UNSUPERVISED PERSON RE-IDENTIFICATION VIA NEAREST NEIGHBOR COLLABORATIVE TRAINING STRATEGY

Qing Tang and Kang-Hyun Jo

University of Ulsan, Ulsan, Republic of Korea
Department of Electrical and Computer Engineering
tangqing@islab.ulsan.ac.kr; acejo@ulsan.ac.kr

ABSTRACT

Because of the lack of human-labeled data, the challenge of unsupervised person re-identification (re-ID) is to learn to generate correct pseudo labels for training. Unlike the human-labeled annotation, the generated pseudo labels contain the noise labels that harm the model's performance. In this paper, we propose the Nearest Neighbors Collaborative Training (NNCT) strategy to mitigate the effects of noisy labels by utilizing information of the nearest neighbor of an image. The proposed NNCT trains the image and its nearest neighbor collaboratively, thereby enhancing the generalization capability of the network and shortening the distance with neighbors. To make training using the up-to-date nearest neighbor possible, we introduce a Pseudo Label Memory Bank (PLMB) to store the up-to-date labels of all images. The experimental results confirm the superiority of the proposed method, which surpasses state-of-the-arts on two mainstream person re-ID datasets, Market-1501, and DukeMTMC-reID in both fully unsupervised learning manner and Unsupervised Domain Adaptation (UDA) manner.

Index Terms— Person re-identification, unsupervised learning, unsupervised domain adaption, pseudo label refinery

1. INTRODUCTION

The person re-identification (re-ID) system aims to retrieve images that contain the same identity. The supervised person re-ID requires substantial labeled training data for satisfying performance. Therefore, some recent works focus on using the unsupervised person re-ID method [1–15] to train the network without human-labeled annotations. It is challenging to capture discriminative features without any supervised information. To make unsupervised training possible, the pseudo label for each image is pre-generated or on-line generated by clustering algorithm [3, 10, 15] or similarity measurements [5, 12]. Unlike human-labeled annotation, such generated pseudo labels contain the noise labels that substantially hinder the model's capability to extract discriminative features because the features are learned based on these pseudo labels. Because of quality of labels, the performance of unsupervised person re-ID still significantly falls behind the supervised person re-ID.

Consequently, the key to improving the unsupervised person re-ID model performance is to generate high-quality pseudo labels which can represent the target-domain distribution. Several studies [7–15] utilize unsupervised domain adaption (UDA) to adapt the model from the labeled source dataset to the unlabeled target dataset. The key of UDA is reducing the gap between the domains of the source and target dataset. ECN [12] found out that the relations among target dataset images also contain critical factors that influence the model performance. Hence, ECN [12] constructed constraints by considering the intra-domain variations in the target domain to push the network to learn relations among images of target datasets. Using available information and constraints has become the mainstream method to improve the unsupervised person re-ID performance.

We observe that humans infer others' identity more accurately by adjusting the view angle. It is because that these multi-view images can be served as additional references to provide more information about the identity. Inspired by it, we intend to utilize additional images as reference information in this paper. However, the person-ID and camera-ID are unknowable in the unsupervised learning task. To simulate this process possible, we propose the Nearest Neighbors Collaborative Training (NNCT) strategy, which trains the model using the image and its neighbor images. Compared with previous work [5] only considers the current image, the proposed NNCT treats the neighbor images as additional references when computing the loss; thereby the NNCT is optimized with more comprehensive information during the loss back-propagation. The prerequisite of our hypothesis is that the model can roughly capture the target domain distribution; hence, the image and its neighbors contain the same identity with high probability.

The architecture of our proposed NNCT model is illustrated in Fig. 1. For an image $x_i \in X$, we first compute its similarity s_i with other images in X. Based on the computed similarity, the pseudo label $\bar{y}_i$ of the image x_i is predicted, and

