images of persons and dynamically generates comparative attention location maps for person ReID. Cheng *et al.* [5] ultimize an improved triplet loss function based on the multi-channel parts-based CNN model, which attempts to pull the instances of the same person closer in the learned feature space. Hermans *et al.* [8] present a hard triplet loss that employs the hard sample mining to select suitable samples for training the embedding function. Similarly, Xiao *et al.* [25] design a margin sample mining loss (MSML) to improve the triplet loss through modified hard sample mining. Zhang *et al.* [26] develop AlignedReID that extracts a global feature learned together with local features by using the hard triplet loss proposed in [8]. In summary, most of the modified triplet losses are still being studied to maximize the inter-class distances that ignore intra-class variations [8, 25]. Although some methods attempt to apply a simple predefined margin to reduce intra-class distance, performance improvements are limited [5]. To this end, we propose new learning objectives to explicitly minimizing intra-class distance while maximizing inter-class distance.

3. Our approach

Fig. 1 illustrates the proposed framework for person ReID. The deep-net model based on ResNet18 is optimized by jointly using three learning objectives, including the classification loss, compact loss, and scatter loss. Given a query image, the optimized deep-net model generates the inherent feature representation, called the inherent code in this work, for the query. A distance metric is then adopted to measure the distances between the query and the gallery images based on their inherent codes. A distance ranking function is applied to complete person ReID.

3.1. Inherent Code Learning for Person ReID

Developing effective learning objectives to improve the discriminative power of the learned feature representations is critical to person ReID. Our idea to enhance the widely-used triplet loss is inspired by the cluster analysis and representation learning [7, 2], especially the cluster properties, which emphasize the notions of high intra-class similarity and low inter-class similarity. To this end, we propose three learning objectives for the deep-net model to learn an inherent code of each individual, which is devoted to minimizing intra-class variations while keeping the codes of different classes separable, *i.e.* emphasizing the inter-class distances.

The first objective, the classification loss illustrated in Fig. 2(a), helps learn a discriminative inherent code by making the codes of different classes separable.

For each training sample of class y, this objective employing *softmax* regression [27, 13] as the loss function in the output layer is defined as

$$\mathcal{L}_{SR} = KL(\mathbf{x}^{(L)}, y), \tag{1}$$

where $KL(\cdot, \cdot)$ represents the KL divergences, and $\mathbf{x}^{(L)}$ is the *posterior* probability of this training sample in the output layer L of the deep-net model.

Since the viewpoints of the query and the gallery images may be different, the variations caused viewpoints will degrade the performance of person ReID. To address this problem, inspired from [10], we propose the compact loss $\mathcal{L}_{CL}$ that makes the most of person orientation information and can learn a viewpoint invariant representation, making

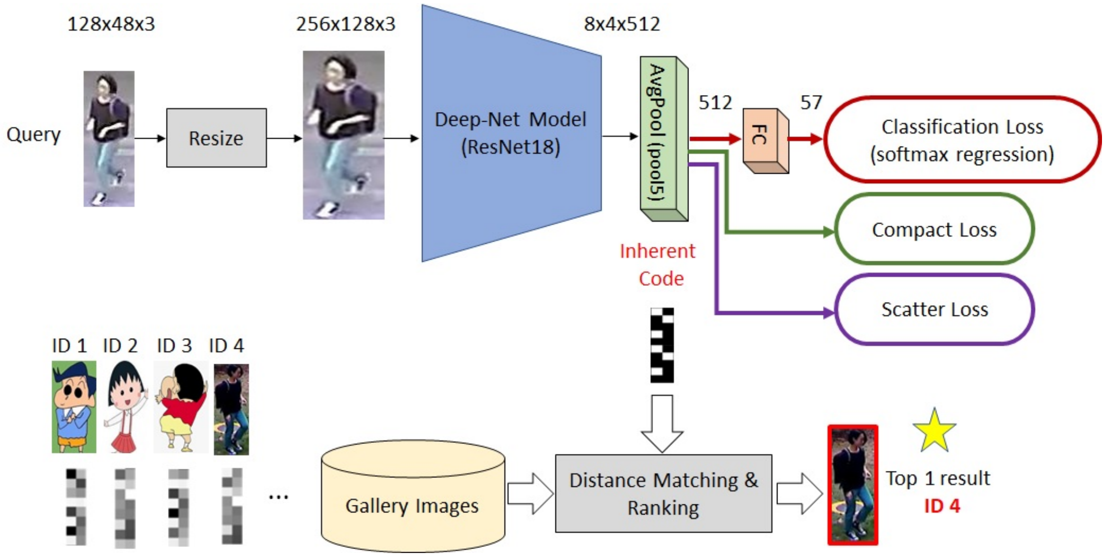

Figure 1. The proposed person re-identification framework that computes the inherent code for the query, with which ReID is carried out by distance matching and ranking.

