Person Re-Identification by Deep Learning Multi-Scale Representations

Yanbei Chen
Queen Mary University of London
yanbei.chen@qmul.ac.uk

Xiatian Zhu
Vision Semantics Ltd.
eddy@visionsemantics.com

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Abstract

Existing person re-identification (re-id) methods depend mostly on single-scale appearance information. This not only ignores the potentially useful explicit information of other different scales, but also loses the chance of mining the implicit correlated complementary advantages across scales. In this work, we demonstrate the benefits of learning multi-scale person appearance features using Convolutional Neural Networks (CNN) by aiming to jointly learn discriminative scale-specific features and maximise multi-scale feature fusion selections in image pyramid inputs. Specifically, we formulate a novel Deep Pyramid Feature Learning (DPFL) CNN architecture for multi-scale appearance feature fusion optimised simultaneously by concurrent per-scale re-id losses and interactive cross-scale consensus regularisation in a closed-loop design. Extensive comparative evaluations demonstrate the re-id advantages of the proposed DPFL model over a wide range of state-of-the-art re-id methods on three benchmarks Market-1501, CUHK03, and DukeMTMC-reID.

1. Introduction

Person re-identification (re-id) aims at matching identity classes of person images across non-overlapping camera views deployed over open surveillance spaces. This is an inherently challenging task because person visual appearance may change dramatically in different camera views due to unknown covariates in human pose, view angle, illumination, occlusion, and background clutter [16]. Existing works focus on designing identity discriminative feature representation [17, 13, 74, 29, 38, 36] or learning matching distance metrics [22, 68, 77, 64, 40, 72, 62, 63, 65, 8] or their combination in a deep learning framework [25, 3, 60, 67, 51, 66]. By aligning local body parts for feature extraction followed by cross-view matching, existing methods often resize all the person bounding box images into a single scale as a canonical pre-processing normalisation step [32, 27, 78], that is, existing re-id models assume a normalised *single-scale* based re-id. This, however, is against that person images are almost always captured in open surveillance spaces over a large range of resolutions (scales) due to the inherent uncontrolled distances between objects and the cameras (Fig. 1). Object re-id is intrinsically a *multi-scale* matching problem.

Figure 1. Illustration of scale alignment of person bounding boxes captured at different scales (resolutions) in public space.

We argue that the single-scale approach to person re-id is suboptimal and *explicit* multi-scale representations are essential. A single-scale representation blurs salient information at different scales useful in object matching. Our consideration is partially inspired by the human visual system that takes into account jointly multi-scale visual information including feature representations at both small (global contextual) and large (local saliency) scales [39, 54]. In general, object/event/scene representation for recognition at explicitly different scales is widely adopted in computer vision [24, 43, 10], in particular the idea of constructing feature pyramids from image pyramid inputs [2, 24, 34]. A pyramid representation aims to be *scale-invariant* in the sense that a scale change in image is counteracted by a scale shift within the feature pyramids. In this work, we investigate multi-scale deep representation learning optimised for person re-id. This is under-studied in the literature.

To this end, we address the following problems: (i) Feature learning behaviours may be different and/or even mutually inconsistent at different scales, therefore a straightforward feature concatenation of multi-scales is unlikely to result in optimal feature fusion; (ii) Any complementary correlation between different pyramid levels is unknown and may not be constant for different images, therefore must be learned and optimised synergistically across data; (iii) People's appearance in open surveillance scenes is diversely captured at an arbitrary scale (unknown). This makes it challenging to learn the underlying correlations among features of different-scales to encode both the finer

and the coarser appearance information. To formulate an end-to-end multi-scale deep re-id model, one straightforward approach is by firstly combining scale-specific feature layers and then back-propagating the supervised loss to all scale-specific branches in a joint learning fashion. This design however ignores the asynchronous learning behaviour in different branches and potentially corrupts the multi-scale feature learning. To ensure *synergistically correlated* feature learning at different scales, we propose a *Deep Pyramidal Feature Learning* (DPFL) CNN architecture for learning explicitly multi-scale deep feature representation. Specifically, the DPFL consists of m scale-specific branches each for learning one input image scale in the pyramid, and an additional scale-fusion branch for learning complementary combination of multi-scale features (Figure 2). Critically, the scale-specific branches are not independent to each other but synergistically correlated. This is the joint effect of (i) simultaneously enforcing separate learning to each branch and (ii) the special design of a closed-loop cross-branch interactive regularisation mechanism. The former aims to maximise scale-specific feature discriminative capability by subjecting them all to the same identity label constraint, whilst the latter is designed to concurrently optimise the underlying complementary advantages across scales. Under such balance between individual learning and correlation learning in a closed-loop form, we allow all branches to be learned concurrently in an end-to-end fashion so as to maximise scale-specific feature learning and optimal discriminative feature selection from multi-scale representations for person re-id.

