Pattern Recognition 121 (2022) 108239

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning hybrid ranking representation for person re-identification

Guile Wu[a,*], Xiatian Zhu[b], Shaogang Gong[a]

[a] Queen Mary University of London, London, United Kingdom
[b] Vision Semantics Limited, London, United Kingdom

ARTICLE INFO

Article history:
Received 26 August 2019
Revised 1 March 2021
Accepted 8 August 2021
Available online 9 August 2021

Keywords:
Person re-identification
Ranking representation
Ranking ensemble

ABSTRACT

Contemporary person re-identification (re-id) methods mostly compute *independently* a feature representation of each person image in the query set and the gallery set. This strategy fails to consider any *ranking context information* of each probe image in the query set represented implicitly by the whole gallery set. Some recent re-ranking re-id methods therefore propose to take a post-processing strategy to exploit such contextual information for improving re-id matching performance. However, post-processing is independent of model training without jointly optimising the re-id feature and the ranking context information for better compatibility. In this work, for the first time, we show that the appearance feature and the ranking context information can be jointly optimised for learning more discriminative representations and achieving superior matching accuracy. Specifically, we propose to learn a hybrid ranking representation for person re-id with a two-stream architecture: (1) In the *external* stream, we use the ranking list of each probe image to learn plausible visual variations among the top ranks from the gallery as the external ranking information; (2) In the *internal* stream, we employ the part-based fine-grained feature as the internal ranking information, which mitigates the harm of incorrect matches in the ranking list. Assembling these two streams generates a hybrid ranking representation for person matching. Extensive experiments demonstrate the superiority of our method over the state-of-the-art methods on four large-scale re-id benchmarks (Market-1501, DukeMTMC-ReID, CUHK03 and MSMT17), under both supervised and unsupervised settings.

1. Introduction

Person re-identification (re-id) aims to match people across non-overlapping camera views deployed over distributed physical locations [1–3]. This capability underpins a wide range of real-world intelligent vision applications, such as video surveillance, motion analysis and anomaly detection. Although many effective re-id methods have been proposed in recent years [4–7], matching people across multiple disjoint camera views remains a challenging task due to arbitrary person pose variation, background clutter, illumination variation, occlusion, etc.

Contemporary person re-id methods are mostly designed to compute the feature representation of each person image in the query set and the gallery set *independently* [8–11]. In other words, the feature representation of a probe image in the query set is *independently* inferred from that of the images in the gallery. This approach does not leverage any useful ranking context information in the gallery set for modelling the overall contextual correlation between the query set and the gallery set. As a result, the returned ranking list usually contains more false matches from the gallery that are visually similar to the probe but semantically incorrect in person label (identity). Recent research in re-ranking based re-id [12–15] shows that given a specific gallery set, a probe image and its true matches in the gallery would generate similar ranking lists where the semantically relevant images are ranked at the top. Importantly, exploiting such ranking context information helps to improve the re-id performance by retrieving more correct matches in higher ranking orders. However, these re-ranking methods all adopt a post-processing (post-rank) strategy, which is independent of model training without jointly optimising the re-id feature and the ranking context information for better compatibility.

In this work, we investigate the ranking context information for learning more discriminative feature representations in person re-id. Specifically, for a given probe image from the query set, the top-ranked candidates from the gallery set resemble similar view variations as the probe image, so the correlations among these candidates reflect the ranking context information between the probe and the gallery. Leveraging such contextual information helps to generate a more discriminative representation for person

* Corresponding author.
E-mail addresses: guile.wu@qmul.ac.uk (G. Wu), eddy.zhuxt@gmail.com (X. Zhu), s.gong@qmul.ac.uk (S. Gong).

https://doi.org/10.1016/j.patcog.2021.108239
0031-3203/

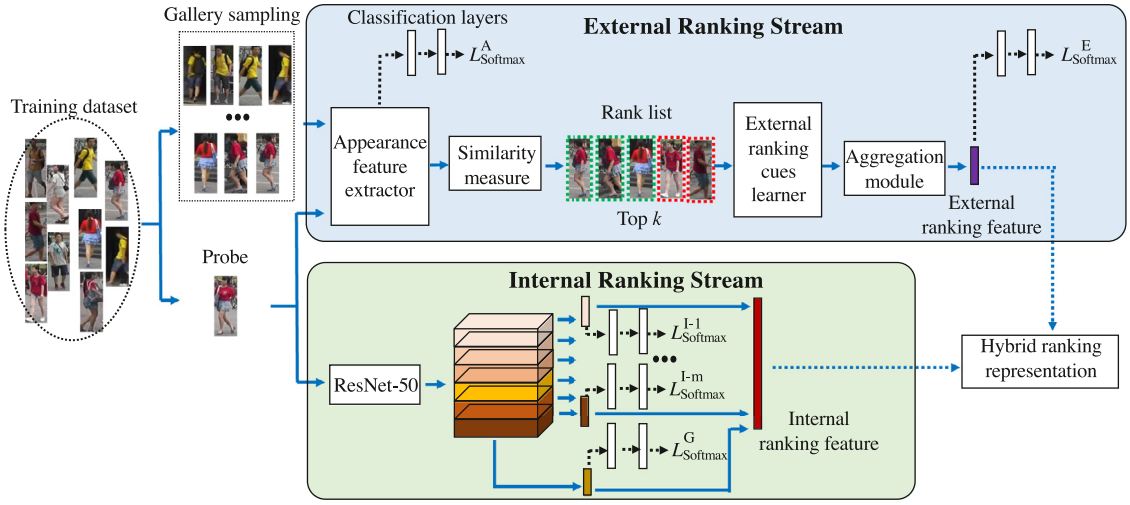


Fig. 1. An overview of the proposed two-stream architecture for learning a hybrid ranking representation for person re-id. The *external* stream (top block) is designed to learn the external ranking information from the ranking list, whilst the *internal* stream (bottom block) is formed to extract the internal part-level fine-grained information. These two streams formulate a hybrid ranking ensemble in a unified architecture.

matching. This would benefit both supervised and unsupervised re-id, due to no need for extra labelling – the ranking list can be automatically generated by matching a probe with the gallery set. Although the ranking list may inevitably be contaminated by incorrect candidates from the gallery (a.k.a. *false positive matches*), we observe that this is partly due to the over coarse description of each person's appearance as a whole. This problem can be alleviated by further utilising structured fine-grained features, which encode finer details of body parts to provide richer information, for minimising false gallery positive matches.

