PERSON RE-IDENTIFICATION BY RANKING ENSEMBLE REPRESENTATIONS

Guile Wu⋆ *Xiatian Zhu*† *Shaogang Gong*⋆

⋆ Queen Mary University of London, UK
† Vision Semantics Limited, UK

ABSTRACT

Existing deep learning algorithms for person re-identification (re-id) typically rely on single-sample classification or pairwise matching constraints. This indicates a breach of deployment due to ignoring the probe-specific matching information against the gallery set encoded in ranking lists. In this work, we address this problem by exploring the idea of RANkinG Ensembles (RANGE) that learns such information from the ranking lists. Specifically, given an off-the-self deep re-id feature representation model, we construct per-probe ranking lists and exploit them to learn inter ranking ensemble representation. To mitigate the harm of inevitable false gallery positives, we further introduce a complementary intra ranking ensemble representation. Extensive experiments show that both supervised and unsupervised re-id benefit from the proposed RANGE method on four challenging benchmarks: MSMT17, Market-1501, DukeMTMC-ReID, and CUHK03.

***Index Terms*—** Person re-identification, ranking list.

1. INTRODUCTION

Person re-identification (re-id) aims to match people across non-overlapping camera views distributed over different locations [1]. Existing deep learning re-id algorithms mostly leverage the classification and matching pair constraints [2, 1, 3]. This is limited in understanding ranking list data as used in deployment, e.g. ignoring useful latent ranking information specific to the probe sample therefore less generalisable.

There are a few existing attempts that exploit the ranking content information for re-id by learning to rank [4] and post-rank [5]. However, these methods are either based on hand-crafted features or assume good performance of trained re-id models, without the advantages of end-to-end deep representation learning for exploiting the potentially informative ranking context. Specifically, for a given probe image, the top-ranked gallery candidates, either true or false matches, resemble similar view variations as the probe image. Such contextual information may be useful for re-id. Importantly, this can generally benefit both existing supervised and unsupervised re-id methods due to no need for extra labelling.

In this work, we explore the largely ignored ranking context information for supervised and unsupervised re-id in deep learning. To this end, we propose a RANkinG Ensembles (RANGE) approach capable of generally benefiting the existing off-the-self re-id models (see Fig. 1). Our contributions are two-fold: **(I)** We propose the idea of exploiting the ranking context encoded in the ranking lists for improved person re-id in both supervised and unsupervised settings. To our best knowledge, this is the first attempt of using the ranking context information in end-to-end deep learning. **(II)** We present a novel deep learning architecture for deriving the ranking ensemble representation. This is achieved by formulating per-probe rank lists to learn the plausible visual variations among top ranks (*inter ranking ensemble*) and the intra body-part visual correlations (*intra ranking ensemble*) simultaneously.

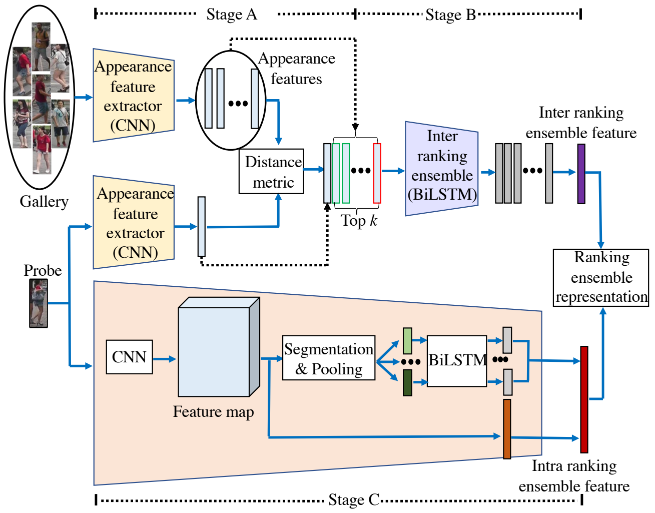
Stage A
Stage B
Gallery
Appearance feature extractor (CNN)
Appearance features
Distance metric
Top k
Inter ranking ensemble (BiLSTM)
Inter ranking ensemble feature
Probe
Appearance feature extractor (CNN)
Ranking ensemble representation
CNN
Feature map
Segmentation & Pooling
BiLSTM
Intra ranking ensemble feature
Stage C