We make two **contributions** in this work: **(I)** We investigate the multi-scale feature learning problem for person re-identification. This is significantly different from typical existing re-id methods considering only a single-scale person appearance information and therefore likely to be suboptimal for re-id matching of cross-view person bounding box images captured at intrinsically different scales. **(II)** We formulate a novel *Deep Pyramidal Feature Learning* (DPFL) CNN architecture design for not only learning scale-specific discriminative features by optimising multiple classification losses on the same person label information concurrently, but also maximising jointly multi-scale complementary fusion selections by multi-scale consensus regularisation in a closed-loop form. This design overcomes the cross-scale feature learning discrepancy challenge by a principled inter-level feature interaction in the pyramid whilst achieving cumulatively multi-scale complementary feature selection over the mini-batch training iterations. Extensive comparative evaluations demonstrate the superiority of the proposed DPFL model over a wide range of state-of-the-art re-id methods on three benchmark datasets Market-1501 [76], CUHK03 [25], and DukeMTMC-reID [78].

2. Related Work

Existing person re-id works mainly focus on feature representations and matching models. Many different hand-crafted person image feature descriptors [13, 74, 61, 35, 70, 38, 29] have been designed in the past decade. They have achieved a sequence of continuous re-id performance boost on benchmarking datasets when integrated with various supervised matching models [22, 41, 75, 28, 68, 29, 30, 40, 72, 63, 73, 62, 80, 42, 71]. Recently, deep learning re-id models [26, 4, 48, 56, 44, 77, 64, 7, 11, 65, 9, 60, 27] start to take over and have obtained impressive performance. This approach is largely inspired by the strong representation auto-learning capacity of deep models benefiting from large sized labelled training data pools; and the establishment of large person re-id datasets [26, 76].

However, all these existing methods typically consider only one resolution scale of person appearance information by a standard scale normalisation process. This not only drops the potentially useful information of other different scales, but also loses the opportunity of mining the correlated complementary advantage across appearance scales. One exception is the multi-scale Triplet CNN (MS-TriCNN) re-id model [33]. In particular, the MS-TriCNN combines multi-scale features by a hard embedding layer and learns a multi-branches CNN model by backpropagating the triplet ranking loss. While sharing the high-level multi-scale feature leaning spirit, the proposed DPFL significantly differs from the MS-TriNet: (1) Beyond the scale concatenation based fusion as MS-TriCNN, DPFL uniquely considers a synergistic cross pyramid scale interaction learning and regularisation by consensus propagation. This is designed to overcome the learning discrepancy challenge in multi-scale feature optimisation. (2) Instead of MS-TriCNN's single loss design, DPFL deploys a multi-loss concurrent supervision mechanism. This allows enforcing and improving scale-specific feature individuality learning. (3) Rather than triplet ranking loss, DPFL employs the Softmax classification loss. This not only reduces substantially the notorious model training complexity, but also improves the model learning scalability when large per-camera imbalanced training data is provided. As shown in our evaluations, these design considerations will contribute collectively to the significant re-id matching performance advantage of our DPFL over the other alternative of multi-scale learning model (MS-TriNet).

3. Multi-Scale Person Re-Identification

3.1. Problem Statement

We aim to learn a deep representation model for generic distance (e.g. L1, L2) based person re-identification without any specific metric transformation. We assume a set of n training images $\mathcal{I} = \{\boldsymbol{I}_i\}_{i=1}^{n}$ with the correspond-

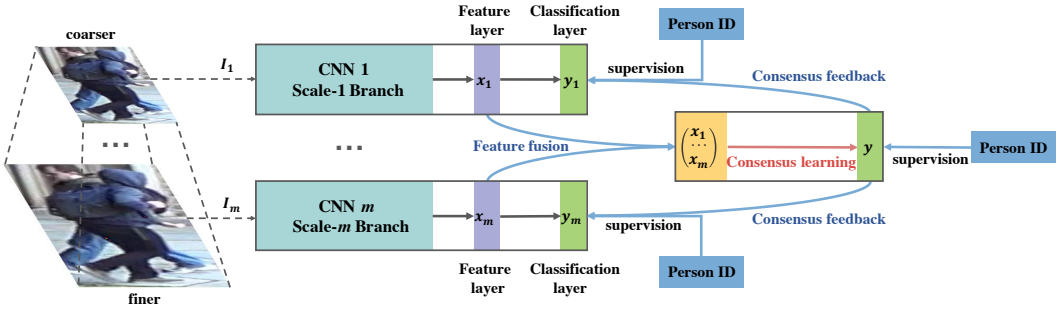

Figure 2. Overview of the proposed Deep Pyramidal Feature Learning (DPFL). The DPFL consists of m scale-specific re-id feature learning branches and one multi-scale feature fusion branch. The training of each branch is supervised by the same identity class label constraint concurrently. Therefore, the multi-scale fusion branch aims to learn the consensus on identity classes across m scales. We call this "Consensus Learning". Importantly, the learned consensus is propagated back to all individual scale-specific branches concurrently to regulate their mini-batch iterative learning behaviour together with groundtruth identity label supervision.

ing identity class labels as $\mathcal{Y} = \{y_i\}_{i=1}^{n}$. These training images capture the visual appearance and variation of n_{id} (where $y_i \in [1, \cdots, n_{\text{id}}]$) different people under multiple non-overlapping camera views. A re-id model needs to learn from these image-identity correspondence relations and importantly transfer the learned knowledge to recognise other unseen person identities in deployment. To that end, we formulate a Deep Pyramidal Feature Learning (DPFL) CNN model that aims to discover and capture concurrently complementary discriminative appearance information about person identity from multiple resolution scales of the bounding box image in order to optimise person re-id matching under significant viewing condition changes across distinct locations. This is in contrast to most existing re-id methods typically depending only on one scale feature representation alone.