To this end, we propose a two-stream RANkingG Ensemble (RANGEv2) approach (see Fig. 1) for person re-id by learning a hybrid ranking representation. In the external stream[1], we jointly optimise the re-id appearance feature and the ranking context information to generate the external ranking representation. This associates the probe and the gallery for retrieving more true matches among the top ranks in the returned ranking list. In the internal stream[2], we leverage part-based features to generate the internal ranking representation, which captures finer details of body parts and provides extra information for minimising false gallery positive matches. Formulating these two streams into an integrated architecture enables to learn a hybrid ranking representation to maximise their respective advantages in a cohort for improving the overall model performance. The *contributions* of this work are:

(I) We propose the idea of jointly learning the ranking context information and the appearance feature to extract a more discriminative feature representation for person re-id. To our best knowledge, this is the first attempt to jointly optimise the two types of information in an online trainable fashion.
(II) We design a novel two-stream architecture to learn a hybrid ranking representation for more effective person re-id. Assembling the external and internal streams helps to maximise their respective advantages in a cohort and improves the model performance.
(III) We introduce an aggregation module to optimise the cumulation of the ranking context information as the external ranking representation.
(IV) Our method achieves superior performance compared with the state-of-the-art alternative methods on four large-scale person re-id benchmarks. We also verify that the proposed idea benefits both supervised and unsupervised cross-domain person re-id.

A preliminary version of this work has been published in [16]. In this work, we present several key improvements as follows:

(I) We reformulate the proposed idea by designing a new two-stream architecture. Each stream is online trainable and their ensemble maximises their respective advantages for improving the model performance.
(II) In the external stream, we introduce a probe-gallery sampling strategy to facilitate the online ranking context learning. And we further propose an aggregation module to better aggregate the ranking context information as the external ranking representation. As such, image-level classification losses and ranking-list-level classification losses can be jointly optimised.
(III) We improve the internal ranking stream by replacing a sequential model with a stronger part-based model for learning internal correlations. We further refine the optimisation objective to obtain more discriminative fine-grained features.
(IV) We conduct more comprehensive experimental evaluations. We demonstrate that our RANGEv2 achieves significantly superior performance than the preliminary model RANGE [16]. For example, in supervised person re-id, RANGEv2 gains an mAP improvement of nearly 10% on CUHK03/MSMT-17 and around 5% on Market-1501/DukeMTMC-ReID.

2. Related work

In recent years, person re-id has been extensively studied [4,6,10,17]. In this section, we discuss two groups of mostly related works, *i.e.* ranking context information based re-id and body part based re-id.

2.1. Ranking context information based person re-id

Given a probe image, the returned ranking list obtained from the gallery set usually encompasses rich ranking context information which is useful for improving the re-id performance. Previous re-id methods typically adopt a post-processing (post-rank)

[1] As the top-ranked gallery images are *not* part of a probe, we refer to it as the "external" information.
[2] As the structured parts are partitioned from a probe, we refer to it as the "internal" information.

strategy to mine such contextual information [12,13,18,19]. In [12], Zhong et al. assume that k-reciprocal nearest neighbours indicate the similarity between a probe and a gallery image, so they propose to encode the k-reciprocal features using the Jaccard distance for re-ranking matches in the ranking list. In [13], Sarfraz et al. exploit the idea of the expanded cross neighbourhood distance between images for refining the initial ranking list to retrieve more true matches at the top. In [20], Liu et al. propose an adaptive iterative re-ranking approach based on the k-reciprocal features and use a deep feature fusion method to exploit diverse information in the learned features for matching. Although these methods have shown effectiveness for improving the performance by re-ranking the initial ranking list, they heavily rely on the pre-trained feature extractor and the accuracy of the initial ranking list. The proposed RANGEv2 model shares the same merit as these methods in exploiting the ranking information for re-id, but the key novelty is that our method can jointly optimise the appearance feature and the ranking context information in an online trainable manner, instead of applying independent post-rank. Besides, some recent studies explore the ranking context information to embed the re-id evaluation measures into the optimisation of a distance metric [18] or to select the most discriminative video fragments for re-id matching [19]. In contrast, the proposed RANGEv2 focuses on learning a more discriminative feature representation for person re-id neither optimising an evaluation metric nor selecting video frames.

2.2. Body part based person re-id

Part-level features are capable of effectively exploiting the fine-grained information of local body regions for person matching. In recent years, body part based re-id models [1,2,8,9] have shown promising performance on challenging re-id benchmarks. In [1], Li et al. propose to refine local parts with the aid of attribute detection and extract discriminative part-based features. In [2], Luo et al. propose to align the pedestrian images using the shortest path between two sets of horizontal part features. In [9], Sun et al. use the uniform partition strategy to generate body parts to learn fine-grained features in a strong baseline model. In our work, we adopt the horizontal uniform partition strategy [9] to learn body part fine-grained information. Instead of computing the part-based feature of a person image as previous methods, we instead jointly optimise all the local part branches with a global part branch in the internal ranking stream, and concatenate these features to form the complementary internal ranking representation. This formulation not only mitigates the harm of false positive matches in the external ranking stream, but also maximises the respective advantages in a cohort and achieves better performance when ensembling these two streams together in a unified framework.

3. Methodology

3.1. Approach overview

In this work, we focus on learning a hybrid ranking representation for more effective person re-id. The overview of our RANGEv2 model is depicted in Fig. 1. The design of RANGEv2 is in a two-stream architecture, where each stream is online trainable as detailed below.