Fig. 1. An overview of the proposed RANGE architecture. **Stage A**: To extract the visual appearance features and generate the initial rank lists. **Stage B**: To extract inter ranking ensemble features. **Stage C**: To extract the intra ranking ensemble features.

Extensive comparative experiments show that the proposed RANGE method improves the state-of-the-art supervised and unsupervised person re-id models on four large scale benchmarks, including MSMT17 [6], Market-1501 [7], DukeMTMC-ReID [8], and CUHK03 [9].

2. LEARNING RANKING CONTEXT

Problem statement We assume an off-the-shelf deep re-id CNN model $\boldsymbol{\theta}_0$ trained by an existing learning algorithm [1, 2] for charactering person appearance. The initial model can be used to generate a ranking list per probe on the training set along with the Euclidean distance metric. Let p_i be an arbitrary probe person image and $\{x_l\}_{l=1}^{k}$ the corresponding top-k re-id matches from the gallery set. This allows to construct a training data with ranking lists.

RANGE formulation We aim to discover the ranking context information underlying to the gallery collection. This is based on the RANGE learning from $\{p_i, \{x_t\}_{t=1}^{k}\}$. The ranking lists are ordered data structures which accommodate extra sequential information, i.e. the contextual information of the gallery specific to the probe. In light of such understanding, we consider a sequential learning strategy. We exploit a recurrent neural network model, BiLSTM [10] in particular. This aims to learn the *inter ranking ensemble* representations per probe (Sec 2.1). Meanwhile, we construct a unified CNN-BiLSTM architecture to learn the *intra ranking ensemble* representation for improving the corresponding inter ranking ensemble features (Sec 2.2). Inter and intra ranking ensemble representations can be correlated and aggregated end-to-end to jointly optimise the model generalisation capability. Importantly, both supervised and unsupervised learning methods allow to benefit from the RANGE method (Sec 2.3).

2.1. Inter Ranking Ensemble Representations

To learn the inter ranking ensemble feature of p_i, we construct an augmented ranking lists by setting p_i as the first element, *i.e.* $\{x_t\}_{t=0}^{k}$ with $x_0 = p_i$. We denote $\{v_t^a\}_{t=0}^{k}$ the corresponding appearance feature vectors extracted by the model $\boldsymbol{\theta}_0$, we formulate them as an input sequence to a BiLSTM model as:

$$\begin{cases} h_t = f(w_1 v_t^a + w_2 h_{t-1}) \\ h'_t = f(w_3 v_t^a + w_5 h'_{t+1}) \\ y_t = g(w_4 h_t + w_6 h'_t) \end{cases} \tag{1}$$

where $\{w_j\}_{j=1}^{6}$ are the shared weights between each input unit, f is the hidden layer activation function, g is the output layer activation function, and h_t and h'_t are the forward and backward hidden states, respectively. y_t denotes the output feature representation.

Remarks The forward and backward hidden states effectively encapsulate the ranking order information for a given probe. The output y_t strongly emphasises the correlation and discriminative selections among the input units (elements in the list). The BiLSTM network output y_t contains the output features h_t from the last layer of the LSTM, so the output features are formulated as latent *inter ranking ensemble* feature vectors $\{v_l^r\}_{l=0}^{k}$, *i.e.* $[v_0^r, v_1^r, ..., v_k^r] = [y_0, y_1, ..., y_k]$.