3.2. Deep Multi-Scale Feature Learning

The overall network design of the proposed DPFL model is depicted in Figure 2. This DPFL model have $(m + 1)$ feed-forward sub-network branches: (1) m branches of scale-specific sub-networks with an identical structure for learning the most discriminative visual features for each individual pyramid scale of person bounding box images; (2) One fusion branch responsible for learning the discriminative feature selection and optimal integration of m scale-specific representations of the same images. We aim to concurrently optimise per-scale discriminative feature representations and discover correlated complementary combination between different scale feature selections in the pyramid. This is achieved by designing a *Deep Pyramidal Feature Learning* model that subjects both scale-specific and scale-fused branches to the same identity label supervision and critically further propagates the multi-scale consensus as a kind of *feedback* to regulate the learning behaviour of scale-specific sub-networks. This design forms a closed-loop "first multi-scale fusion then consensus propagation" scheme. In particular, the DPFL model has three parts: (I) Single Scale Feature Learning; (II) Multi-Scale Consensus Learning; (III) Feature Regularisation by Consensus Propagation. We describe the detailed architecture components design below.

(I) Single Scale Feature Learning We construct the scale-specific branches using the 42-layers Inception-V3 CNN architecture design [53] due to its high computational cost-efficiency (higher modelling capacity at a smaller parameter size) and the capability for learning more discriminative visual features at varying spatial scales. Other architectures, e.g. MobileNet [21], ResNet [18] or VGG-Net [49], can be readily applied. The base network choice is independent of our DPFL model design.

For single scale model training, we utilise the Softmax *classification* loss function so as to optimise person identity discrimination given training labels of multiple person classes extracted from pair-wise labelled re-id dataset. Formally, we predict the posterior probability $\tilde{y}_i$ of training image $\boldsymbol{I}_i$ over the given identity label y_i:

$$p_i = p(\tilde{y}_i = y_i|\boldsymbol{I}_i) = \frac{\exp(\boldsymbol{w}_{y_i}^{\top}\boldsymbol{x}_i)}{\sum_{k=1}^{n_{\text{id}}}\exp(\boldsymbol{w}_k^{\top}\boldsymbol{x}_i)} \quad (1)$$

where $\boldsymbol{x}_i$ refers to the feature vector of $\boldsymbol{I}_i$ from the corresponding branch, and $\boldsymbol{w}_k$ the prediction function parameter of training identity class k. The per-scale model training loss on a batch of n_{bs} images is computed as:

$$l_{\text{brch}} = -\frac{1}{n_{\text{bs}}}\sum_{i=1}^{n_{\text{bs}}}\log\Big(p(\tilde{y}_i = y_i|\boldsymbol{I}_i)\Big) \quad (2)$$

Loss Function Choice Instead of the more common pairwise or triplet loss functions [25, 3, 51, 9, 19], we select the classification loss due to: **(i)** Significantly *simplified* training data batch construction, e.g. random sampling with no notorious tricks required, as shown by seminal deep classification methods [23, 49] and recent re-id methods [27, 67]. This makes our DPFL model more scalable in real-world applications with very large training population

sizes when available. This also eliminates the *undesirable* need for carefully forming pairs and/or triplets in preparing training splits, as in most existing methods, due to the inherent imbalanced negative and positive pair size distributions. **(ii)** Visual psychophysical findings suggest that representations optimised for classification tasks generalise well to novel categories [12]. We consider that re-id tasks are about model generalisation to unseen test identity classes given training data on *independent* seen identity classes. The DPFL model learning exploits this general classification learning principle beyond the strict pair-wise relative verification loss in most existing re-id models.

(II) Multi-Scale Consensus Learning We perform multi-scale consensus learning on person identity classes from m scale-specific branches. To this end, we firstly perform feature fusion across scales. In the DPFL instantiation by Inception-V3, we achieve the feature fusion on the highest convolutional feature maps (of shape $c\times c\times 2048$) by an operation of averaging-pooling$\rightarrow$vector-concatenation$\rightarrow$dropout. The spacial size c is proportional to the input image resolution scale. This produces a $2048\times m$-dimensional fused feature representation for multi-scale consensus learning. For design simplicity and cost efficiency, we directly deploy an identity classification layer (i.e. *consensus learning layer*) to the multi-scale fused feature. We similarly utilise the Softmax classification loss (Eq. (2)) for consensus classification learning as in the scale-specific branches.