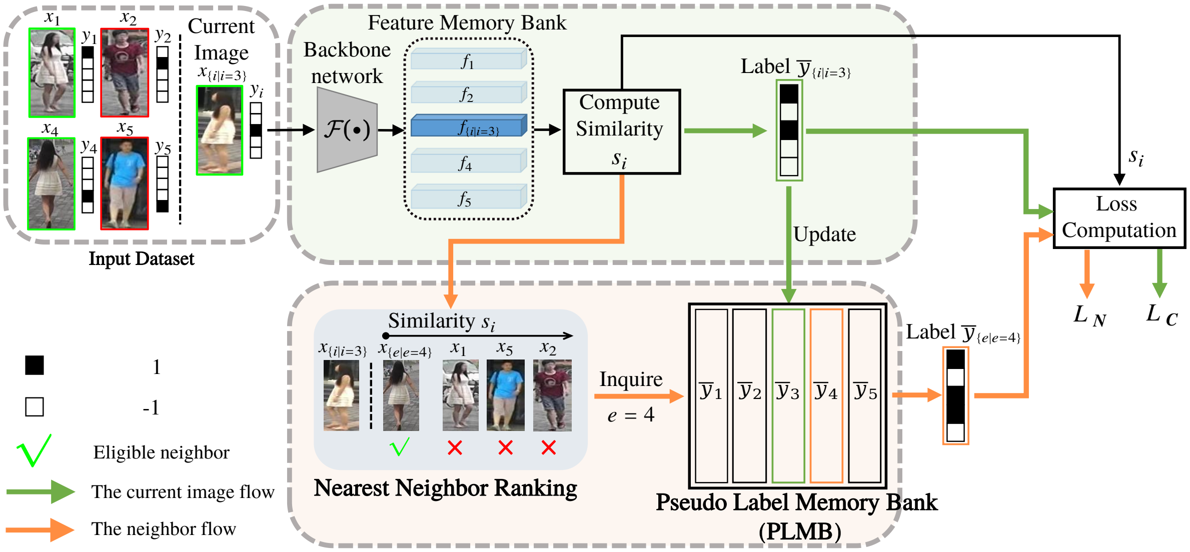

Fig. 1. The framework of the proposed NNCT. The k-nearest neighbor are selected as the eligible neighbor, $k = 1$ in the figure.

the eligible neighbors of x_i are selected. The Pseudo Label Memory Bank (PLMB) is constructed to store the up-to-date labels of all images, which is updated by $\bar{y}_i$ in each training iteration. With the help of PLMB, the up-to-date pseudo label of eligible neighbor can be inquired, notated as $\bar{y}_e$. Except for predicted $\bar{y}_i$, we use $\bar{y}_e$ to train the re-ID model collaboratively. During the loss back-propagation, the network $\mathcal{F}(\cdot)$ learns to extract more discriminative features for person re-ID by leveraging the additional reference information provided by neighbors.

Our contributions are highlighted as follows. (1) We propose Nearest Neighbor collaborative Training (NNCT) strategy to leverage the eligible neighbors as additional reference information to mitigate the effects of noisy labels for unsupervised person re-ID. (2) We propose a lookup table called PLMB to store and inquire the up-to-date pseudo labels for all images. (3) The performance of the proposed NNCT surpasses state-of-the-arts methods on the Market-1501 and DukeMTMC-reID dataset in both the fully supervised learning method and the UDA-based method.

2. PROPOSED APPROACH

Our proposed NNCT framework is designed based on the multi-class label-based unsupervised re-ID methods described in [5]. The NNCT is divided into three stages: feature similarity computation, the current image flow, and the neighbor flow.

2.1. The Overview of NNCT

The NNCT framework is shown in Fig. 1. Given a set of unlabeled person images $x_{\{i|i=1,2,...,n\}} \in X$, we regard each image as an individual category, labeled as a n-dimensional single-class label y_i. n is the number of images in input dataset X. Then, d-dimensional feature f_i of x_i are extracted by backbone network $\mathcal{F}(\cdot)$ to form the feature memory bank $\mathcal{M}$. $\mathcal{M}$ serves as a feature-storage for all images in X. The size of $\mathcal{M}$ is $n \times d$. Using $\mathcal{M}$, the similarity between x_i and the other image x_j is computed as,

$$s_i[j] = f_i \times f_j^{\top}, \quad j = 1, ..., n. \tag{1}$$

where s_i is an n-dimensional vector.

After similarity computation, the model is mainly divided into two flows. We first introduce the current image flow, and the details of the proposed neighbor flow will be presented in the next subsection. Based on s_i, the n-dimensional multi-class label $\bar{y}_i$ is predicted by Memory-based Positive Label Prediction (MPLP) described in [5]. Based on the s_i, the MPLP aims to predict whether the image x_j contains same identity with x_i or not. There are two steps in MPLP. Firstly, if $s_i[j] > 0.6$, x_j would be treated as a positive sample for x_i In other words, $\bar{y}_i[j] = 1$; otherwise, $\bar{y}_i[j] = -1$. Then, the $\bar{y}_i$ is further filtered by cycle consistency. For any $\bar{y}_i[j] = 1$, if $\bar{y}_j[i] = 1$ simultaneously, x_j would be considered as a positive sample for x_i. After these two steps, the multi-class pseudo label $\bar{y}_i$ of x_i is obtained. In the current image flow, we compute the current image loss $\mathcal{L}_C$ to regress the s_i to the current image's pseudo label $\bar{y}_i$.