Specifically, **(I)** in the external stream, we construct each training mini-batch with T randomly selected person identity and uniformly sample S images of each identity, so there are $T\times S$ images in each mini-batch. Given a probe image x_o, the remaining images in the mini-batch are defined as the gallery samples $\{x_{j\neq o}\}_{j=1}^{T\times S}$. Then, we use a feature extractor $\mathcal{F}_A$ (*e.g.* ResNet-50 [21]) to compute an appearance feature vector v^a of each image:

$$\{v_j^a\}_{j=1}^{T\times S} = \mathcal{F}_A(\{x_j\}_{j=1}^{T\times S}) \tag{1}$$

where $\{x_j\}$ is the jth image and $\{v_j^a\}$ is the corresponding appearance feature.

Next, we use a generic distance metric (*e.g.* Euclidean distance $\mathcal{D}(\cdot)$) to measure the similarity between v_o^a and $\{v_{j\neq o}^a\}_{j=1}^{T\times S}$. We select the top-$k$ ranked candidates from the gallery $\{v_{j\neq o}^a\}_{j=1}^{T\times S}$ to construct a ranking list and employ an external ranking information learner $\mathcal{F}_E$ to encode the ranking context information in the ranking list:

$$\{v_t^e\}_{t=1}^{k} = \mathcal{F}_E(\{v_t^{ar}\}_{t=1}^{k}) \tag{2}$$

where $\{v_t^{ar}\}$ is the tth appearance feature in the ranking list, and $\{v_t^e\}$ is the tth encoded external ranking feature outputted by $\mathcal{F}_E$. In this work, we use the BiLSTM [22] model as the external ranking context information learner $\mathcal{F}_E$.

At last, we use an aggregation module $\mathcal{F}_M$ to aggregate $\{v_t^e\}_{t=1}^{k}$ and generate the external ranking representation f_{ext}:

$$f_{ext} = \mathcal{F}_M(\{v_t^e\}_{t=1}^{k}) \tag{3}$$

Meanwhile, **(II)** in the internal stream, there are $T\times S$ images in each mini-batch. Given each training image x_o, we employ the ResNet-50 to extract the global appearance feature v^g and use horizontal partition to generate m part-level features $\{v_j^p\}_{j=1}^{m}$. We concatenate all the local part-level features and the global image-level feature to form the fine-grained internal ranking representation f_{int}:

$$f_{int} = \mathcal{F}_I(\{v_j^p\}_{j=1}^{m}, v^g) \tag{4}$$

where $\mathcal{F}_I$ denotes the vector concatenation operation. At last, we use $D(\cdot)$ to measure the pairwise similarity of f_{ext} in the external ranking space and that of f_{int} in the internal ranking space, and aggregate these two types of scores for re-id at test time:

$$D_s = \alpha\mathcal{D}(f_{int}) + (1-\alpha)\mathcal{D}(f_{ext}) \tag{5}$$

where D_s is the final pairwise score for re-id matching, α is a fusion weight for balancing the internal and external ranking representations. When $\alpha = 0$, only the external ranking stream is used, while $\alpha = 1$, only the internal ranking stream is used.

Model deployment. At test time, in the external stream, we extract the appearance features of each person image in the query and the gallery, and then generate the external ranking representation based on the original ranking lists; In the internal stream, we directly extract the internal ranking features of each image. The pairwise similarities are measured using Eq. (5).

3.2. External ranking representation

Traditionally, given a probe x_o, we can retrieve candidates from the gallery and obtain a ranking list. Existing re-id methods either ignore probe-gallery specific ranking lists [4,8,9] or employ offline re-ranking strategies [12,13], but fail to optimise the ranking context information from the gallery along with the deep appearance feature computation. To this end, we formulate an external ranking stream to encode the ranking context information more effectively for re-id.

In a training mini-batch, we begin by extracting an appearance feature vector of each sample using Eq. (1). And then, for x_o, we select the top-k candidates from its corresponding gallery set (*i.e.* all in-batch samples except x_o) as its ranking list $\{v_t^{ar}\}_{t=1}^{k}$, where $k > S$. Thus, ideally, we can have $S-1$ true positive samples and $k-S+1$ false positive samples in each ranking list for a given probe image. As each ranking list contains both true and false positive matches,

the proposed model can well learn to encode the latent structured ranking information of each task, which is closer to the re-id deployment where true and false positive matches are mixed in a ranking list with the plausible variations latently included. We formulate $\{v_t^{ar}\}_{t=1}^k$ as an input sequence to a BiLSTM model as:

$$\begin{cases} h_t = \mathcal{F}_{L_f}(W_1 v_t^{ar} + W_2 h_{t-1}) \\ h'_t = \mathcal{F}_{L_f}(W_3 v_t^{ar} + W_4 h'_{t+1}) \\ v_t^e = \mathcal{F}_{L_o}(W_5 h_t + W_6 h'_t) \end{cases} \tag{6}$$

where $\{W_j\}_{j=1}^6$ are the shared weights between each input unit, $\mathcal{F}_{L_f}$ is the hidden layer function, $\mathcal{F}_{L_o}$ is the output layer function, and h_t and h'_t are the forward and backward hidden states, respectively.

Remarks. The forward and backward hidden states effectively encapsulate the ranking order information for a given probe, so the outputs can learn the correlations and the discriminative selections among the input units (elements in the list). In our design, $\{v_t^e\}_{t=1}^k$ contain the output features from the last layer of the BiLSTM network, so $\{v_t^e\}_{t=1}^k$ are formulated as the latent *external ranking ensemble* feature vectors.

To obtain the external ranking representation f_{ext}, we discuss the choices for the aggregation module function $\mathcal{F}_M$.

Average Pooling. Since $\{v_t^e\}_{t=1}^k$ encode the correlations among samples of a ranking list, a natural choice for implementing $\mathcal{F}_M$ is by average pooling:

$$f_{ext} = \frac{1}{k}\sum_{t=1}^{k} v_t^e \tag{7}$$

Neighbour Weighted Pooling. Average pooling considers each candidate in a ranking list as equally important, but since these candidates are neighbours of a probe in the gallery, we can reassign the weights based on the pairwise similarity between a probe and its neighbours as:

$$\beta_t = \frac{e^{-\mathcal{D}(v_o^a, v_t^{ar})}}{\sum_{l=1}^{k} e^{-\mathcal{D}(v_o^a, v_l^{ar})}} \tag{8}$$

$$f_{ext} = \sum_{t=1}^{k} \beta_t v_t^e \tag{9}$$

where β_t is the neighbour weight for the tth encoded external ranking feature v_t^e.