(III) Feature Regularisation by Consensus Propagation We propose regularising the scale-specific and therefore the entire feature learning by multi-scale person identity consensus in a closed-loop. Specifically, we further propagate the consensus as extra *feedback* information to regularise the batch learning of all scale-specific branches concurrently. Inspired by the teacher-student learning spirit [20], we do this propagation by exploiting the sample-wise probability prediction $\tilde{P} = [\tilde{p}_1, \cdots, \tilde{p}_i, \cdots, \tilde{p}_{n_{\text{id}}}]$ (i.e. person identity consensus) with the elements defined as:

$$\tilde{p}_i = \tilde{p}(\tilde{y}_i = y_i|\boldsymbol{I}_i) = \frac{\exp(\frac{z_i}{T})}{\sum_k \exp(\frac{z_k}{T})} \tag{3}$$

where z is the logit (i.e. unnormalised log-probability) and T is a temperature with higher values giving softer probability distributions over classes. We set $T = 1$ (i.e. the probability prediction over all training identity classes) in our experiments. More specifically, we utilise the consensus probability $\tilde{P}$ as the teacher signal (called "soft target" [20] *versus* the groundtruth one-hot "hard target") to guide the learning process of all scale-specific branches (student) concurrently via enforcing an additional regularisation in Eq. (2) as:

$$l_{\text{scale}} = l_{\text{brch}} + \lambda\mathcal{H}(\tilde{P}, P) \tag{4}$$

where the hyperparameter λ controls the importance trade-off between the two terms. $P = [p_1, \cdots, p_i, \cdots, p_{n_{\text{id}}}]$ defines the probability prediction over all n_{id} identity classes by the corresponding scale-specific branch (Eq. (1)). $\mathcal{H}(\tilde{P}, P)$ is the consensus regularisation term that denotes the cross-entropy between two distributions $\tilde{P}$ and P, i.e.

$$\mathcal{H}(\tilde{P}, P) = -\frac{1}{n_{\text{id}}}\sum_{i=1}^{n_{\text{id}}}\left(\tilde{p}_i\ln(p_i) + (1-\tilde{p}_i)\ln(1-p_i)\right) \tag{5}$$

We fix $\lambda = 1$ in Eq. (4) in our evaluations, i.e. both the "soft targets" and "hard targets" contribute equally to the learning process of each student (scale-specific) branch[1].

Discussion The proposed DPFL model shares some spirit of Knowledge Distillation (KD) by teacher-student learning [6, 20]. This is because, the consensus feedback propagation in DPFL can be considered as a kind of knowledge transfer via aligning higher-entropy soft targets.

The additional knowledge is a result of per-batch multi-scale consensus learning on-the-fly. However, DPFL differs significantly from KD in the following perspectives: **(a)** Objective: KD aims to achieve model compression by transferring the knowledge learned by a large teacher model or ensemble to a small deep model. The rational behind is that, small models may have similar representation capacity but are harder to train as compared to large counterparts [5]. In contrast, DPFL aims to obtain the most discriminative pyramidal representation via interactive multi-scale feature selection learning. **(b)** Dynamics: KD requires to explicitly pre-train a powerful teacher model. In contrast, DPFL collectively exploits the per-batch outputs of all student models to generate the teacher signals, e.g. a committee of student models as a whole play a *virtual* teacher role. Consequently, DPFL performs knowledge transfer *dynamically* in an interactive manner rather than *statically* as KD.

3.3. Model Optimisation

The proposed DPFL model can be optimised by back-propagating the gradients of per-branch loss design by using the standard Stochastic Gradient Descent algorithm. As a result, our method can be readily integrated with many existing deep neural network architectures [53, 49, 21, 23, 18] without the heavy need for modifying the optimisation algorithm. Since all branches in DPFL are interacted and correlated in a closed-loop form, we need to properly handle the operation order. We present the entire DPFL optimisation process in Alg. 1.

3.4. Re-ID by Multi-Scale DPFL Features

After the DPFL model is trained, we deploy the multi-scale fused ($2048\times m$-D with m the scale number) feature

[1] More sophisticated balancing ways, e.g. a batch-wise ramp-up function of quantifying the consensus regularisation term, can be considered but may lead to unessential distractive complexity to the overall design.

Algorithm 1 DPFL model optimisation.

Input: Multi-scale training data $\mathcal{I}$, Identity labels Y, Training iterations τ;

Output: Learned DPFL model $\mathcal{M}$;

Initialisation: Randomly initialise $\mathcal{M}$;

for iteration t in $[1:\tau]$

Single scale feature extraction

– Feedforward image pyramid inputs;

Multi-scale consensus learning

– Multi-scale feature fusion;

– Multi-scale consensus learning (Eq. (2));

Feature regularisation by consensus propagation

– Align consensus on scale-specific branches (Eq. (5));

Single scale branches update

– Backpropagate identity classification loss with the consensus regularisation (Eq. (4));

Fusion branch update

– Backpropagate identity classification loss (Eq. (2));

end for

return $\mathcal{M}$.

representation for person re-id. We utilise *only* a generic distance metric *without* camera-pair specific distance metric learning, e.g. L2 distance. Specifically, given a test probe image $\boldsymbol{I}^p$ from one camera view and a set of test gallery images $\{\boldsymbol{I}_i^g\}$ from other non-overlapping camera views: (1) We first compute their corresponding $2048{\times}m$-D feature vectors by forward-feeding multi-scale images into the trained DPFL model, denoted as $\boldsymbol{x}^p$ and $\{\boldsymbol{x}_i^g\}$. (2) We then compute the cross-camera matching distances between $\boldsymbol{x}^p$ and $\boldsymbol{x}_i^g$ by some generic matching metric, e.g. L2 distance. (3) We lastly rank all gallery images in ascendant order by their matching distances to the probe image. The proportions of true matches (in the galley) of probe person images in Rank-1 and among the higher ranks indicate the goodness of the learned DPFL features for person re-id tasks.