To obtain the inter ranking ensemble representation v_i^R of x_i, we use the average pooling strategy as:

$$v_i^R = \frac{1}{k+1}\sum_{l=0}^{k} v_{l,i}^r \tag{2}$$

Objective loss function To train this BiLSTM network with labelled training data in the supervised re-id, we employ the hardest positive $v_{i,p}^R$ and negative $v_{i,n}^R$ of x_i in the feature space v^R. We adopt the triplet ranking loss function [11] as:

$$L_r = \max\Big(0, \alpha + d(v_i^R, v_{i,p}^R) - d(v_i^R, v_{i,n}^R)\Big) \tag{3}$$

where α denotes a margin and d is the Euclidean distance.

2.2. Intra Ranking Ensemble Representations

While the inter ranking ensemble can effectively learn the variations and correlation between a probe and the gallery, there may be more false positives than true positives. This is likely to contaminate the inter ranking ensemble representations. To overcome this problem, we further develop the intra ranking ensemble representation so that the initial ranking lists can be improved.

Given x_i, the feature map obtained from the CNN model $\boldsymbol{\theta}_0$ is horizontally divided into m stripes to compute feature vectors by average pooling. Unlike existing part-based re-id methods [12, 13, 14], we progressively input these feature vectors into another BiLSTM model to learn bi-directional intra correlations using Eq (1). With m output feature vectors from the BiLSTM, *i.e.* $[v_1^p, ..., v_m^p] = [y_1^p, ..., y_m^p]$, the intra ranking ensemble features of x_i is obtained as the concatenation of m feature vectors, *i.e.* $v_i^P = [v_0^p \oplus v_1^p \oplus ... \oplus v_m^p]_i$, where v_0^p and $\oplus$ denote the holistic visual feature vector and vector concatenation, respectively. In the intra ranking ensemble, the CNN and the BiLSTM are jointly trained as a CNN-BiLSTM model. The objective is to optimise the softmax cross-entropy loss function:

$$L_p = -\frac{1}{K}\sum_{i=1}^{K} y_i \log \frac{\exp(W_c[v_0^p \oplus ... \oplus v_m^p]_i)}{\sum_{q=1}^{Q} \exp(W_q[v_0^p \oplus ... \oplus v_m^p]_i)} \tag{4}$$

where K and Q denote the mini-batch size and the total number of person identity ($c \in Q$), y_i is the ground truth distribution, and W_c and W_q are to-be-learned model weights.

2.3. Supervised and Unsupervised RANGE

Supervised RANGE In supervised re-id, we have n training person images $X = \{x_1, ..., x_n\}$ of $N_{id} = \{1, ..., n_{id}\}$ different people together with their corresponding identity labels $Y = \{y_1, ..., y_n\}$ ($y_i \in N_{id}$). We first train a CNN re-id model (*e.g.* ResNet-50 with the Cross Entropy loss) for person appearance feature vectors extraction $V^a = \{v_i^a\}_{i=1}^{n}$. Then, we compute the pairwise Euclidean distances D^a for ranking p_i against the gallery x_g in X.

Inter and intra ranking ensemble representations To compute the inter ranking ensemble representation v_i^R, we start by generating the i-th candidate rank list S_i^c, ranking v_i^a in the first position and the other candidates in an ascending order of D^a. Top $k+1$ candidates in this list are then used as inputs to the BiLSTM for inferring the inter ranking ensemble vectors $\{v_l^r\}_{l=0}^k$ from the gallery. We finally obtain the v_i^R by Eq (2). Meanwhile, we compute the intra ranking ensemble v_i^P of x_i.

Re-id deployment Given a test probe, we utilise the inter and intra ranking ensembles to compute the pairwise Euclidean distances D^r and D^p, respectively. We then aggregate the two types of distance score for re-id as

$$D^* = D^r + \beta D^p \tag{5}$$

where β is the fusion weight. When $\beta = 0$, only the inter ranking ensemble is used for re-id. We denote this score fusion as ***RANGE-s***.