Based on the similarity s_i, the eligible neighbor image x_e is selected from X by the nearest neighbor ranking. e represents the index of eligible neighbor in X. Thanks to the PLMB, the up-to-date pseudo label of eligible neighbor can be inquired to further regress the similarity score s_i.

2.2. The Neighbor Flow

2.2.1. Nearest Neighbor Ranking

As mentioned in Section 1, the prerequisite of our proposed collaborative training strategy is that the selected eligible neighbor contains a same identity as x_i with high probability. If not, the image which contains a different identity as x_i will be used to train the x_i, which hinder the model's capability.

The nearer neighbors are more related to x_i, having higher probabilities of sharing the same multi-class label. Thus, the nearest neighbor ranking algorithm is used to rank all images in X according to its similarity s_i. Then, the k-nearest neighbors are selected as eligible neighbors. The model performance with different k will be tested in Section 3. As shown in Fig. 1, $x_{\{e|e=4\}}$ is the selected eligible neighbor which is the first nearest neighbor of $x_{\{i|i=3\}}$.

2.2.2. Pseudo Label Memory Bank (PLMB)

In order to make training with eligible neighbor possible and accelerate inquiry speed on whole target dataset X, we propose a Pseudo Label Memory Bank (PLMB), notated as $\mathcal{B}$, which store the pseudo labels $\bar{y}_{\{i|i=1,2,\ldots,n\}}$ of all images $x_{\{i|i=1,2,\ldots,n\}} \in X$. Thus, the PLMB contains n slots, in which each slot storing a n-dimensional pseudo label $\bar{y}_i$. The size of $\mathcal{B}$ is $n \times n$. In the initialization, $\mathcal{B}$ is an identity matrix, we initialized it using the pre-defined single-class label y_i.

After obtaining the index of eligible neighbor using nearest neighbor ranking, the pseudo label of the neighbor $\bar{y}_e$ is inquired from $\mathcal{B}$ as,

$$\bar{y}_e = \mathcal{B}[e] \tag{2}$$

where e represents the index of eligible neighbor in X.

To store and inquire up-to-date pseudo labels for all images, PLMB is updated using generated pseudo label $\bar{y}_i$ during each training iteration through,

$$\mathcal{B}[i] \leftarrow \bar{y}_i \tag{3}$$

Thanks to the PLMB, $\bar{y}_e$ can be efficiently inquired to further regress the similarity score s_i according to information of neighbor. Both of obtained multi-class labels $\bar{y}_i$ and $\bar{y}_e$ are used for training the NNCT.

2.3. Overall Loss

The overall loss can be represented as the sum of $\mathcal{L}_C$ and $\mathcal{L}_N$ as,

$$\mathcal{L} = \mathcal{L}_C + \lambda^n \mathcal{L}_N \tag{4}$$

where λ^n is the parameter weighting $\mathcal{L}_C$ and $\mathcal{L}_N$, defaults as 0.5. The $\mathcal{L}_C$ are the loss of similarly s_i and label $\bar{y}_i$, and $\mathcal{L}_N$ are the loss of similarly s_i and label of eligible neighbor $\bar{y}_e$. $\mathcal{L}_C$ and $\mathcal{L}_N$ are computed using Memory-based Multi-label Classification Loss (MMCL) [5] as,

$$\mathcal{L}_C = \sum_{j=1}^{n} \left\| s_i[j] - \bar{y}_i[j] \right\|^2 \tag{5}$$

$$\mathcal{L}_N = \sum_{1}^{k} \sum_{j=1}^{n} \left\| s_i[j] - \bar{y}_e^k[j] \right\|^2 \tag{6}$$

where k is denoted as k-nearest neighbor are selected as the eligible neighbor. The model performances of different k are reported in Section 3.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

Market-1501 (Market) [16] has six cameras and 32,668 person images of 1,501 identities in total. DukeMTMC-reID (Duke) [17, 18] has eight cameras and 36,411 person images of 1,404 identities in total. The CamStyle [20] is used as a data augmentation strategy. Two evaluation metrics are used to measure system performance. The first one is the Cumulative Matching Characteristic (CMC) curves. The CMC represents the probability of top-k ranked gallery samples containing the query identity. The CMCs (%) of rank-1 (R-1), rank-5 (R-5), and rank-10 (R-10) are reported in this paper. The second evaluation metric is the Mean Average Precision (mAP) (%).