4. Experiments

Datasets For comparative evaluations, we utilised 3 benchmarking person re-id datasets, including Market-1501 [76], DukeMTMC-reID [78], and CUHK03 [25]. Figure 3 shows some examples of person bounding box images from these datasets. In particular, different data collection protocols (including surveillance environments) were employed in constructing these datasets: **(a)** Market-1501 has $2{\sim}6,617$ images per person captured by 6 camera views deployed around a university supermarket, with all bounding boxes automatically detected by the Deformable Part Model (DPM) [14]. **(b)** DukeMTMC-reID contains $2{\sim}426$ images per person captured by 8 camera views. This dataset was constructed from the multi-camera tracking dataset DukeMTMC [46] by random selection of manually labelled tracklet bounding boxes [78]. The raw surveillance video data were captured on a university campus. **(c)** CUHK03 consists of $4{\sim}10$ images per person from 6 camera views deployed on a university campus. This dataset was constructed by both manual labelling and auto-detection (DPM) with the latter posing more re-id challenging due to more severe bounding box misalignment and background clutters. These datasets collectively represent a wide variety of real-world person re-id deployment scenarios with different population sizes and image quality in diverse challenging viewing conditions.

(a) **Market-1501** (b) **DukeMTMC** (c) **CUHK03**

Figure 3. Example cross-view image pairs of three re-id datasets.

Evaluation Protocol We adopted the standard supervised person re-id settings to evaluate the proposed DPFL model. The training/test data splits and testing settings of each dataset is summarised in Table 1. Specifically, on Market-1501, we used the standard training/test split (750/751) [76] and evaluated both single-query and multi-query test evaluation settings. On DukeMTMC-reID, we followed [78] by splitting all 1,404 person identities into two halves 702/702 for model training and test, respectively and testing re-id tasks in the single-query setting. On CUHK03, we considered two identity split settings: (1) Repeating 20 times of random 1367/100 training/test splits and reported the averaged accuracies [25]; (2) A 767/700 training/test split introduced in [79]. The single-shot evaluation setting is utilised for both split settings.

For re-id performance measure, we used the cumulative matching characteristic (CMC) and mean Average Precision (mAP). The CMC is computed on each individual rank k as the probe cumulative percentage of truth matches appearing at ranks $\leq k$. The mAP is to measure the recall of multiple truth matches, computed by first computing the area under the Precision-Recall curve for each probe, then calculating the mean of Average Precision over all probes [76].

Implementation Details We implemented the proposed DPFL model in the Tensorflow [1] framework. For model learning, we pre-train the base network Inception-V3 [53] on the ImageNet object classification images [47] for model initialisation warmup before be trained on each target person re-id dataset. By default, we utilised $m = 2$ resolution scales in the pyramid: 299×299 (large) and 225×225 (small). The mini-batch size n_{bs} is set to 8. We trained the DPFL models until convergence (i.e. the loss value stagnates) by setting the maximal iterations $100,000$ for all the

Table 1. Statistics and evaluation protocol on three person re-id datasets. Two benchmarking split settings: 1,367/100 [25] and 767/700 [79], are considered for CUHK03. SS: Single-Shot; SQ: Single-Query; MQ: Multi-Query.

Dataset	Cameras	Identities	Identity Split		Person Bounding Box Split			Test Setting
			Training	Test	Training	Test Gallery	Test Probe	
Market1501 [76]	6	1,501	751	750	12,936	19,732	3,368	SQ, MQ
DukeMTMC-reID [78]	8	1,404	702	702	16,522	17,661	2,228	SQ
CUHK03 [25]	6	1,467	1,367/767	100/700	13,132/7,368	489/5,328	475/1,400	SS

datasets. We used the Adam optimiser [45] with an initial learning rate of 0.0002 and the momentum term $\beta_1 = 0.5$, $\beta_2 = 0.999$.

4.1. Comparisons to State-Of-The-Arts

Evaluation on Market-1501 We compare the re-id performance of 17 existing methods against the proposed DPFL model on the Market-1501 benchmark [76]. Since all person bounding boxes were generated by auto-detection, this dataset represents a more scalable re-id deployment scenario than other conventional re-id datasets with manually labelled bounding boxes. Table 2 shows the clear superiority of our DPFL model over all competitors. Specifically, compared to the only multi-scale alternative MS-TriNet, our model's performance is substantially better, e.g. improving Rank-1 by 43.5% (88.6-45.1) for single-query and 36.8% (92.2-55.4) for multi-query. Our DPFL outnumbers the deep local-global joint CNN model JLML [27] by 3.5% (88.6-85.1) for single-query and 2.5% (92.2-89.7) for multi-query in Rank-1; 7.1% (72.6-65.5) for single-query and 6.2% (80.7-74.5) for multi-query in mAP. Our method outperforms TriNet by a clear margin even when they applied 10 times test-time data augmentation. In contrast to TriNet profiting effectively (improving Rank-1 by 2.4% and mAP by 3.6%) from this computation-intensive augmentation scheme at test time, the DPFL gains only marginal benefits ($\leq$ 0.3% increase in both mAP and Rank-1). This indicates the favourable robustness of our model against the inevitable local patch misalignment and background clutter in auto-detected person bounding box images for more reliable re-id matching.