Attentive Weighted Pooling. While neighbour weighted pooling reassigns the weights based on the neighbour similarity, it is limited in learning the correlations among candidates. To this end, we reassign the weights based on attention selection [17]. We use two fully connected layers to generate the 1-dimension scalar value for each encoded feature vector and employ a softmax function $\mathcal{F}_S$ to generate the attentive weights for aggregation:

$$\gamma = \mathcal{F}_S(W_8 max(0, (W_7\{v_t^e\}_{t=1}^k + b_7) + b_8)) \tag{10}$$

$$f_{ext} = \gamma\{v_t^e\}_{t=1}^k \tag{11}$$

where W_7 and W_8 are weights of the fully connected layers, b_7 and b_8 are corresponding biases, γ contains the attentive weight for the tth encoded external ranking feature v_t^e. The above variants of the aggregation module are evaluated in Section 4.

3.3. Internal ranking representation

While the external ranking stream can effectively learn the variation and correlation between a probe image and the gallery set, a ranking list may be a mixture of true and false positive matches. This is likely to contaminate the external ranking ensemble representation. As part-based appearance deep features [8] encode more effective fine-grained information, they are complementary to the external ranking representation and therefore can be used to minimise false positive matches in the gallery. Therefore, we further develop an internal ranking stream to learn the complementary part-based features and to jointly refine the ranking results.

Specifically, given x_o, we first use ResNet-50 to extract the global appearance feature v^g. Among many strategies for obtaining the regional features (such as uniform partitioned parts [9], saliency parts [3], skeleton parts [23] and attention parts [17]), we select the horizontal partition, which is both simple and effective, to generate m parts $\{v_j^p\}_{j=1}^m$ and use Eq. (4) to extract the part-level fine-grained internal ranking representation. Here, different from existing methods, the extracted part-based representations are used as complementary features to mitigate the harm of false gallery positive matches inevitable in the external ensembles.

3.4. Optimisation objective

To optimise each stream, we stack the classification layers for the appearance feature extractor, the external ranking stream, the internal part branches, and the internal global branch (see Fig. 1), to compute their classification losses. In particular, each classification module consists of two fully connected layers and a softmax operation. The first layer is a bottleneck layer with BatchNorm, while the second layer maps each feature vector to a probability vector. In the external ranking stream, the training objective is formulated as:

$$\begin{aligned} \mathcal{L}_{ext} &= \mathcal{L}^A_{soft\max} + \mathcal{L}^E_{soft\max} \\ &= -\frac{1}{T\times S}\sum_{i=1}^{T\times S} y_i \log \frac{\exp\left(W_c^A(v_i^a)\right)}{\sum_{n=1}^{N}\exp\left(W_n^A(v_i^a)\right)} \\ &\quad -\frac{1}{T\times S}\sum_{i=1}^{T\times S} y_i \log \frac{\exp\left(W_c^E(f_{ext,i})\right)}{\sum_{n=1}^{N}\exp\left(W_n^E(f_{ext,i})\right)} \end{aligned} \tag{12}$$

where $\mathcal{L}^A_{softmax}$ is the per-image classification loss, and $\mathcal{L}^E_{softmax}$ is the ranking-list-level classification loss, $\mathcal{L}_{ext}$ specifies the external ranking ensemble loss, y_i denotes the ground truth distribution, W_c^A and W_c^E compute logits with the classification layers in the external stream. In the internal ranking stream, the training objective is formulated as:

$$\begin{aligned} \mathcal{L}_{int} &= \sum_{j=1}^{m}\mathcal{L}^{I_j}_{soft\max} + \mathcal{L}^G_{soft\max} \\ &= -\frac{1}{T\times S}\sum_{j=1}^{m}\sum_{i=1}^{T\times S} y_i \log \frac{\exp\left(W_c^I(v_{j,i}^p)\right)}{\sum_{n=1}^{N}\exp\left(W_n^I(v_{j,i}^p)\right)} \\ &\quad -\frac{1}{T\times S}\sum_{i=1}^{T\times S} y_i \log \frac{\exp\left(W_c^G(v_i^g)\right)}{\sum_{n=1}^{N}\exp\left(W_n^G(v_i^g)\right)} \end{aligned} \tag{13}$$

where $\mathcal{L}^{I_j}_{softmax}$ is the local part-level classification loss, $\mathcal{L}^G_{softmax}$ is the global image-level classification loss, $\mathcal{L}_{int}$ denotes the internal ranking representation learning loss, y_i denotes the ground truth distribution, W_c^I and W_c^G compute logits with the classification layers in the internal stream. The entire training process of our model is summarised in Algorithm 1.

4. Experiment

4.1. Datasets and evaluation protocol

Datasets. To evaluate the proposed method for person re-id, we used four large-scale benchmarks (*i.e.* Market-1501 [24], DukeMTMC-ReID [25], CUHK03 [12,26] and MSMT17 [27]). We adopted the standard split setting and evaluation protocol. The statistics of these re-id benchmarks along with their split settings are shown in Table 1 and some examples from these benchmarks are shown in Fig. 2.

Specifically, **Market-1501** is a large-scale re-id benchmark with 1501 person identities and 32,668 images captured from 6 non-overlapping camera views. We followed the standard split setting [24] for experiments, *i.e.* using 751 identities for training and

Algorithm 1 Hybrid Ranking Representation for Person Re-ID.

Input: Training samples $\mathcal{X}$,Identity class labels $\mathcal{Y}$.
Output: Learned RANGEv2 model θ, includingthe appearance feature extractor θ^A,the external ranking information learner θ^E,and the internal stream model θ^I.