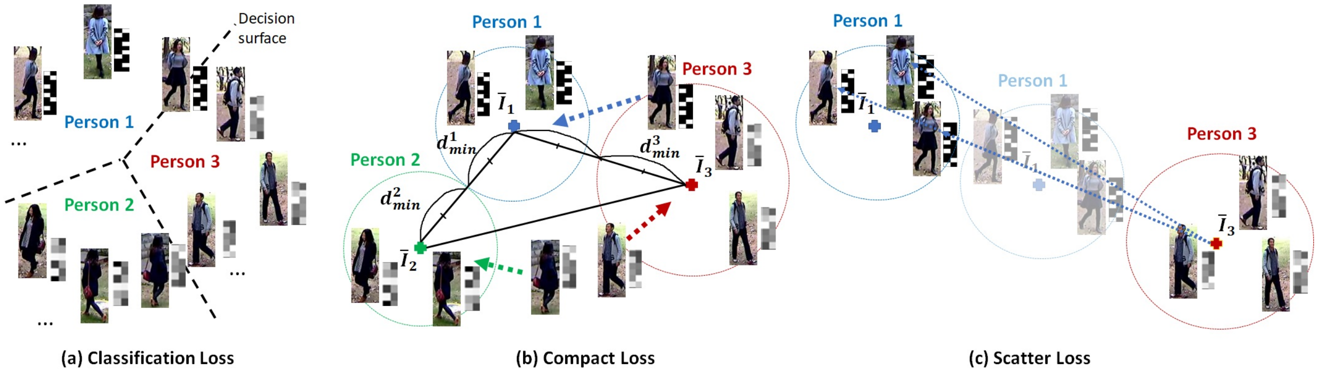

Figure 2. Illustration of the three loss functions developed for inherent code learning.

the images of a person taken with diverse viewpoints share similar inherent codes. Formally, to mitigate intra-class variations due to subject appearance variations, we devise the term $\mathcal{L}_{CL}$ as follows,

$$\mathcal{L}_{CL} = \frac{1}{2}\sum_{p,i} g(\|I_{p,i} - \bar{I}_p\|_2^2 - (d^p_{min})^2), \quad (2)$$

where $I_{p,i}$ represents the inherent code of person p of the ith viewpoint extracted at the output of pool5 layer of ResNet18, and $\bar{I}_p$ is the mean (so-called center) of inherent codes of person p under various viewpoints calculated at the beginning of every batch. The value of d^p_{min} is set to half the minimum distance between $\bar{I}_p$ and $\bar{I}_j$ for subject $j \neq p$. The function $g(w)$ is defined as $g(w) = max(0, w)$. By minimizing $\mathcal{L}_{CL}$, the inherent codes of images of a person are prone to gather together in a multi-dimensional sphere, where the radius and the center of the sphere are specified by d^p_{min} and $\bar{I}_p$, respectively. As illustrated in Fig. 2(b), this objective function reduces the intra-class variations of the learned inherent codes and improves person ReID.

The inherent code learned through the first two objectives is already able to reduce the intra-class scatter and distinguish images of different classes (persons). However, these objectives do not explicitly enforce high inter-class distances. The discriminative power of the inherent codes can be further improved. To this end, we propose the scatter loss $\mathcal{L}_{SL}$ to increase the inter-class distances of the learned inherent codes. The loss $\mathcal{L}_{SL}$ is devised as

$$\mathcal{L}_{SL} = \frac{1}{2}\sum_{p,i,j} \|I_{p,i} - \bar{I}_j\|_2^2, \quad (3)$$

where $\bar{I}_j$ is the mean of the inherent codes of person j for each $j \neq p$. As shown in Fig. 2(c), maximizing $\mathcal{L}_{SL}$ enlarges inter-class distances of the learned inherent codes,

Table 1. Rank-1, Rank-5, Rank-10, Rank-20 accuracy (%) and mAP of five different methods on the PKU-ReID dataset.

