FEATURE CONSISTENCY FOR DOMAIN GENERALIZED PERSON RE-IDENTIFICATION

[1]*Ci-Siang Lin (林棋祥)*, [1]*Yuan-Chia Cheng (鄭元嘉)*,

[1]*Yu-Chiang Frank Wang (王鈺強)*, [2]*Chiou-Shann Fuh (傅楸善)*

[1]Graduate Institute of Communication Engineering,
National Taiwan University, Taiwan

[2]Department of Computer Science and Information Engineering
National Taiwan University, Taiwan

E-mail: {d08942011, r08942154, ycwang}@ntu.edu.tw, fuh@csie.ntu.edu.tw

ABSTRACT

Given a query image containing a person of interest, re-ID aims at matching gallery images with the same identity across different camera views. Existing works typically require collection of a large amount of labeled image data, which is not practical for real-world applications due to limited resources and privacy issues. Without observing any label information from target domain data during training, unsupervised domain adaptation (UDA) approaches extract and adapt useful information from source to target domain. Nevertheless, data collection and model updating are still required. In this paper, we target at an even more challenging and practical setting, domain generalized (DG) person re-ID. That is, while a number of labeled source-domain datasets are available, we do not have access to any target-domain training data. In order to learn domain-invariant features without knowing the target domain of interest, we present an episodic learning scheme which advances meta learning strategies to exploit the observed source-domain labeled data. The learned features would exhibit sufficient domain-invariant properties while not overfitting the source-domain data or ID labels. Our experiments on four benchmark datasets confirm the superiority of our method over the state-of-the-arts.

Index Terms— Domain Generalization, Person Re-Identification, Deep Learning

1. INTRODUCTION

Person re-identification (re-ID) [1] has been among active research topics in computer vision due to its wide applications to person tracking [2], video surveillance systems [3] and smart cities. Given a query image containing a person of interest, re-ID aims at matching gallery images with the same identity across different camera views. A number of works [4, 5, 6, 7, 8] have been proposed to recognize the identical identity suffering from the variation of viewpoints, postures, occlusions or background clutters. However, most of these approaches require collection of a large amount of labeled image data from the scenes of interest, which is not practical for real-world applications due to limited resources and privacy issues.

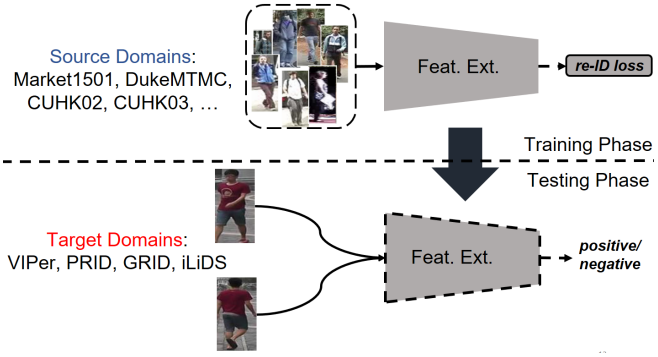

Fig. 1. Domain generailized person re-identification.

Since data labeling is quite time-consuming, one popular solution is to utilize another labeled source domain data to tackle re-ID at the domain of interest. However, deep learning methods often suffer from a significant performance drop when directly applied to a new dataset or environment due to different data distributions or statistics from training data. To tackle the problem, unsupervised domain adaptation (UDA) approaches [9, 10, 11] have been proposed for cross-dataset re-ID. Without observing any label information from target domain data during training, these UDA approaches extract and adapt useful information from source to target domain. Nevertheless, data collection and model updating are still required.

In this paper, we target at an even more challenging and practical setting, domain generalized (DG) person re-ID. As illustrated in Figure 1, while a number of labeled source-domain datasets are available, we do not have access to the any target-domain data during training, whether labeled or not. Different from cross-dataset re-ID, the goal of DG per-

son re-ID is to improve the generalization and robustness of the learned model with data from multiple source domains only. While a number of works [12, 13, 14, 15] have been proposed for solving DG classification tasks, the label spaces are shared across domains. This is very different from the setting for re-ID. For any two different re-ID datasets, their person identities are disjoint, i.e., any two images from different domain represent different people. Hence, we could not assume that person identities across domains remain the same. As a result, existing DG classification methods cannot be easily applied to tackle the re-ID tasks. Few existing works [16, 17] have attempted DG person re-ID. DualNorm [16] takes advantage of Batch Normalization (BN) [18] and Instance Normalization (IN) [19] to alleviate the domain difference. While DIMN [17] learns the mapping between person images and the ID classifiers via meta-learning, their model do not learn domain invariant representations, and thus cannot generalize well to unseen target-domain data.