3.2. Implementation Details

The experiments are performed on a desktop with an Intel Core i5-6600 3.30-GHz CPU and one NVIDIA GeForce Titan 1080Ti GPU with 11 GB of memory. The experiments are implemented on PyTorch. The training batch size is 64. The ResNet-50 [21] or Osnet [22] are adopted as the backbone network, where Osnet achieves better performances by extracting multi-scale features. We remove the subsequent layers after the pooling-5 layer of ResNet-50 or Osnet and add a batch normalization layer. These two backbone networks are pre-trained on ImageNet [23]. During training, the initial learning rate is 0.1. The learning rate is divided by ten after 40 epochs. The network is trained in an end-to-end fashion by the Stochastic Gradient Descent (SGD).

3.3. Comparison to the state-of-the-arts

We compare our collaborative training strategy NNCT with state-of-the-arts fully unsupervised learning methods in Table 1. The results show that the NNCT clearly outperforms CAMEL, DECAMEL, BUC, DBC, and MPLP+MMCL on both datasets. The baseline model performances are reported as “w/o NNCT” to investigate the necessity of the proposed neighbor flow, which is trained without the neighbor flow. Specifically, by adopting ResNet-50, we observe 4.1% and

Table 1. **Unsupervised person re-ID performance comparison with state-of-the-art methods on Market-1501 and DukeMTMC-ReID Dataset.** Results that surpass all methods are **bold**. "w/o NNCT": Baseline model, trained without the neighbor flow. The k-NNCT represents the model with k-nearest neighbors are selected. The results with underline mean that it exceeds the baseline model "w/o NNCT".

Method	reference	Market					Duke				
		source	R-1	R-5	R-10	mAP	source	R-1	R-5	R-10	mAP
CAMEL [1]	ICCV17	None	54.5	-	-	26.3	-	-	-	-	-
DECAMEL [2]	TPAMI18	None	60.2	76.0	81.1	32.4	-	-	-	-	-
BUC [3]	AAAI19	None	66.2	79.6	84.5	38.3	None	47.4	64.6	68.4	27.5
DBC [4]	BMVC19	None	69.2	83.0	87.8	41.3	None	51.5	64.6	70.1	30.3
MPLP+MMCL [5]	CVPR20	None	80.3	89.4	92.3	45.5	None	65.2	75.9	80.0	40.2
w/o NNCT (ResNet-50)	Proposed	None	80.0	89.4	92.3	44.3	None	63.5	73.7	77.7	37.4
1-NNCT (ResNet-50)		None	82.0	90.0	92.9	48.4	None	64.8	75.7	79.2	40.7
w/o NNCT (Osnet)		None	80.3	89.9	93.0	45.0	None	66.3	77.7	81.1	41.9
1-NNCT (Osnet)		None	**85.2**	**92.3**	**94.3**	**57.6**	None	**70.2**	**80.3**	**83.6**	**47.5**
PTGAN [6]	CVPR18	Duke	38.6	-	66.1		Market	27.4	-	50.7	-
HHL [7]	ECCV18	Duke	62.2	78.8	84.0	31.4	Market	46.9	61.0	66.7	27.2
SAL [8]	TIP20	Duke	65.3	79.7	84.6	38.7	Market	67.6	80.9	84.7	48.5
ATNet [9]	CVPR19	Duke	55.7	73.2	79.4	25.6	Market	45.1	59.5	64.2	24.9
SML [11]	CVPR19	MSMT	67.7	81.9	-	40.0	MSMT	67.1	79.8	-	48.0
ECN [12]	CVPR19	Duke	71.5	87.6	91.6	43.0	Market	63.3	75.8	80.4	40.4
UCDA [13]	ICCV19	Duke	64.3	-	-	34.5	Market	55.4	-	-	36.7
PAST [10]	ICCV19	Duke	78.4	-	-	54.6	Market	72.4	-	-	**54.3**
PDA-Net [14]	ICCV19	Duke	75.2	86.3	90.2	47.6	Market	63.2	77.0	82.5	45.1
SSG [15]	ICCV19	Duke	80.0	90.0	92.4	58.3	Market	73.0	80.6	83.2	53.4
MPLP+MMCL [5]	CVPR20	Duke	84.4	92.8	95.0	60.4	Market	72.4	**82.9**	85.0	51.4
w/o NNCT (ResNet-50)	Proposed	Duke	85.0	92.3	95.0	55.4	Market	70.2	81.0	84.9	49.0
1-NNCT (ResNet-50)		Duke	84.8	92.6	95.0	55.9	Market	71.3	80.8	84.0	49.8
w/o NNCT (Osnet)		Duke	85.7	93.5	95.5	57.1	Market	70.8	81.0	84.7	48.6
1-NNCT (Osnet)		Duke	88.0	**94.3**	**96.3**	65.3	Market	**73.6**	**82.9**	**86.0**	52.7
2-NNCT (Osnet)		Duke	**88.2**	94.1	96.1	**66.3**	Market	73.3	82.6	85.8	54.0
3-NNCT (Osnet)		Duke	87.0	93.6	95.4	65.8	Market	73.1	82.5	85.6	53.0