Alternatively, we can concatenate v_i^R and v_i^P as $v_i^* = [v_i^R \oplus \gamma v_i^P]$ of x_i, with γ is the concatenation weight. The final re-id distance D^* is then computed with v_i^*. We call this feature fusion as ***RANGE-f***.

Unsupervised RANGE We explore the benefits of RANGE for unsupervised cross-domain re-id. In this case, the initial model $\boldsymbol{\theta}_0$ is typically weak for performing re-id in the unseen target domain due to the potentially significant domain discrepancy, resulting in more false matches in the top ranks.

To this end, we adapt a CNN model pre-trained in a labelled source domain to an unlabelled target domain for more accurately estimating v_i^* and D^*. This is achieved by adaptive clustering and fine-tuning [15, 16], with the aim of improving the ranking ensemble representations. More specifically, assume there are a total of N unlabelled training data, after the performing adaptive clustering, n samples are then clustered into a number of j clusters, where $j \ll n < N$. We label these clustered samples with cluster labels with the remaining discarded. Then, we fine-tune the pre-trained model to optimise the ranking ensemble representations with the triplet loss (one-pass solution instead of iterative clustering for computation reduction). The updated model is then used to revise v_i^* and D^*. We conduct person re-id deployment as above.

3. EXPERIMENTS

Datasets To evaluate the benefits of RANGE for both supervised and unsupervised re-id, we selected four large-scale benchmarks (*i.e.* MSMT17 [6], Market-1501 [7], DukeMTMC-ReID [8], and CUHK03 [9]). We used the standard data split setting (Table 1) and the *single query* test.

Evaluation protocol We adopted the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as the performance evaluation metrics.

Table 1. Statistics of four re-id datasets and test settings.

Benchmark	Image	ID	Train	Test
MSMT17	126,441	4,101	1041	3,060
Market-1501	32,668	1,501	751	750
DukeMTMC-ReID	36,411	1,404	702	702
CUHK03	14,097	1,467	767	700

Table 2. Comparisons to the state-of-the-art supervised re-id methods on Market-1501 and DukeMTMC-ReID. The top 1 and 2 results are in **red** and **blue**.

Methods	Reference	Market		DukeMTMC	
		mAP	R1	mAP	R1
Backbone	-	70.5	87.6	59.2	76.9
RANGE-f	Ours	**81.9**	**91.0**	**69.7**	**81.3**
RANGE-s	Ours	**81.0**	90.7	**70.1**	**81.9**
SVDNet [19]	ICCV17	62.1	82.3	56.8	76.7
PDC [12]	ICCV17	63.4	84.1	-	-
DPFL [20]	ICCVW17	72.6	88.6	60.6	79.2
DaF [21]	BMVC17	72.4	82.3	-	-
Reranking [5]	CVPR17	63.6	77.1	-	-
JLML [3]	IJCAI17	65.5	85.1	56.4	73.3
CRAFT [22]	TPAMI18	42.3	68.7	-	-
BraidNet [23]	CVPR18	69.5	83.7	59.5	76.4
DML [24]	CVPR18	68.8	87.7	-	-
MLFN [2]	CVPR18	74.3	90.0	62.8	81.0
HAN [1]	CVPR18	75.7	**91.2**	63.8	80.5

Implementation details We implemented the RANGE model using Pytorch. We employed the ResNet-50 [17] pre-trained on ImageNet as the backbone network. Other off-the-self re-id networks, such a MLFN [2], and HA-CNN [1], can be readily used. We adopted the SGD for optimisation with the initial learning rate as 10^{-2} (decayed to 10^{-3} after 20 training epochs). After the model is trained, 2048-D initial feature vectors are extracted from the last convolutional layer. For the inter ranking ensemble representations, we set the number of forward-backward recurrent layer to 1 and set the learning rate to 10^{-4}. The input sequence length for BiLSTM is $k+1 = 5$ and the margin $\alpha = 0.8$. We set $\beta = 0.25$ and $\gamma = 0.15$. The output feature vector is 1024-D. For the intra ranking ensemble, we set $m = 4$ and the output feature dimension of each stipes is 256-D. The final output feature vector is 2048-D. We adopted DBSCAN [18] for adaptive clustering and set $\beta = 0.4$ in unsupervised cross-domain re-id.