To address the challenging DG person re-ID tasks without observing target-domain training data, we design an episodic scheme which advances meta learning strategies to exploit the observed source-domain labeled data. With proposed cross-consistency, the learned features would exhibit sufficient domain-invariant properties while not overfitting the source-domain data or ID labels. This arms us to apply the learned model to any target domains of interest in no need of extra data collection and model updating. Compared to prior works, our experiments confirm that our model is able to improve the performance and thus is practically favorable

We now highlight the contributions of our work below:

- To best of our knowledge, we are among the first to derive domain invariant yet identity-discriminative features for re-ID under such an unique and challenging DG person re-ID setting.
- Without the access to any labeled or unlabeled data from the domain of interest, our model learns a domain invariant latent space and is able to be directly applied to novel domain of interest in no need of any data collection and model updating.
- Experimental results on four benchmark datasets quantitatively verifies that our approach performs preferably well against the state-of-the-art methods.

2. PROPOSED METHOD

2.1. Notations and Problem Formulation

We first define the notations to be used in this paper. For domain generalized person re-ID, assume that we have the access to data from N_d labeled source domains (datasets), i.e., source-domain data $D_S = \{D_i\}_{i=1}^{N_d}$, and the ith source domain D_i contains a set of N_i images $X_i = \{x_u^i\}_{u=1}^{N_i}$ with the associate label set $Y_i = \{y_u^i\}_{u=1}^{N_i}$, where $x_u^i \in R^{H\times W\times 3}$ and $y_u^i \in R$ denote the uth image and its corresponding identity label from the ith source domain D_i, respectively. Note that for any pair of source domains D_i and D_j, we consider their ID labels are disjoint. This unique property makes domain generalized person re-ID more challenging than traditional domain adaptation and generalization problems, where all source domains share the same label space. Thus, previous methods for domain generalization cannot be easily applied to solve the issue.

To achieve DG re-ID, we present an end-to-end meta learning framework, as illustrated in Fig. 2. Our framework aims at learning domain-invariant features $\boldsymbol{v}$ by learning feature and re-ID encoders F and E, based on feature extractors pretrained on each source domain. During each episode, three domains D_i, D_j and D_k are randomly sampled from the source domain datasets. Feature extractors F_j and F_k pretrained on D_j and D_k are used to extract the domain-biased features $\tilde{\boldsymbol{v}}^{i\to j}$ and $\tilde{\boldsymbol{v}}^{i\to k}$ for images x^i from domain D_i, respectively. With the proposed meta learning scheme, our encoder E would derive domain-invariant yet identity-discriminative features for re-ID purposes The details of our proposed learning framework will be discussed in the following sub-sections.

2.2. Meta Learning for Domain-Invariant Representation

We now detail how we advance meta learning strategies for deriving domain-invariant features, without observing the target-domain data during training. Our proposed framework starts with pretrained domain-specific feature extractor F_i using labeled data in each source domain D_i (e.g., using Triplet-Loss [4] as most re-ID works do). Thus, a total of N_d pretrained domain-specific feature extractors are obtained and will later be utilized in our meta learning framework.

As shown in Fig 2, our goal is to learn a domain-invariant feature extractor F, followed by a domain generalized encoder E, for deriving DG re-ID. To accomplish this, we present an episodic learning scheme that utilizes the pretrained F_i for learning the above network modules. To be more precise, in each episode during training, we randomly select data from three source domains D_i, D_j and D_k. For an input anchor image x_a^i from domain D_i, we particularly apply the pretrained domain-specific feature extractors F_j and F_k to output the associated domain-biased features $\tilde{\boldsymbol{v}}_a^{i\to j}$ and $\tilde{\boldsymbol{v}}_a^{i\to k}$. In other words, we have $\tilde{\boldsymbol{v}}_a^{i\to j} = F_j(x_a^i)$ and $\tilde{\boldsymbol{v}}_a^{i\to k} = F_k(x_a^i)$. Since $\tilde{\boldsymbol{v}}_a^{i\to j}$ and $\tilde{\boldsymbol{v}}_a^{i\to k}$ are both derived from the same image x_a^i, we require the domain generalized encoder E to output the final domain-invariant features $\boldsymbol{v}_a^{i\to j}$ and $\boldsymbol{v}_a^{i\to k}$. To enforce E to preserve the domain-invariant yet re-ID preserved information from $\boldsymbol{v}_a^{i\to j}$ and $\boldsymbol{v}_a^{i\to k}$, we propose to calculate the consistency loss $\mathcal{L}_{consis}$ on the above feature pair:

$$\mathcal{L}_{\text{consis}} = E_{x_a^i\sim X_i}\|\boldsymbol{v}_a^{i\to j} - \boldsymbol{v}_a^{i\to k}\|_2. \quad (1)$$