Evaluation on DukeMTMC-reID We evaluate the performance of the DPFL on the large DukeMTMC-reID dataset in single-query setting[2]. As opposite to Market-1501, the person bounding box images were manually cropped in a labour-intensive manner. While being less scalable in processing big video data, this effort is still indispensable in many deployment scenarios given imperfect auto-detection performance by enabling to accommodate missing detections and diverse varying-sized person occurrences in uncontrolled open space. On the contrary, the auto-detected person bounding boxes can be largely incomplete due to high missing detection rates especially with small person appearances or dense crowds. Table 3 shows that the DPFL outperforms all hand-crafted low-level feature based and deep CNN feature based alternative methods for re-id matching. The best competitor SVDNet is surpassed by our model in Rank-1 and mAP by 2.5% (79.2-76.7) and 3.8% (60.6-56.8), respectively. This suggests the consistent superiority of the proposed multi-scale pyramidal feature learning method over existing single-scale feature learning methods in re-id tasks with more comprehensive person bounding box images and more diverse imaging resolutions.

Table 2. Comparative evaluation on Market-1501 [76]. $^{(m+)}$: Applying m times test-time data augmentation. '*': Methods from arXiv papers (unpublished). '-': No reported result available.

Metric (%)	Single-Query		Multi-Query	
	Rank-1	mAP	Rank-1	mAP
BoW [76]	34.4	14.1	42.6	19.5
KISSME [22]	40.5	19.0	-	-
MFA [69]	45.7	18.2	-	-
kLFDA [68]	51.4	24.4	52.7	27.4
SSDAL [50]	39.4	19.6	49.0	25.8
LOMO+XQDA [29]	43.8	22.2	54.1	28.4
DNS [72]	61.0	35.7	71.6	46.0
CAN [32]	60.3	35.9	72.1	47.9
Gated-SCNN [57]	65.9	39.7	76.0	48.5
S-LSTM [58]	-	-	61.6	35.3
TMA [37]	47.9	22.3	-	-
HL [55]	59.5	-	-	-
CRAFT [8]	68.7	42.3	77.0	50.3
JLML [27]	85.1	65.5	89.7	74.5
MS-TriNet [33]	45.1	-	55.4	-
DeepTransfer* [15]	83.7	65.5	89.6	73.8
TriNet* [19]	82.5	65.5	-	-
TriNet$^{(10+)}$* [19]	84.9	69.1	90.5	76.4
DPFL	88.6	72.6	92.2	80.4
DPFL$^{(2+)}$	**88.9**	**73.1**	**92.3**	**80.7**

Evaluation on CUHK03 We evaluate the re-id performance of the DPFL in comparisons to 21 existing methods on CUHK03 with two (1367/100 and 767/700) identity split settings. Unlike Market-1501 and DukeMTMC-reID, this dataset provides both manually labelled and auto-detected (by the DPM model [14]) bounding boxes of the same people population. This allows a like-to-like comparison of model generalisation on distinct-quality person images.

[2] As this dataset was newly constructed for person re-id from the multi-target multi-camera tracking benchmark DukeMTMC [46], there are only a small number of results reported in a few unpublished arXiv papers [78, 31, 52]. Following these works, we utilise the single-query evaluation setting.

Table 3. Comparative evaluation on DukeMTMC-reID [78]. '*': Method from arXiv papers (unpublished). '+': Using additional per-person semantic attribute annotations.

Metric (%)	Rank-1	mAP
BoW+KISSME [76]	25.1	12.2
LOMO+XQDA [29]	30.8	17.0
ResNet50 [18]	65.2	45.0
ResNet50+LSRO* [78]	67.7	47.1
AttIDNet*+ [31]	70.7	51.9
SVDNet* [52]	76.7	56.8
DPFL	**79.2**	**60.6**

Table 4. Comparative evaluation on CUHK03 [26].

Setting	**1367/100 training/test split**			
Metric (%)	Labelled		Detected	
	Rank-1	mAP	Rank-1	mAP
kLFDA [68]	45.8	-	-	-
LOMO+XQDA [29]	52.2	-	46.3	-
BoW+XQDA [76]	-	-	23.0	-
MLAPG [30]	58.0	-	51.2	-
GOG+XQDA [38]	67.3	-	65.5	-
HER [62]	60.8	-	-	-
CRAFT [8]	84.3	-	-	-
FPNN [26]	20.7	-	19.9	-
CIND-Net [4]	54.7	-	45.0	-
SICI [60]	-	-	52.2	-
DNS [72]	62.6	-	54.7	-
S-LSTM [59]	-	-	57.3	46.3
Gated-SCNN [57]	-	-	61.8	51.3
CAN [32]	77.6	-	69.2	-
Fused Model [51]	72.4	-	72.0	-
FT-JSTL+DGD [67]	75.3	-	-	-
JLML [27]	83.2	-	80.6	-
DPFL	**86.7**	**82.8**	**82.0**	**78.1**
Setting	**767/700 training/test split**			
BoW+XQDA [76]	7.9	7.3	6.4	6.4
LOMO+XQDA [29]	14.8	13.6	12.8	11.5
IDE(C) [79]	15.6	14.9	15.1	14.2
IDE(C)+XQDA [79]	21.9	20.0	21.1	19.0
IDE(R) [79]	22.2	21.0	21.3	19.7
IDE(R)+XQDA [79]	32.0	29.6	31.1	28.2
DPFL	**43.0**	**40.5**	**40.7**	**37.0**