(I) External ranking stream
Initialise: Construct a probe-gallery mini-batch with $T \times S$ images,
for epoch=1→Max-epoch **do**
  for t=1→Batch-number **do**
    Forward to extract the appearance features v^a (Eq. (1));
    Compute the per-image classification loss $\mathcal{L}^A_{softmax}$ (Eq. (12));
    Compute the pairwise similarity to generate ranking lists;
    Forward to learn the ranking context information (Eq. (6));
    Extract the external ranking feature f_{ext} (Eq. (7)–(11));
    Compute the ranking context loss $\mathcal{L}^E_{softmax}$ (Eq. (12));
    Update $\{\theta^A, \theta^E\}$ with Eq. (12);
  end for
end for
(II) Internal ranking stream
Initialise: Construct a mini-batch with $T \times S$ images,
for epoch=1→Max-epoch **do**
  for t=1→Batch-number **do**
    Forward to extract the internal feature f_{int} (Eq. (4));
    Compute the internal ranking loss $\mathcal{L}_{int}$ (Eq. (13));
    Update θ^I with Eq. (13);
  end for
end for
return $\theta = \{\theta^A, \theta^E, \theta^I\}$

Table 1
Data statistics of four person re-id benchmarks.

Benchmark	Image	ID	Train	Test
MSMT17	126,441	4101	1041	3060
Market-1501	32,668	1501	751	750
DukeMTMC-ReID	36,411	1404	702	702
CUHK03 (new protocol)	14,097	1467	767	700

(a) Market (b) Duke (c) CUHK03 (d) MSMT17

Fig. 2. Example person image pairs from four person re-id benchmarks.

the remaining 750 identities for testing. **DukeMTMC-ReID** is a subset re-id benchmark of the DukeMTMC benchmark containing 1404 identities and 36,411 images capture by 8 cameras. Following [25], we used 702 identities with 16,522 images for training and 702 identities with 2228/17,661 images for query/gallery testing. **CUHK03** contains 1467 identities with 14,097 images. It provides both manually labelled and DPM-detected bounding boxes for experiments. The traditional protocol [26] uses 1376/100 identities for training/testing, while the new protocol uses 767/700 identities for training/testing. We used the new and more challenging protocol with both labelled (CUHK03-L) and detected (CUHK03-D) datasets in our experiments. **MSMT17** is a newly released large-scale re-id benchmark with 126,441 images and 4101 identities. We used 1041 identities for training and 3060 identities for testing as [27].

Evaluation Metrics. For all experiments, we used the *single query* evaluation. We adopted the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as the performance evaluation metrics.

4.2. Implementation details

We used ResNet-50 [21] pre-trained on ImageNet as the backbone CNN model. **(I) For the external stream**, we employed the backbone model as the appearance feature extractor and used BiLSTM [22] as the ranking context information learner. We used the Stochastic Gradient Descent (SGD) as the optimiser with the initial learning rate at 0.005, which decayed by 0.1 every 40 training epochs (initial learning rate for classification layers was set to 0.05). The model was trained for 100 epochs. We set $T = 8$ and $S = 4$ to construct the training probe-gallery batch, and set $k = 5$, *i.e.* selected top-5 candidates in a ranking list. The number of forward-backward recurrent layers was set to 1 and the feature dimension of hidden states was set to 256, so the external ranking feature representation was 512-D. By default, we reported the results using the attentive weighted pooling, while the experiments for comparing different aggregation module variants are provided in Section 4.7. **(II) For the internal stream**, we partitioned the outputted feature map from the backbone model into $m = 6$ parts, so there were m local part branches and one global branch in this stream. We set the same optimiser learning rate as the external stream and trained the model for 60 epochs. The output feature dimension of each branch was 2048-D, leading to a 14336-D internal ranking feature representation. We set $\alpha = 0.25$ as the fusion weight.

4.3. Model component evaluation and analysis

To comprehensively evaluate the effectiveness of each component of the proposed method, we conducted the experiments on all the four benchmarks. From Table 2, we can see that both the external and internal streams can improve the performance of the backbone model (*i.e.* ResNet-50) separately, and their combination (*i.e.* the full RANGEv2 model) achieves the best accuracies in most cases. Specifically, on Market-1501, DukeMTMC-ReID and MSMT17, the external stream performs better than the internal stream, which indicates that the ranking context information is relatively more effective. However, on CUHK03-D and CUHK03-L, an opposite phenomenon is observed. This suggests somewhat data-dependent advantages of the proposed two designs. Compared with the preliminary method (RANGE [16]), RANGEv2 gains nearly 10% improvements on MSMT17, CUHK03-L and CUHK03-D, and around 5% improvements on Market-1051 and DukeMTMC-ReID in the mAP score. Such great gains are attributed to the online optimisation of each ingredient stream, in addition to their improved designs.

4.4. Supervised learning re-id evaluation

Comparison with the State-of-the-Art Methods. In Table 3, we compared RANGEv2 with recently proposed re-id methods with state-of-the-art performance. As shown in Table 3, RANGEv2 achieves superior performance over the state-of-the-art alternatives [4,7,28,29]. Specifically, on Market-1501, RANGEv2 and JDGL [4] achieve the best performance. In terms of mAP, RANGEv2 ranks the first (86.8%) with a 0.8% margin over the state-of-the-art. In terms of rank-1 accuracy, RANGEv2 achieves the second-best result (94.7%) approaching to the best result (94.8%) obtained

Table 2
Component effectiveness evaluation.

Component	Market		Duke		CUHK03-D		CUHK03-L		MSMT17	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
RANGE [16]	81.9	91.0	69.7	81.3	57.0	52.9	56.2	53.0	41.0	68.7
Backbone	73.5	90.4	60.4	77.9	40.8	41.8	43.0	43.2	35.4	66.8
External	86.1	92.1	76.2	81.6	55.4	49.4	58.3	52.4	**54.4**	72.4
Internal	78.9	92.8	71.7	85.6	61.5	**65.7**	64.1	**66.9**	44.0	71.0
RANGEv2	**86.8**	**94.7**	**78.2**	**87.0**	**64.6**	61.6	**67.4**	64.3	51.3	**76.4**

Table 3
Comparison with the state-of-the-art methods on Market-1501, DukeMTMC-ReID, MSMT17 and CUHK03.The best results are shown in **RED BOLD**, while second-best in **BLUE BOLD**.