3.1. Comparison to the State-of-the-Art Methods

Supervised re-id Table 2 and 3 compare the supervised re-id performance of the proposed RANGE method with state-of-the-art methods. We make the following observations: **(I)** When using ResNet-50 as the baseline model, re-id performance can be significantly improved from the proposed

Table 3. Comparisons to the state-of-the-art supervised re-id methods on MSMT17 and CUHK03.

Methods	Reference	MSMT17		CUHK03			
				Labelled		Detected	
		mAP	R1	mAP	R1	mAP	R1
Backbone	-	31.0	61.7	44.9	45.8	43.7	45.7
RANGE-f	Ours	**41.0**	**68.7**	**57.0**	**52.9**	**56.2**	**53.0**
RANGE-s	Ours	**41.5**	**68.6**	**54.8**	51.6	**54.2**	51.8
DaF [21]	BMVC17	-	-	31.5	27.5	30.0	26.4
SVDNet [19]	ICCV17	-	-	37.8	40.9	37.3	41.5
PDC [12]	ICCV17	29.7	58.0	-	-	-	-
DPFL [20]	ICCVW17	-	-	40.5	43.0	37.0	40.7
MLFN [2]	CVPR18	37.0	66.3	49.2	**54.7**	47.8	**52.8**
HAN [1]	CVPR18	35.6	63.5	41.0	44.4	38.6	41.7

Table 4. Comparisons to the state-of-the-art unsupervised cross-domain re-id. **D2M**: DukeMTMC (source) ⇒ Market (target). **M2D**: Market (source) ⇒ DukeMTMC (target).

Methods	Reference	D2M		M2D	
		mAP	R1	mAP	R1
Backbone	-	16.3	40.4	11.9	23.1
RANGE	Ours	**32.5**	**58.5**	**21.8**	**34.6**
LOMO [25]	CVPR15	8.0	27.2	4.8	12.3
BoW [7]	ICCV15	14.8	35.8	8.3	17.1
RKSL [26]	ICIP16	11.0	34.0	-	-
CAMEL [27]	ICCV17	26.3	54.5	-	-
PUL [15]	TOMM18	20.1	44.7	16.4	30.4
PTGAN [6]	CVPR18	-	38.6	-	27.4
TJ-AIDL [28]	CVPR18	**26.5**	**58.2**	**23.0**	**44.3**

RANGE method, thanks to the exploitation of ranking context cues. **(II)** The RANGE outperforms or is on par with existing re-id methods on all test benchmarks. For example, on Market-1501 RANGE-f and RANGE-s significantly improve the state-of-the-art mAP by 6.2% (81.9%-75.7%) and 5.3% (81.0%-75.7%) respectively, although the baseline is significantly inferior to the state-of-the-arts. On the largest test MSMT17, RANG-f improves the state-of-the-art method (*i.e.* MLFN) by 2.3% and 4.5% in terms of rank-1 accuracy and mAP. This suggests the importance of considering the gallery contextual information in re-id similar as other orthogonal perspectives like attention modelling in HAN [1]. **(III)** Overall, RANGE brings more significant improvements on mAP, which indicates that RANGE benefits the model to retrieve more related candidates as both probe and gallery representations are exploited based on ranking context in the gallery.

Unsupervised cross-domain re-id Table 4 reports the unsupervised cross-domain re-id performance of RANGE in comparison to existing alternatives (the results of RANGE-s and RANGE-f are close, so we report RANGE here). We have similar observations: RANGE improves the baseline ResNet-50 significantly, and yields similar performance as the best competitor TJ-AIDL. This suggests the benefits of our model when no labels are available for the target domain.

Table 5. RANGE component analysis on Market-1501 and DukeMTMC-ReID. Setting: supervised re-id.