Methods	Rank 1	Rank 5	Rank 10	Rank 20	mAP
ResNet50$_{\text{htri [8]}}$	55.3	78.3	86.0	93.2	47.7
ResNet18$_{\text{htri [8]}}$	66.9	85.5	93.6	97.8	55.6
Ours$_{\text{set1}}$	67.3	87.1	94.1	97.4	51.5
Ours$_{\text{set2}}$	70.2	88.2	92.8	96.7	53.5
Ours$_{\text{set3}}$	**71.5**	**90.8**	**95.6**	**98.9**	**56.8**

i.e. pushing the inherent codes belonging to different persons far away from each other, thereby enhancing the discriminative power.

The total loss $\mathcal{L}$ for training is a weighted combination of all the three loss functions, *i.e.*

$$\mathcal{L} = \mathcal{L}_{SR} + \beta\mathcal{L}_{CL} - \gamma\mathcal{L}_{SL}, \tag{4}$$

where β and γ are non-negative hyper-parameters, and are set to 5×10^{-5} and 1×10^{-6} in the experiments, respectively. The total loss $\mathcal{L}$ can be back propagated to learn the parameters of the deep-net model through classical backpropagation.

After training, the trained deep-net model can compile the inherent code for an input image. During testing, the inherent code of a query image is generated from the output of the pool5 layer of deep-net model through forward propagation, as illustrated in Fig. 1. The L2 distance is then used to compute the dissimilarity between the inherent codes of the query and each gallery image. Finally, all the gallery images are ranked in the ascending order of distances, and the top one is retrieved as the result of person ReID.

4. Experimental results

To evaluate the proposed loss functions, we conduct experiments on the recently published dataset PKU-ReID [16]. This dataset contains 114 individuals, including 1,824 images captured with two disjoint camera views. For each person, eight images are captured with eight different orientations under one camera view and normalized to 128×48 pixels. For evaluation, we split the dataset into two equal-size parts randomly. One contains 57 individuals for training, and the rest contains other 57 individuals for testing. During testing, we further split the test set into two subsets, including the query and the gallery, with disjoint cameras. To the best of our knowledge, PKU-ReID is the first dataset that collects person appearances with eight orientations.

We compare the proposed loss functions with the popular hard triplet loss [8], based on the ResNet framework. Regarding the experimental setup, we use ResNet18 and ResNet50 pre-trained on ImageNet [18] as the base models. Random horizontal flipping and cropping are used to augment the data. The margin of the hard triplet loss is set to 0.3. For training, the batch size and the learning rate are set to 32 and 2×10^{-4}, respectively. The Adam solver [11] is adopted. Both models are trained by 300 epochs. Techniques such as batch normalization and L2 weight decay regularization are used to alleviate overfitting. For testing, the feature representation in the two ResNet frameworks is extracted from the output of pool5 layer. For person ReID, the sum of squared distances is used to compare the distance between query image and gallery image in the mapped feature space. The rank-1, 5, 10, 20 accuracy and mean average precision (mAP) are adopted as the performance measure, where the rank-i accuracy is the mean accuracy that images of the same identity appear in the top-i returns.

We first evaluate the performance of the hard triplet loss under the ResNet frameworks, named ResNet18$_{\text{htri}}$ and ResNet50$_{\text{htri}}$. The results in Table 1 demonstrate that ResNet18$_{\text{htri}}$ outperforms ResNet50$_{\text{htri}}$ in our experiments. This could be due to that the use of more parameters in ResNet50$_{\text{htri}}$ is prone to trap itself into the problems of data sparseness and overfitting. Although techniques such as batch normalization and L2 weight decay regularization have been used in ResNet50$_{\text{htri}}$ to alleviate overfitting, performance improvements are still limited. Driven by these findings, we choose ResNet18 as the default model for our approach. Regarding our approach, we first evaluate the performance of the designed learning objectives. Compared to Ours$_{\text{set1}}$ where only the classification loss is active, Ours$_{\text{set2}}$ improves the rank-1 accuracy and mAP by additionally considering the compact loss. It reveals that compact loss can effectively make images of a person with different viewpoints have similar inherent codes, thus reducing the matching error caused by feature variations. Ours$_{\text{set3}}$ (full version of our approach) outperforms Ours$_{\text{set2}}$ by additionally including the scatter loss in the learning objective. The expected improvement results from the scatter loss which enlarges the inter-class distances, thereby increasing the discriminative power of the learned inherent code. The results in Table 1 demonstrate that the proposed Ours$_{\text{set3}}$ is superior to the hard triplet loss-based Resnet framework (ResNet18$_{\text{htri}}$). One main reason is that the hard triplet loss does not emphasize the intra-class similarity in the embedded function construction. It may lose the discriminative power for person ReID, since the visual appearances of the same person in different images may be very dissimilar due to illuminations, viewpoints and human pose changes. The proposed learning objectives used in the ResNet framework, Ours$_{\text{set3}}$, effectively reduce the intra-class variations while increasing inter-class scatter for learning robust codes. Owing to such complementary designs, Ours$_{\text{set3}}$ pushes ahead the rank-1 accuracy by about