Methods	Ref.	Market		Duke		MSMT		CUHK03			
								Labelled		Detected	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
Reranking[12]	CVPR17	70.4	81.4	-	-	-	-	40.3	38.1	37.4	34.7
DaF [14]	BMVC17	72.4	82.3	-	-	-	-	31.5	27.5	30.0	26.4
PCB [9]	ECCV18	77.4	92.3	66.1	81.8	40.4	68.2	-	-	54.2	61.3
PCB+RPP [9]	ECCV18	81.6	93.8	69.2	83.3	-	-	-	-	57.5	**63.7**
MLFN [28]	CVPR18	74.3	90.0	62.8	81.0	37.0	66.3	**49.2**	**54.7**	47.8	52.8
PLNet [3]	TIP19	69.3	88.2	-	-	-	-	-	-	-	-
APR [5]	PR19	66.9	87.0	55.6	73.9	-	-	-	-	-	-
AlignedReID++ [2]	PR19	79.1	91.8	69.7	82.1	43.7	69.8	-	-	**59.6**	61.5
HPM [8]	AAAI19	82.7	94.2	74.3	**86.6**	-	-	-	-	57.5	**63.9**
BTASB [29]	CVPRW19	85.9	94.5	**76.4**	86.4	-	-	-	-	-	-
VPM [7]	CVPR19	80.8	93.0	72.6	83.6	-	-	-	-	-	-
JDGL [4]	CVPR19	**86.0**	**94.8**	74.8	**86.6**	**52.3**	**77.2**	-	-	-	-
HAN [17]	IJCV19	76.7	91.6	64.1	80.6	46.2	66.2	46.1	46.5	45.5	47.5
Deep-Person [10]	PR20	79.6	92.3	64.8	80.9	-	-	-	-	-	-
APDR [1]	PR20	80.1	93.1	69.7	84.3	-	-	-	-	-	-
RANGEv2	Ours	**86.8**	**94.7**	**78.2**	**87.0**	**51.3**	**76.4**	**67.4**	**64.3**	**64.6**	61.6

by JDGL. On DukeMTMC-ReID, RANGEv2 performs the best in both rank-1 accuracy and mAP. It clearly improves the state-of-the-art by 1.8% (78.2%-76.4%) on mAP and 0.4% (87.0%-86.6%) on rank-1 accuracy. On MSMT17, JDGL achieves the best performance (52.3% on mAP and 77.2% on rank-1 accuracy) while RANGEv2 achieves the second-best results (51.3% and 76.4% in terms of mAP and rank-1 accuracy, respectively). On CUHK03 (labelled), RANGEv2 achieves 67.4% on mAP and 64.3% on rank-1 accuracy, significantly outperforming the state-of-the-art. On CUHK03 (detected), RANGEv2 ranks the first in terms of mAP (64.6%) by improving the state-of-the-art by 5.0%, and achieves a competitive rank-1 accuracy (61.6%).

Comparison with Re-Ranking Based Re-ID. One key idea of the proposed method is to exploit the ranking context information for more effective re-id. Traditional re-ranking methods exploit this information by only considering the top-ranked candidates in the gallery and revising the scores of candidates to refine the ranking list, i.e. taking a post-rank strategy. Different from existing alternative methods, our model is a joint learning approach for optimising the appearance features and the ranking context features concurrently. As shown in Table 4, compared with the state-of-the-art re-ranking methods, RANGEv2 achieves significantly better performance. For example, on Market-1501, RANGEv2 improves the state-of-the-art by 13.6% (86.8%-73.2%) on mAP and 12.3% (94.7%-82.4%) in the rank-1 accuracy, respectively. We attribute this margin to two reasons: (I) The online joint optimisation of the appearance feature and the ranking context information representation, and (II) the complementary external and internal ranking feature representations. Interestingly, compared with the state-of-the-art re-id methods which use the re-ranking strategies (*e.g.* PSE + ECN, PSE + K-reciprocal, PCB + ARR-DFF, AlignedReID++ + K-reciprocal, and HAN + K-reciprocal), RANGEv2 still achieves the best mAP and rank-1 accuracy. This validates that the proposed external stream in RANGEv2 can more effectively learn the ranking context information. Critically, RANGEv2 is compatible with existing re-ranking methods. As shown in the last row in Table 4, K-reciprocal can work well together with RANGEv2 (adjust weights for each component), which further improves the performance of RANGEv2 by 4.5% (91.3%-86.8%) in mAP and 0.4% (95.1%-94.7%) in rank-1 accuracy on Market-1501, and by 6.0% (84.2%-78.2%) in mAP and 1.7% (88.7%-87.0%) in rank-1 accuracy on DukeMTMC-ReID. The small rank-1 margin implies that RANGEv2 is effective for encoding the ranking context information even without using re-ranking.

Comparison with Body Part Based Re-ID. The proposed method is built on an observation that body part based internal ranking representations are complementary to the ranking-list-based external ensemble representations. In RANGEv2, we adopt the uniform partition for design simplification. In comparison, existing state-of-the-art methods use various partition strategies, including uniform partition [2,8,9,32], skeleton partition [23], saliency detection [3], attention parts [17,31] and at-

Table 4
Comparison with the state-of-the-art re-ranking based re-id methods.

Method	Ref.	Strategy	Market		Duke	
			mAP	R1	mAP	R1
K-NN [30]	CVPR12	Offline	60.3	79.5	-	-
SCA [15]	TIP16	Offline	68.9	79.8	-	-
DaF [14]	BMVC17	Offline	72.4	82.3	-	-
ARR-DFF [20]	arXiv18	Offline	73.2	82.4	63.7	68.3
ECN [13]	CVPR18	Offline	71.1	82.3	-	-
K-reciprocal [12]	CVPR17	Offline	70.4	81.4	-	-
RANGEv2	Ours	**Online**	**86.8**	**94.7**	**78.2**	**87.0**
PSE+ECN [13]	CVPR18	Offline	84.0	90.3	79.8	85.2
PSE+K-reciprocal [13]	CVPR18	Offline	83.5	90.2	78.9	84.4
PCB+ARR-DFF [20]	arXiv18	Offline	85.6	91.8	80.2	86.4
AlignedReID+ + K-reciprocal [2]	PR19	offline	89.4	92.8	82.8	86.2
HAN+K-reciprocal [17]	IJCV19	offline	89.6	93.1	81.3	84.6
Backbone+K-reciprocal	Ours	Offline	89.1	92.8	80.4	84.2
RANGEv2+K-reciprocal	Ours	**Online**+Offline	**91.3**	**95.1**	**84.2**	**88.7**

Table 5
Comparison with the state-of-the-art part-based re-id methods.