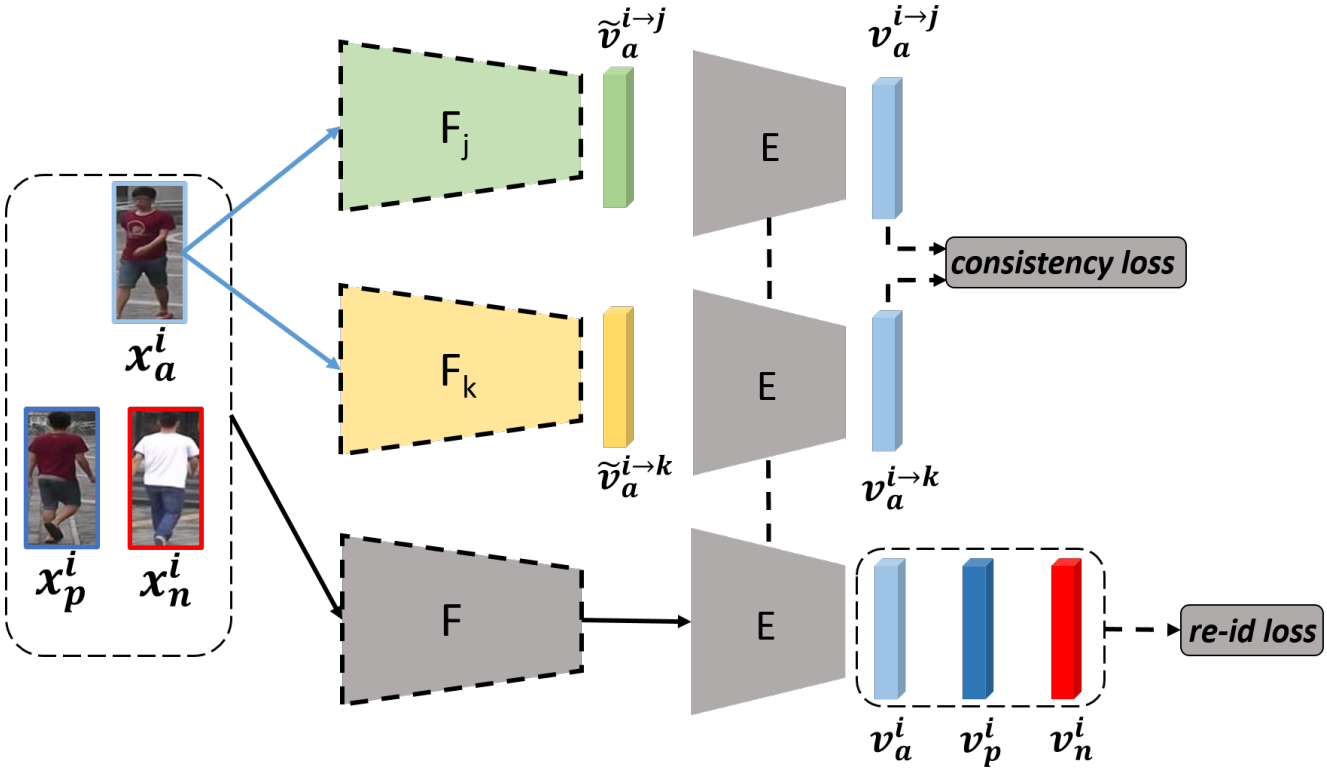

Fig. 2. Overview of our proposed framework. During training, triplet images (i.e., anchor x_a^i, and its positive x_p^i and negative neighbor x_n^i) from source domain D_i are forwarded through feature extractors F_j, F_k pretrained on domains D_j and D_k, also through a global feature extractor F. The features derived by pretrained extractors are viewed as domain-biased features $\tilde{v}$, and the global encoder E is expected to observe the resulting consistency loss $\mathcal{L}_{consis}$. The features derived by F are forwarded through E to guide triplet loss $\mathcal{L}_{tri}$. During testing, only F and E are needed to extract as domain-invariant features v for performing re-ID.

Note that the pretrained feature extractors F_j and F_k are fixed, which would not be updated by back-propagated gradients calculated from $\mathcal{L}_{consis}$. This is the key technique for each domain-specific feature extractors to preserve its own domain-specific properties, while allowing the domain generalized encoder E to extract domain-invariant features during the episodic learning process. It is worth repeating that, during learning of domain-invariant features, no target-domain data are observed.