Metric (%)	Market		DukeMTMC	
	mAP	R1	mAP	R1
Backbone	70.5	87.6	59.2	76.8
Inter ranking ensemble	81.1	88.9	65.5	78.6
Intra ranking ensemble	60.3	83.4	52.9	74.3
RANGE	**81.9**	**91.0**	**69.7**	**81.3**

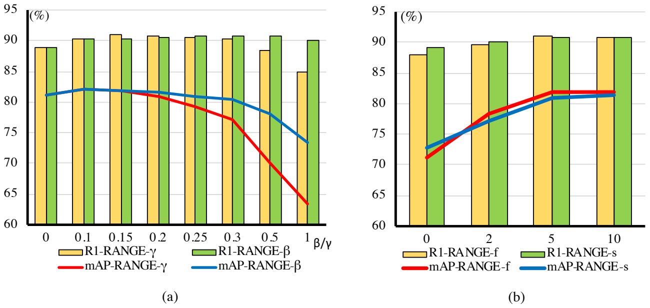

Fig. 2. Analyses about (a) fusion weight and (b) rank length on Market-1501. Setting: supervised re-id.

3.2. Component and Parameter Analyses

Component analysis Table 5 shows that inter ranking ensemble clearly improves the re-id performance of the backbone, while intra ranking ensemble gives extra benefits to further refine rank lists.

Model parameter analysis Fig. 2(a) shows that the fusion weights affect the re-id performance. This is because inter ranking ensemble with gallery contextual information gives more importance. As shown in Fig. 2(b), the top 5 ranks suffice to capture the gallery contextual information.

4. CONCLUSION

In this work, we presented a RANkinG Ensembles (RANGE) approach to exploiting the ranking context information of the gallery population in re-id. Starting with any trained re-id model, the RANGE constructs per-probe ranking lists on the same training data for discovering additional discriminative re-id information. The proposed method can benefit both supervised and unsupervised re-id learning algorithms in a unified formulation. Extensive experiments on four large scale benchmarks with varying challenging covariates have demonstrates the benefits and advantages of the proposed RANGE method in enhancing the person re-id matching accuracy. We also conducted in-depth component analysis to give insights on the superiority of our RANGE design.

5. REFERENCES

[1] Wei Li, Xiatian Zhu, and Shaogang Gong, "Harmonious attention network for person re-identification," in *CVPR*, 2018.

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang, "Multi-level factorisation net for person re-identification," in *CVPR*, 2018.

[3] Wei Li, Xiatian Zhu, and Shaogang Gong, "Person re-identification by deep joint learning of multi-loss classification," in *IJCAI*, 2017.

[4] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, 2015.

[5] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017.

[6] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, "Person transfer gan to bridge domain gap for person re-identification," in *CVPR*, 2018.

[7] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[8] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017.

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[10] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *TSP*, 1997.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[12] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017.

[13] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *ECCV*, 2018.

[14] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[15] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *TOMM*, 2018.

[16] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *arXiv preprint arXiv:1807.11334*, 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[18] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *AAAI*, 1996.

[19] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017.

[20] Yanbei Chen, Xiatian Zhu, and Shaogang Gong, "Person re-identification by deep learning multi-scale representations," in *ICCVW*, 2017.

[21] Rui Yu, Zhichao Zhou, Song Bai, and Xiang Bai, "Divide and fuse: A re-ranking approach for person re-identification," in *BMVC*, 2017.

[22] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai, "Person re-identification by camera correlation aware feature augmentation," *TPAMI*, 2018.

[23] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang, "Person re-identification with cascaded pairwise convolutions," in *CVPR*, 2018.

[24] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu, "Deep mutual learning," in *CVPR*, 2018.

[25] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[26] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong, "Towards unsupervised open-set person re-identification," in *ICIP*, 2016.

[27] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *ICCV*, 2017.

[28] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *CVPR*, 2018.