Method	Ref.	Strategy	Market		Duke	
			mAP	R1	mAP	R1
Spindle [23]	CVPR17	skeleton	-	76.9	-	-
AACN [31]	CVPR18	attention	66.9	85.9	59.3	76.8
CA3Net [32]	ACMMM18	uniform	80.0	93.2	70.2	84.6
PCB [9]	ECCV18	uniform	77.4	92.3	66.1	81.8
PCB+RPP [9]	ECCV18	uniform	81.6	93.8	69.2	83.3
PLNet [3]	TIP19	saliency	69.3	88.2	-	-
AlignedReID+ [2]	PR19	uniform	79.1	91.8	69.7	82.1
HAN [17]	IJCV19	attention	76.7	91.6	64.1	80.6
HPM [8]	AAAI19	uniform	82.7	94.2	74.3	86.6
APDR [1]	PR20	attribute	80.1	93.1	69.7	84.3
RANGEv2	Ours	uniform	**86.8**	**94.7**	**78.2**	**87.0**

tribute parts [1]. As shown in Table 5, compared with the state-of-the-art part-based re-id methods, RANGEv2 achieves significantly better performance. In particular, on Market-1501, RANGEv2 improves the state-of-the-art by 4.1% (86.8%-82.7%) in mAP and 0.5% (94.7%-94.2%) in rank-1 accuracy, while on DukeMTMC-ReID, RANGEv2 improves the state-of-the-art by 3.9% (78.2%-74.3%) in mAP and 0.4% (87.0%-86.6%) in rank-1 accuracy. These results demonstrate that the ranking-list-based features are complementary to part-based fine-grained features, consequently the fused ranking feature representations bring superior performance.

4.5. Unsupervised cross-domain re-id evaluation

To further evaluate the effectiveness of our method, we conducted unsupervised cross-domain re-id experiments. Supervised re-id assumes the availability of person identity class labels in the target domain, however, this supervision is not necessarily accessible in many real-world deployments. One effective solution is by unsupervised cross-domain knowledge transfer [33–36]. This task is more challenging because the model trained on the source domain is typically weak when directly transferred to the unseen target domain, due to the potentially significant domain discrepancy. As a consequence, more false matches will be retrieved in the top ranks.

To demonstrate that the proposed RANGEv2 is effective in unsupervised re-id where more false matches are retrieved in the top ranks, we formulate a training pipeline as shown in Fig. 3. Specifically, we first pretrain the proposed RANGEv2 in a source labelled domain. Then, in the target unlabelled domain, we use the pretrained model to extract the initial ranking feature representations of N_u unlabelled samples. Based on the fusion scores, we use adaptive clustering (*e.g.* DBSCAN [37]) to gather n_u samples into j_u clusters, where $j_u \ll n_u < N_u$. Subsequently, we reassign the pseudo labels for the target unlabelled domain based on the clusters, with $N_u - n_u$ isolated samples discarded. Next, we fine-tune the pretrained model to optimise the ranking feature representations in the target domain using the triplet loss [38], resulting in the final model for deployment. For simplification and computation efficiency, we only use a single clustering and fine-tune process, rather than performing more expensive iterative clustering as [33].

Table 6
Evaluation on unsupervised cross-domain person re-id. '*': The target domain is Market-1501 but the source domain involves multiple datasets instead of DukeMTMC-ReID. **D2M**: DukeMTMC (source) ⇒ Market (target). **M2D**: Market (source) ⇒ DukeMTMC (target).

Methods	Ref.	D2M		M2D	
		mAP	R1	mAP	R1
PUL [33]	TOMM18	20.1	44.7	16.4	30.4
PTGAN [27]	CVPR18	-	38.6	-	27.4
SPGAN [36]	CVPR18	22.8	51.5	22.3	41.4
SPGAN+LMP [36]	CVPR18	26.9	58.1	26.4	**46.9**
TJ-AIDL [34]	CVPR18	26.5	58.2	23.0	44.3
HHL [35]	ECCV18	31.4	**62.2**	**27.2**	**46.9**
CSGLP [39]	TIFS19	31.5	61.2	27.1	**47.8**
DECAMEL* [40]	TPAMI20	32.4	60.2	-	-
RANGE [16]	Ours	**32.5**	58.5	21.8	34.6
Backbone	Ours	25.2	52.9	15.3	28.7
RANGEv2	Ours	**33.5**	**61.8**	**27.4**	41.7

As shown in Table 6, RANGEv2 obtains significant improvements (around 10%) in mAP and rank-1 accuracy as compared with the backbone model. This indicates that the proposed ranking feature representation remains effective for the more challenging unsupervised cross-domain person re-id. Moreover, compared with the state-of-the-art methods [35,39,40], when using DukeMTMC-ReID as the source domain and Market-1501 as the target domain (*i.e.* DukeMTMC-ReID ⇒ Market-1501), RANGEv2 ranks the first in mAP (33.5%) and the second in rank-1 (61.8%). For the Market-1501 ⇒ DukeMTMC-ReID case, RANGEv2 yields the best mAP (27.4%) and obtains a competitive rank-1 accuracy (41.7%).

