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 [2, 24, 34]. 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.

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 [15, 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, 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 benefitting 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 $\mathcal{I} = \{I_i\}_{i=1}^{n}$ with the correspond-
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 \(\hat{y}_i\) of training image \(I_i\) over the given identity label \(y_i\):

\[
p_i = p(\hat{y}_i = y_i | I_i) = \frac{\exp(w_{\hat{y}_i}^T x_i)}{\sum_{k=1}^{n_id} \exp(w_k^T x_i)}
\]

where \(x_i\) refers to the feature vector of \(I_i\) from the corresponding branch, and \(w_k\) 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:

\[
b_{brch} = -\frac{1}{n_{bs}} \sum_{i=1}^{n_{bs}} \log \left( p(\hat{y}_i = y_i | I_i) \right)
\]

Loss Function Choice Instead of the more common pair-wise 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
The proposed DPFL model shares some spirit in ability distributions over classes. We set $T$ is a temperature with higher values giving softer probability prediction over all $n_{id}$ identity classes by the corresponding scale-specific branch (Eq. (1)). $H(\hat{P}, P)$ is the consensus regularisation term that denotes the cross-entropy between two distributions $\hat{P}$ and $P$, i.e.

$$H(\hat{P}, P) = -\frac{1}{n_{id}} \sum_{i=1}^{n_{id}} (\hat{p}_i \ln(p_i) + (1 - \hat{p}_i) \ln(1 - p_i))$$

where the hyperparameter $\lambda$ controls the importance trade-off between the two terms. $P = [p_1, \ldots, p_{n_{id}}]$ defines the probability prediction over all $n_{id}$ identity classes by the corresponding scale-specific branch. 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.

**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×$m$-D with $m$ the scale number) feature
This dataset was constructed from the multi-camera track-bounding boxes automatically detected by the Deformable views deployed around a university supermarket, with all 2 images per person captured by 8 camera views.

Datasets of the learned DPFL features for person re-id tasks. We utilise only a generic distance metric without camera-pair specific distance metric learning, e.g. L2 distance. Specifically, given a test probe image $I^p$ from one camera view and a set of test gallery images $\{I^g_i\}$ from other non-overlapping camera views:

1. We first compute their corresponding $2048 \times m \times D$ feature vectors by forward-feeding multi-scale images into the trained DPFL model, denoted as $x^p$ and $\{x^g_i\}$.
2. We then compute the cross-camera matching distances between $x^p$ and $x^g_i$ 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 track-

Algorithm 1 DPFL model optimisation.

Input: Multi-scale training data $\mathcal{X}$, Identity labels $\mathcal{Y}$, Training iterations $\tau$.
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 $I^p$ from one camera view and a set of test gallery images $\{I^g_i\}$ from other non-overlapping camera views: (1) We first compute their corresponding $2048 \times m \times D$ feature vectors by forward-feeding multi-scale images into the trained DPFL model, denoted as $x^p$ and $\{x^g_i\}$. (2) We then compute the cross-camera matching distances between $x^p$ and $x^g_i$ 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.

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: $299 \times 299$ (large) and $225 \times 225$ (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