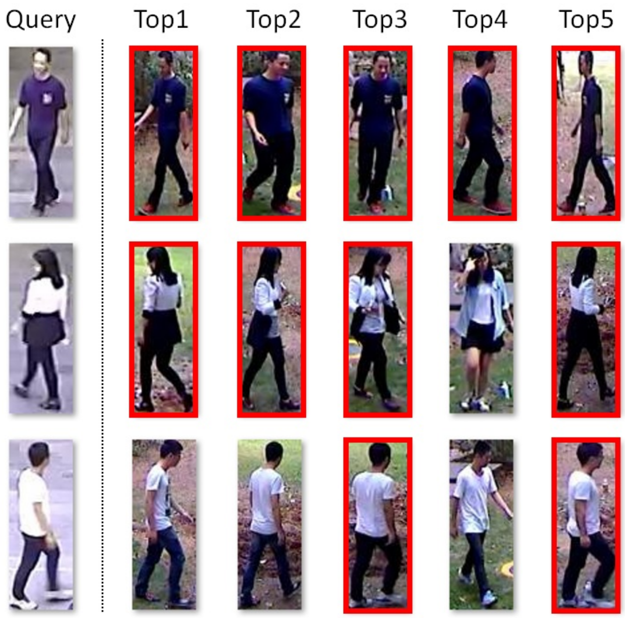

Figure 3. An example of our (Ours$_{\text{set3}}$) person ReID results. The red bounding boxes indicate the correct matches. From top to bottom, three examples represent the simple, moderate, and hard cases in the PKU-ReID dataset, respectively.

4.6%, compared to ResNet18$_{\text{htri}}$. The mAP is improved from 55.6% to 56.8%. To gain insight into the quantitative performance, Fig. 3 shows an example of our (Ours$_{\text{set3}}$) person ReID results.

5. Conclusions and future work

We have introduced new learning objectives for the deep-net model, including compact loss and scatter loss, to learn a unique and identifiable feature representation for each individual to better solve person ReID. The experimental results have demonstrated that the proposed learning objectives outperform the popular hard triplet loss, and can offer satisfactory person ReID results. Based on the promising outcomes, our future work will focus on developing a more effective strategy to improve the fineness of the learned feature representation, such as integrating local feature learning into our deep-net framework. We also plan to expand the scale of the experiments in the future.

Acknowledgment: This work was supported in part by MOST grants 108-2634-F-001-003 and 107-2218-E-155-011-MY2.

References

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

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017.

[4] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017.

[5] 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.

[6] 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, October 2015.

[7] B. S. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster analysis*. U.K: Wiley, 2011.

[8] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. In *ArXiv e-prints*, 2017.

[9] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

[10] D. H. Kim, W. J. Baddar, and Y. M. Ro. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In *ACM MM*, 2016.

[11] D. Kingma and J. B. Adam. A method for stochastic optimization. In *ICLR*, 2015.

[12] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[13] J.-C. Lin, W.-L. Wei, and H.-M. Wang. Demv-matchmaker: Emotional temporal course representation and deep similarity matching for automatic music video generation. In *ICASSP*, 2016.

[14] H. Liu, J. Feng, M. Qi, J. Jiang, and S.-C. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, May 2017.

[15] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR*, 2019.

[16] L. Ma, H. Liu, L. Hu, C. Wang, and Q. Sun. Orientation driven bag of appearances for person re-identification. In *ArXiv e-prints*, 2016.

[17] T. Matsukawa and E. Suzuki. Person re-identification using cnn features learned from combination of attributes. In *ICPR*, 2016.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. B. *et al.* Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[19] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018.

[20] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016.

[21] Y. Suh, J. Wang, S. Tang, T. Mei, and K.-M. Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.

[22] X. Sun and L. Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, 2019.

[23] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human reidentification. In *ECCV*, 2016.

[24] L. Wu, Y. Wang, J. Gao, and D. Tao. Deep co-attention based comparators for relative representation learning in person re-identification. In *ArXiv e-prints*, 2018.

[25] Q. Xiao, H. Luo, and C. Zhang. Margin sample mining loss: a deep learning based method for person re-identification. In *ArXiv e-prints*, 2017.

[26] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. In *ArXiv e-prints*, 2018.

[27] Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao. Cross-media hashing with neural networks. In *ACM MM*, 2014.