Table 4 shows that our DPFL model outperforms clearly all competitors on both versions of person images under both split settings. For example, the DPFL outperforms the JLML by 3.5% (Labelled) / 1.4% (Detected) in Rank-1 given the 1367/100 split. For the harder split 767/700, our model achieves more significant advantages over the best alternative IDE(R)+XQDA: 11.0% (Labelled) / 9.6% (Detected) in Rank-1, and 10.9% (Labelled) / 8.8% (Detected) in mAP. This further validates the performance advantage of our pyramidal feature learning method over single-scale feature learning based alternative methods under different re-id settings. On the other hand, it is observed that auto-detected person bounding boxes indeed present more re-id matching challenges than manually labelled ones, with lower re-id performance on the former obtained across all methods. This is highly expected, due to more severe misalignment and background noise in auto-detected person images introduced by inaccurate detection.

4.2. Further Analysis and Discussions

Next, we provide detailed model component analysis in terms of performance contributions on the DukeMTMC-reID and Market-1501 in the single-query re-id setting.

Table 5. Evaluating generalisation to different CNN models.

Dataset		DukeMTMC-reID		Market-1501	
Metric (%)		Rank-1	mAP	Rank-1	mAP
Inception-V3	Scale-299	70.1	48.9	85.7	66.5
	Scale-225	65.5	42.8	83.3	62.8
	DPFL	**79.2**	**60.6**	**88.6**	**72.6**

Single-Scale *versus* Multi-Scale Features We evaluate the re-id performance advantage of our multi-scale features over independently learned single-scale features. Results of models initialised by Inception-V3 in Table 5 show that the DPFL multi-scale features outperform significantly either single-scale features, e.g. surpassing the scale-299 feature on DukeMTMC-reID and Market-1501 by 9.1% (79.2-70.1) and 2.9% (88.6-85.7) in Rank-1, 11.7% (60.6-48.9) and 6.1% (72.6-66.5) in mAP, respectively. This suggests the effectiveness of our proposed multi-scale consensus regularised feature learning method in improving open space re-id matching.

Table 6. Evaluating different multi-scale feature fusion methods.

Dataset		DukeMTMC-reID		Market-1501	
Metric (%)		Rank-1	mAP	Rank-1	mAP
Inception-V3	Independent-Scales	72.2	50.3	87.2	69.5
	Joint-Scales	72.9	51.3	83.4	61.1
	DPFL	**79.2**	**60.6**	**88.6**	**72.6**

Multi-Scale Feature Fusion Approaches We compared the DPFL with two baseline multi-scale fusion methods: **(a)** Independent-Scales: Independently train individual scale-specific deep CNN models (Figure 4 (a)); and utilise the concatenation of all scale-specific feature vectors for re-id matching in deployment. **(b)** Joint-Scales: A vanilla multi-scale joint learning CNN framework capable of applying the identity classification supervision learning on the fusion

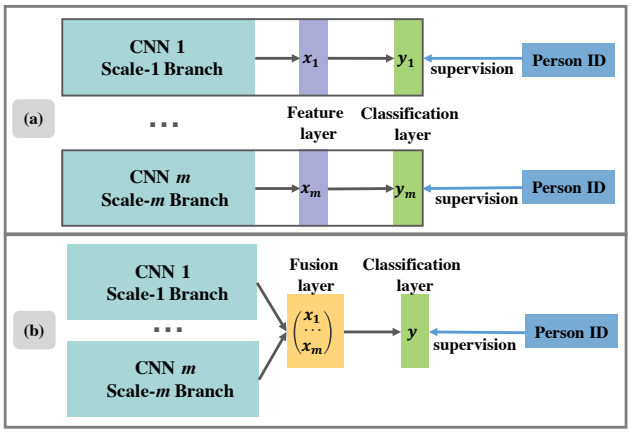

Figure 4. Illustration of two baseline multi-scale feature fusion CNN model designs: **(a)** Independent-Scales and **(b)** Joint-Scales.

of all the scale-specific features in end-to-end training (Figure 4 (b)). In re-id deployment, we similarly use the fused feature. This method shares a similar multi-scale fusion design principle as MS-TriNet [33] although a different loss function is employed.

From the results shown in Table 6, we have the following observations: (1) The DPFL outperforms both alternative multi-scale fusion methods. This suggests a clear advantage of the proposed method in maximising correlated complementary benefits of multi-scale re-id features. (2) On DukeMTMC-reID, both Independent-Scales and Joint-Scales improve re-id matching performance but only mildly. On one hand, this suggests the advantages of multi-scale features over single-scale counterparts in re-id matching. On the other hand, this also indicates that no cross-scale interaction in feature learning (Independent-Scales) or a simple multi-scale concatenation in joint learning (Joint-Scales) may result in suboptimal multi-scale feature optimisation. (3) On Market-1501, Independent-Scales consistently improves over single-scale features, but Joint-Scales even suffers a considerable (-5.4%) mAP drop as compared to the Scale-299 feature alone. This indicates that multi-scale joint end-to-end learning is non-trivial and a straightforward feature fusion alone may bring adversarial effects. A plausible reason is the underlying learning behaviour discrepancy at different scales. For instance, the large-scale branch model needs to reason more detailed localised appearance information from more raw pixels and therefore probably takes a slower learning pace. (4) The DPFL model can be considered as a synergistic combination design of Independent-Scales, Joint-Scales, and importantly the proposed multi-scale consensus propagation mechanism (Figure 2). Our model is clearly superior on both datasets, indicating that the proposed multi-scale consensus regularisation is an effective approach to overcoming the limitations of both alternatives in learning multi-scale re-id discriminative features.