We also note that, we choose not to apply adversarial learning techniques (e.g., DANN [20]) for deriving domain-invariant features. This is because that, the person identities are disjoint across source domains. If one applies adversarial learning or similar domain adaptation techniques for eliminating the domain differences, it is likely that the person ID or pose information is also confused by such learning strategy, which is not desirable for re-ID tasks.

2.3. Domain Generalized Person Re-ID

To have the encoder E describe global identity features, we additionally train a global (or domain-invariant) feature extractor F to cooperate with the encoder E. To utilize label information observed from source-domain data for learning domain-invariant features for re-ID, we adopt the triplet loss on the derived domain invariant space (i.e., feature space output by domain generalized encoder E). For each domain invariant feature v_a^i derived from the input anchor image x_a^i, a triplet tuple is composed of v_p^i with the same identity label as x_a^i and v_n^i with different identity label as x_a^i. Then, the distances d_p and d_n for such positive and negative pairs are defined as:

$$d_{\mathrm{p}} = \|\boldsymbol{v}_a^i - \boldsymbol{v}_p^i\|_2, \tag{2}$$

$$d_{\mathrm{n}} = \|\boldsymbol{v}_a^i - \boldsymbol{v}_n^i\|_2, \tag{3}$$

where $\boldsymbol{v}_a^i$, $\boldsymbol{v}_p^{i]}$ and $\boldsymbol{v}_n^i$ denote the domain-invariant latent features of x_a^i, x_p^i and x_n^i derived from the encoder E. With the above definitions, we have the domain generalized triplet

loss, $\mathcal{L}_{tri}$, which is calculated as:

$$\mathcal{L}_{tri} = E_{(x_a^i, y_a^i)\sim(X_i,Y_i)} \max(0, m + d_{\mathrm{p}} - d_{\mathrm{n}}), \quad (4)$$

where $m > 0$ is the margin enforcing the separation between positive and negative image pairs.

Take computational feasibility into consideration, we follow [4] and select only the hardest positive and negative samples to calculate the triplet loss. By minimizing $\mathcal{L}_{tri}$, capture the identity information to extract domain invariant but identity-discriminative features v_a^i . It is worth repeating that, the pretrained domain specific feature extractors F_j and F_k will not be updated by this loss.

With the above meta learning framework together with the introduced losses, the total loss $\mathcal{L}_{total}$ of our model is

$$\mathcal{L}_{total} = L_{tri} + \lambda \cdot L_{consis} \quad (5)$$

where λ is the hyperparameter. To perform person re-ID on the unseen domain in the testing phase, the query and gallery images are forwarded through feature extractor F and the encoder E to derive domain-invariant re-ID features, which are applied for matching query/gallery images via nearest neighbor search in Euclidean distances.

3. EXPERIMENTS

3.1. Datasets and Experimental Settings

To evaluate our proposed method, following [17] we use existing large-scale re-ID datasts as the source domains and test the performance on several target datasets which are not observed during training. To be specific, the source domains include Market1501 [21], DukeMTMC-reID [22], CUHK02 [23] and CUHK03 [24], with a total of 6596 identities and 87191 images. The target datasets include GRID [25], i-LIDS [26], PRID [27] and VIPeR [28]. We follow the single-shot setting with the number of probe/gallery images set as: GRID: 125/900; i-LIDS: 60/60; PRID 100/649; VIPeR 316/316 respectively, and the average rank-1 accuracy over 10 random splits is reported based on the standard evaluation protocol and the cumulative matching curve (CMC). To be clear, we reproduce all methods and performances due to the lack of legal access to certain datasets used in original papers.

3.2. Implementation Details

We implement our method using PyTorch. We use the backbone model proposed in [16] as our global feature extractor F and use ResNet-50 [29] pretrained on ImageNet for domain specific modules F_i. The global encoder E consists of two fully-connected (FC) layers with BatchNorm [18] layers while the global classifier C is a single FC layer. We resize all the input images to $256 \times 128 \times 3$ (denoting height, width and channel, respectively). Random clipping and cropping are adapted for data augmentation. The margin m is set as 0.3 and we fix λ_{consis} as 0.01. We train our model for 150 epochs with the SGD optimizer. The initial learning rate is set as 0.01 and is decreased to 0.001 at 100 epochs. Label smoothing is used to prevent overfitting.

Table 1. Performance comparisons of domain generalization based methods in terms of averaged Rank-1 accuracy. Note that target-domain data are only seen during testing.