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

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>
<thead>
<tr>
<th>Dataset</th>
<th>Cameras</th>
<th>Identities</th>
<th>Identity Split</th>
<th>Person Bounding Box Split</th>
<th>Test Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market1501 [76]</td>
<td>6</td>
<td>1,501</td>
<td>Training 751</td>
<td>Training 12,936</td>
<td>SQ, MQ</td>
</tr>
<tr>
<td>DukeMTMC-reID [78]</td>
<td>8</td>
<td>1,404</td>
<td>Test 750</td>
<td>Test 19,732</td>
<td>SQ</td>
</tr>
<tr>
<td>CUHK03 [25]</td>
<td>6</td>
<td>1,467</td>
<td>1,367/767</td>
<td>767/700</td>
<td>SS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric (%)</th>
<th>Single-Query</th>
<th>Multi-Query</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
</tr>
<tr>
<td>BoW [76]</td>
<td>34.4</td>
<td>14.1</td>
</tr>
<tr>
<td>KISSME [22]</td>
<td>40.5</td>
<td>19.0</td>
</tr>
<tr>
<td>MFA [69]</td>
<td>45.7</td>
<td>18.2</td>
</tr>
<tr>
<td>kLFDA [68]</td>
<td>51.4</td>
<td>24.4</td>
</tr>
<tr>
<td>SSDAL [50]</td>
<td>39.4</td>
<td>19.6</td>
</tr>
<tr>
<td>LOMO+XQDA [29]</td>
<td>43.8</td>
<td>22.2</td>
</tr>
<tr>
<td>DNS [72]</td>
<td>61.0</td>
<td>35.7</td>
</tr>
<tr>
<td>CAN [32]</td>
<td>60.3</td>
<td>35.9</td>
</tr>
<tr>
<td>Gated-SCNN [57]</td>
<td>65.9</td>
<td>39.7</td>
</tr>
<tr>
<td>S-LSTM [58]</td>
<td>-</td>
<td>61.6</td>
</tr>
<tr>
<td>TMA [37]</td>
<td>47.9</td>
<td>22.3</td>
</tr>
<tr>
<td>HL [55]</td>
<td>59.5</td>
<td>-</td>
</tr>
<tr>
<td>CRAFT [8]</td>
<td>68.7</td>
<td>42.3</td>
</tr>
<tr>
<td>JLML [27]</td>
<td>85.1</td>
<td>65.5</td>
</tr>
<tr>
<td>MS-TriNet [33]</td>
<td>45.1</td>
<td>-</td>
</tr>
<tr>
<td>DeepTransfer* [15]</td>
<td>83.7</td>
<td>65.5</td>
</tr>
<tr>
<td>TriNet* [19]</td>
<td>82.5</td>
<td>65.5</td>
</tr>
<tr>
<td>TriNet(10+)* [19]</td>
<td>84.9</td>
<td>69.1</td>
</tr>
<tr>
<td>DPFL</td>
<td>88.6</td>
<td>72.6</td>
</tr>
<tr>
<td>DPFL(2+)</td>
<td>88.9</td>
<td>73.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank-1</th>
<th>mAP</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>42.6</td>
<td>19.5</td>
<td>89.6</td>
<td>72.6</td>
</tr>
</tbody>
</table>

\(^{m+}\): Applying \(m\) times test-time data augmentation. '+'*: Methods from arXiv papers (unpublished). '-'*: No reported result available.
Table 3 Comparative evaluation on DukeMTMC-reID [78]. ‘*’: Method from arXiv papers (unpublished). ‘+’: Using additional per-person semantic attribute annotations.

<table>
<thead>
<tr>
<th>Metric (%)</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW+KISSME [76]</td>
<td>25.1</td>
<td>12.2</td>
</tr>
<tr>
<td>LOMO+XQDA [29]</td>
<td>30.8</td>
<td>17.0</td>
</tr>
<tr>
<td>ResNet50 [18]</td>
<td>65.2</td>
<td>45.0</td>
</tr>
<tr>
<td>ResNet50+LSRO* [78]</td>
<td>67.7</td>
<td>47.1</td>
</tr>
<tr>
<td>AttIDNet++ [31]</td>
<td>70.7</td>
<td>51.9</td>
</tr>
<tr>
<td>SVSNet*</td>
<td>76.7</td>
<td>56.8</td>
</tr>
<tr>
<td><strong>DPFL</strong></td>
<td><strong>79.2</strong></td>
<td><strong>60.6</strong></td>
</tr>
</tbody>
</table>

Table 4. Comparative evaluation on CUHK03 [26].