Table 7. Evaluating generalisation to different CNN models.

Dataset		DukeMTMC-reID		Market-1501	
Metric (%)		Rank-1	mAP	Rank-1	mAP
Inception-V3	Scale-299	70.1	48.9	85.7	66.5
	Scale-225	65.5	42.8	83.3	62.8
	DPFL	**79.2**	**60.6**	**88.6**	**72.6**
MobileNet	Scale-224	73.8	53.9	87.5	66.4
	Scale-160	72.5	51.7	87.6	63.9
	DPFL	**77.6**	**58.6**	**90.0**	**70.6**

Generalisation to Different CNN Models We evaluate the benefits of the DPFL approach when integrated with other CNN architectures in addition to Inception-V3. We select the light MobileNet architecture [21] for particularly testing the potentials in mobile vision applications. Table 7 shows the generic capability of our DPFL method in extracting the multi-scale complementary benefits from different scales of person images when combining with either large Inception-V3 or small MobileNet CNN architectures.

5. Conclusion

We presented a novel Deep Pyramid Feature Learning (DPFL) CNN model by aiming to learn multi-scale appearance information for person re-identification. In contrast to existing re-id approaches that only employ single scale appearance features, the proposed model is capable of extracting and exploiting discriminative scale-specific features and optimal cross-scale complementary benefits by jointly learning multiple scales of person images in a pyramid subject to individual classification objective functions with a specially designed cross-scale consensus regularisation in an end-to-end training deep CNN model. This is made possible by the proposed multi-scale consensus learning and propagation mechanism. Extensive comparative evaluations on three re-id benchmark datasets were conducted to validate the advantages of the proposed DPFL model over a wide range of state-of-the-art methods on both manually labelled and auto-detected person bounding box images. We lastly provided component evaluations and analysis in terms of re-id performance so as to give the insights into the DPFL model design.

Acknowledgements

This work was partially supported by the China Scholarship Council, Vision Semantics Ltd., and the Royal Society Newton Advanced Fellowship Programme (NA150459).

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv*, 2016.

[2] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden. Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41, 1984.

[3] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[4] E. Ahmed, M. J. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] J. Ba and R. Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, pages 2654–2662, 2014.

[6] C. Bucilu, R. Caruana, and A. Niculescu-Mizil. Model compression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541, 2006.

[7] J. Chen, Z. Zhang, and Y. Wang. Relevance metric learning for person re-identification by exploiting listwise similarities. *Image Processing, IEEE Transactions on*, 24(12):4741–4755, 2015.

[8] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[9] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

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

[12] S. Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(04):449–467, 1998.

[13] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[15] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[16] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. Springer, January 2014.

[17] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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

[20] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[22] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[24] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[25] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[27] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference of Artificial Intelligence*, 2017.

[28] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[30] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *IEEE International Conference on Computer Vision*, 2015.

[31] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[32] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017.

[33] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet cnn for person re-identification. In *ACM on Multimedia Conference*, 2016.

[34] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[35] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Workshop of European Conference on Computer Vision*, 2012.

[36] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[37] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *European Conference on Computer Vision*, 2016.

[38] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[39] D. Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977.

[40] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[41] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[42] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[43] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[44] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, 2010.

[45] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[46] 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.

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

[48] 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 *European Conference on Computer Vision*, 2016.

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[50] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491. Springer, 2016.

[51] A. Subramaniam, M. Chatterjee, and A. Mittal. Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*, pages 2667–2675, 2016.

[52] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. *arXiv preprint arXiv:1703.05693*, 2017.

[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[54] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4):766, 2006.

[55] E. Ustinova and V. Lempitsky. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, pages 4170–4178, 2016.

[56] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, 2016.

[57] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016.

[58] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, 2016.

[59] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016.

[60] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[61] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *British Machine Vision Conference*, 2014.

[62] H. Wang, S. Gong, and T. Xiang. Highly efficient regression for scalable person re-identification. In *British Machine Vision Conference*, 2016.

[63] H. Wang, S. Gong, X. Zhu, and T. Xiang. Human-in-the-loop person re-identification. In *European Conference on Computer Vision*, 2016.

[64] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, 2014.

[65] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January 2016.

[66] C. Weihua, C. Xiaotang, Z. Jianguo, and H. Kaiqi. A multi-task deep network for person re-identification. In *AAAI Conference on Artificial Intelligence*, 2017.

[67] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[68] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*. 2014.

[69] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[70] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *European Conference on Computer Vision*, 2014.

[71] H. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE International Conference on Computer Vision*, 2017.

[72] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[73] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[74] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[75] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[76] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015.

[77] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, March 2013.

[78] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017.

[79] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[80] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng. Fast open-world person re-identification. *IEEE Transactions on Image Processing*, 2017.