Target	GRID	i-LIDS	PRID	VIPeR	Avg.
DIMN [17]	23.4	44.8	13.1	29.9	27.8
DualNorm [16]	29.2	58.3	54.3	**38.6**	45.1
Ours	**32.8**	**62.0**	**57.3**	38.1	**47.5**

Table 2. Performance comparisons of domain adaptation based methods in terms of averaged Rank-1 accuracy. Note that baseline and DANN observe only source-domain data during training without the access to any information from target domain.

Target	GRID	i-LIDS	PRID	VIPeR	Avg.
Baseline	18.8	52.5	14.8	32.0	29.5
DANN [20]	29.0	57.2	56.8	37.8	45.2
Ours	**32.8**	**62.0**	**57.3**	**38.1**	**47.5**

3.3. Comparisons Against State-of-the-Art

In Table 1, we compare our proposed method with two state-of-the-arts [16, 17] which attempted the DG setting for re-ID. From this table, we see that our method performed favorably well and observed performance margins over the state-of-the-art methods. We achieved the averaged **Rank-1 accuracy** of **47.5%** on the four target datasets. Compared to DIMN [17], our method learned domain invariant representations, while DMIN learned a mapping between a person image and its identity classifier weight without deriving a domain invariant latent space, and thus failed to generalize to target domain. Compared to DualNorm [16] which simply adopted BN and IN layers to alleviate the domain shift, our method meta-learned to capture identity information under the proposed episodic scheme, and thus our averaged Rank-1 accuracy was higher by **2.4%**. From the experiment, the effectiveness of our model for domain generalized person re-ID was quantitatively verified.

As mentioned in Sec. 2.2, we did not apply the adversarial training strategy to derive domain invariant representations since person identities are disjoint for any two different re-ID datasets, and thus the adversarial training strategy might produce pose-invariant or camera-invariant features instead of domain-invariant ones. To verify this, we compare our method to a simple baseline and DANN [20], which derived domain invariant features via adversarial loss and is a popular UDA method. For the baseline, we simply trained a single ResNet-50 to predict the person identities for all do-

Table 3. Ablation studies analyzing the importance of each introduced loss function.

Target	GRID	i-LIDS	PRID	VIPeR	Avg.
Ours w/o $\mathcal{L}_{consis}$	30.6	60.3	55.7	**40.1**	46.7
Ours	**32.8**	**62.0**	**57.3**	38.1	**47.5**

mains. Although DANN was originally proposed for UDA, we repurposed it for domain generalization by adding an auxiliary classifier to the baseline model with a gradient-reversal layer for domain confusion. As shown in Table 2, our averaged Rank-1 accuracy was higher than DANN by **2.3%**. This demonstrated our cross-domain consistency loss is practically more preferable than the adversarial loss.

3.4. Ablation Studies

To further analyze the importance of each introduced loss function, we conduct ablation studies shown in Table 3. When $\mathcal{L}_{consis}$ is turned off, our model would fail to generalize to unseen domain since there is no explicit constraint for learning domain invariant representations. From the above experiment, we confirmed that each introduced loss function is vital and beneficial to domain generalized person re-ID.

4. CONCLUSIONS

In this paper, we addressed the challenging domain generalized person re-ID problem, in which target-domain data is not available during training. With the proposed meta learning framework, we utilized episodic training strategy with pre-trained domain specific feature extractors, and learn domain-invariant yet identity-discriminative features with re-ID performance guarantees. Our experiments on multiple benchmark datasets confirmed that our approach performed favorably against state-of-the-art domain adaptation and domain generalization methods on this challenge task.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[2] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *CVPR*, 2008.

[3] F. M. Khan and F. Brémond, "Person re-identification for real-world surveillance systems," *arXiv preprint arXiv:1607.05975*, 2016.

[4] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[5] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *CVPR*, 2018.

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

[7] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Multimedia*, 2018.

[8] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, et al., "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *NeurIPS*, 2018.

[9] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *ECCV*, 2018.

[10] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *CVPR*, 2019.

[11] Y. Chen, X. Zhu, and S. Gong, "Instance-guided context rendering for cross-domain person re-identification," in *ICCV*, 2019.

[12] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *ICCV*, 2015.

[13] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *AAAI*, 2018.

[14] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," in *NeurIPS*, 2018.

[15] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *ICCV*, 2019.

[16] J. Jia, Q. Ruan, and T. M. Hospedales, "Frustratingly easy person re-identification: Generalizing person re-id in practice," *arXiv preprint arXiv:1905.03422*, 2019.

[17] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *CVPR*, 2019.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[19] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, 2016.

[21] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[22] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017.

[23] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.

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

[25] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *CVPR*, 2009.

[26] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *BMVC*, 2009.

[27] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *SCIA*, 2011.

[28] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.