3.3% mAP drops on Market and Duke, respectively. The results demonstrate that our proposed collaborative training strategy helps model performance by utilizing the neighbor information without any labeled dataset.

The results of our proposed NNCT with UDA-based method follows the same training manner as described in [12]. The UDA-based NNCT transfers the knowledge from the labeled source dataset to the unlabeled target dataset by training the network on both source and target datasets. In Table 1, the proposed NNCT achieves the best performance on Market and Duke. On Market, we obtain rank-1 =88.2%, mAP =66.3%. On Duke, we obtain rank-1 =73.3%, mAP =54.0%. It demonstrates the promising performance of our proposed collaborative training. It is also interesting to observe that, the performance of "w/o NNCT (Osnet)" and "w/o NNCT (ResNet-50)" are close, but the performance of "1-NNCT (Osnet)" significantly surpasses the "1-NNCT (ResNet-50)" with using our proposed NNCT. It is because that the Osnet provides more accurate neighbor information than ResNet-50 by utilizing the multi-scale features in each layer.

Moreover, We test the model using different k and report the result as "1-NNCT", "2-NNCT", and "3-NNCT". "2-NNCT" achieves the best performance in Market and Duke. It is because that selecting more neighbors boosts the performance by providing more assistant information but also easy to harm the performance because of increasing the noise labels.

4. CONCLUSION

This paper introduces a collaborative training strategy NNCT to address the noisy pseudo labels for unsupervised person re-ID. To make training with eligible neighbors possible, we construct a PLMB to store and inquire the up-to-date labels of neighbors. Through the experiments, the effectiveness of our proposed collaborative training is demonstrated. The proposed NNCT surpasses state-of-the-arts in fully unsupervised learning-based methods and UDA-based methods.

5. REFERENCES

[1] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 994-1002.

[2] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person reidentification by deep asymmetric metric embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 956-973, Apr. 2020.

[3] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A Bottom-up Clustering Approach to Unsupervised Person Re-Identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 8738-8745.

[4] G. Ding, S. Khan, Q. Yin, and Z. Tang. "Dispersion based clustering for unsupervised person re-identification," in *BMVC*, 2019.

[5] D. Wang and S. Zhang, "Unsupervised Person Re-Identification via Multi-Label Classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10978-10987.

[6] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person reidentification. In CVPR, 2018

[7] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In ECCV, 2018.

[8] K. Jiang, T. Zhang, Y. Zhang, F. Wu and Y. Rui, "Self-Supervised Agent Learning for Unsupervised Cross-Domain Person Re-Identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 8549-8560, 2020.

[9] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, Adaptive transfer network for cross-domain person re-identification, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 72027211.

[10] X. Zhang, J. Cao, C. Shen and M. You, "Self-Training With Progressive Augmentation for Unsupervised Cross-Domain Person Re-Identification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8221-8230.

[11] H. Yu, W. Zheng, A. Wu, X. Guo, S. Gong and J. Lai, "Unsupervised Person Re-identification by Soft Multilabel Learning," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2143-2152.

[12] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 598-607.

[13] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In ICCV, 2019.

[14] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In ICCV, 2019.

[15] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S. Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In ICCV, 2019.

[16] L. Zheng, S. Wang, L. Shen, L. Tian, J. Bu, and Q. Tian, "Person Re-identification Meets Image Search," *arXiv preprint arXiv:1502.02171.*

[17] Z. Zheng and L. Zheng and Y. Yang, "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3774-3782.

[18] E. Ristani, F. Solera, R. Zou, R. Cucchiara and C. Tomasi, "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking," in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[19] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79-88

[20] Z. Zhong, L. Zheng, Z. Zheng, S. Li and Y. Yang, "Camera style adaptation for person reidentification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5157-5166.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770778, 2016

[22] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, "Omni-Scale Feature Learning for Person Re-Identification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3701-3711.

[23] J. Deng, W. Dong, R. Socher, L. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.