<table>
<thead>
<tr>
<th>Setting</th>
<th>1367/100 training/test split</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric (%)</td>
<td>Labelled</td>
</tr>
<tr>
<td>kLFDA [68]</td>
<td>45.8</td>
</tr>
<tr>
<td>LOMO+XQDA [29]</td>
<td>52.2</td>
</tr>
<tr>
<td>BoW+XQDA [76]</td>
<td>-</td>
</tr>
<tr>
<td>MLAPG [30]</td>
<td>58.0</td>
</tr>
<tr>
<td>GOG+XQDA [38]</td>
<td>67.3</td>
</tr>
<tr>
<td>HER [62]</td>
<td>60.8</td>
</tr>
<tr>
<td>CRAFT [8]</td>
<td>84.3</td>
</tr>
<tr>
<td>FPN [26]</td>
<td>20.7</td>
</tr>
<tr>
<td>CIND-Net [4]</td>
<td>54.7</td>
</tr>
<tr>
<td>SCIC [60]</td>
<td>-</td>
</tr>
<tr>
<td>DNS [72]</td>
<td>62.6</td>
</tr>
<tr>
<td>S-LSTM [59]</td>
<td>-</td>
</tr>
<tr>
<td>Gated-SCNN [57]</td>
<td>-</td>
</tr>
<tr>
<td>CAN [32]</td>
<td>77.6</td>
</tr>
<tr>
<td>Fused Model [51]</td>
<td>72.4</td>
</tr>
<tr>
<td>FT-JSTL+DGD [67]</td>
<td>75.3</td>
</tr>
<tr>
<td>JLML [27]</td>
<td>83.2</td>
</tr>
<tr>
<td><strong>DPFL</strong></td>
<td><strong>86.7</strong></td>
</tr>
</tbody>
</table>

Table 5. Evaluating generalisation to different CNN models.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DukeMTMC-reID</th>
<th>Market-1501</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric (%)</td>
<td>Rank-1</td>
<td>mAP</td>
</tr>
<tr>
<td>Inception-V3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale-299</td>
<td>70.1</td>
<td>69.5</td>
</tr>
<tr>
<td>Scale-225</td>
<td>65.5</td>
<td>62.8</td>
</tr>
<tr>
<td><strong>DPFL</strong></td>
<td><strong>79.2</strong></td>
<td><strong>60.6</strong></td>
</tr>
</tbody>
</table>

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 6. Evaluating different multi-scale feature fusion methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DukeMTMC-reID</th>
<th>Market-1501</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric (%)</td>
<td>Rank-1</td>
<td>mAP</td>
</tr>
<tr>
<td>Inception-V3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent-Scales</td>
<td>72.2</td>
<td>50.3</td>
</tr>
<tr>
<td>Joint-Scales</td>
<td>72.9</td>
<td>51.3</td>
</tr>
<tr>
<td><strong>DPFL</strong></td>
<td><strong>79.2</strong></td>
<td><strong>60.6</strong></td>
</tr>
</tbody>
</table>

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.

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
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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DukeMTMC-reID</th>
<th>Market-1501</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric (%)</td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
</tr>
<tr>
<td>Scale-299</td>
<td>70.1</td>
<td>85.7</td>
</tr>
<tr>
<td>Scale-225</td>
<td>65.5</td>
<td>83.3</td>
</tr>
<tr>
<td>DPFL</td>
<td>79.2</td>
<td>88.6</td>
</tr>
<tr>
<td>MobileNet</td>
<td>73.8</td>
<td>87.5</td>
</tr>
<tr>
<td>Scale-160</td>
<td>72.5</td>
<td>87.6</td>
</tr>
<tr>
<td>DPFL</td>
<td>77.6</td>
<td>90.0</td>
</tr>
<tr>
<td><strong>V3</strong></td>
<td><strong>58.6</strong></td>
<td><strong>72.6</strong></td>
</tr>
</tbody>
</table>

**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


[34] D. G. Lowe. Distinctive image features from scale-invariant keypoint.


    In *European Conference on Computer Vision*, 2016.


    *ICLR*, 2016.


    In *European Conference on Computer Vision*, 2016.


    In *European Conference on Computer Vision*, 2016.


    In *European Conference on Computer Vision*, 2016.


    In *European Conference on Computer Vision*, 2016.