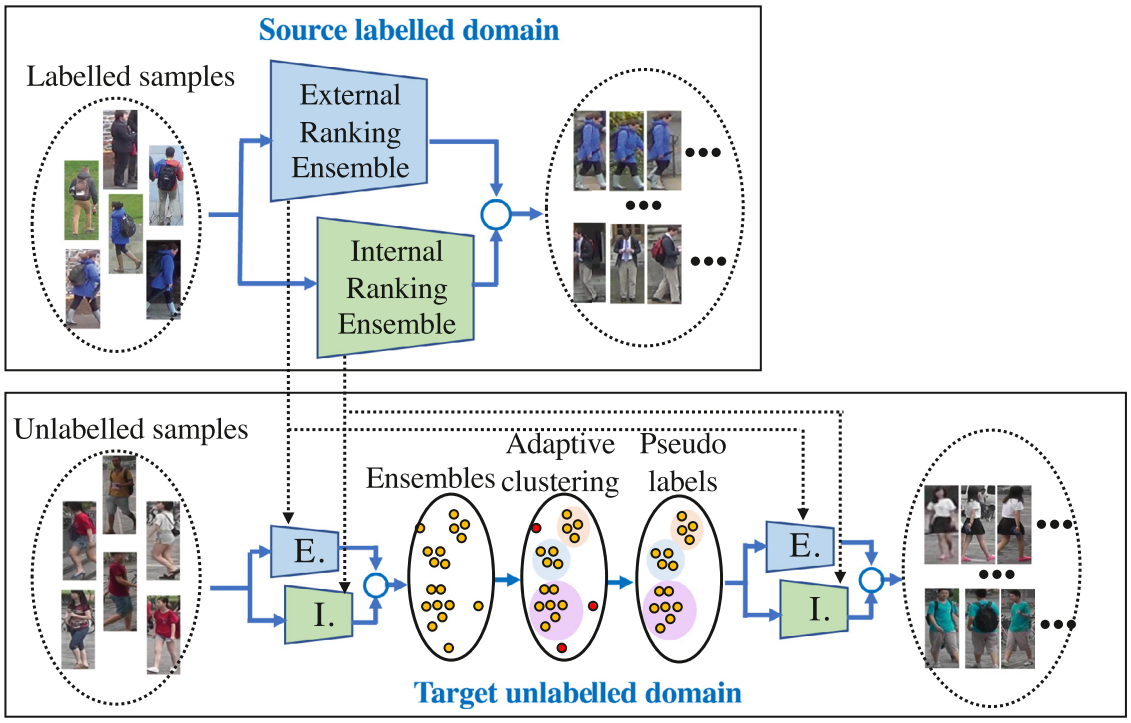


Fig. 3. Illustration of RANGEv2 for unsupervised cross-domain person re-id.

Fig. 4. Qualitative person re-id examples. The first block presents the probe persons, followed by the ranking results of the backbone model (middle) and the proposed RANGEv2 model (right). The images with **green**/**red** bounding boxes are the true/false matches of the corresponding probe person. Candidates in the ranking lists produced by RANGEv2 are more accurate and have more consistency in appearance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.6. Qualitative retrieve evaluation

To qualitatively evaluate the performance of RANGEv2, we visualise several examples of the person re-id results in Fig. 4. We have two main observations as follows:

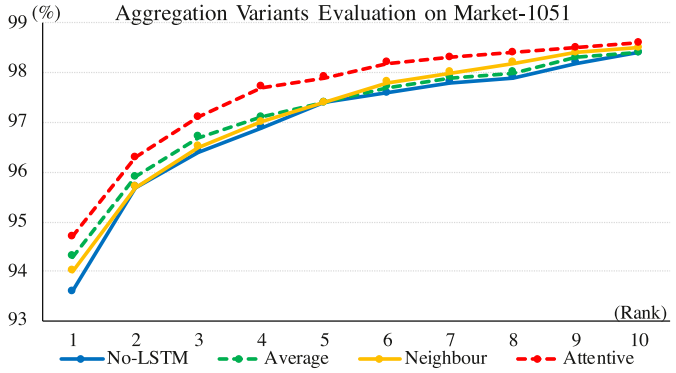
(I) In each ranking list, with the backbone model, the retrieved results are generated mostly by the appearance similarity between the probe and each gallery candidate independently, with some candidates in the top ranks visually very dissimilar to the others (see the first and fourth rows). In contrast, the retrieved results of the proposed RANGEv2 not only accurately account for the appearance similarity of probe-gallery pairs, but also well encode the latent structured correlations among candidates in a ranking list. As a result, RANGEv2 can capture more true candidates on the top ranks, with each candidate being more consistent to one another. These results demonstrate that the hybrid ranking representation produced by our RANGEv2 are able to reliably learn the ranking context information, therefore delivering better performance.

(II) As shown in the first, second and fifth rows, the backbone model is incapable of learning discriminative representations of the probe images, where it retrieves false top-1 results and more false results in each ranking list. In comparison, RANGEv2 not only retrieves the true candidates in the first rank, but also retrieves more true candidates in the ranking lists. This shows that the proposed ranking representation is more effective and discriminative for person re-id.

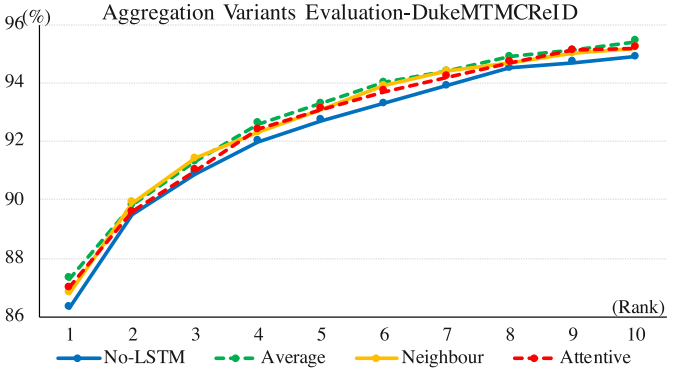
4.7. Parameter analysis

Aggregation Module Variant Comparison. To evaluate the effectiveness of different aggregation module variants in the external ranking stream, we conduct the experiments for comparative evaluation among average pooling, neighbour weighted pooling, attentive weighted pooling and no-LSTM. Here, the former three variants refer to those we introduce in Section 3.2, while no-LSTM means directly pooling the appearance features without using the external ranking information learner and the aggregation module. As shown in Fig. 5, no-LSTM performs slightly worse, suggesting the effectiveness of the external ranking information learner and the aggregation module. On Market-1501, the attentive variant performs significantly better than the other two aggregation designs, while on DukeMTMC-ReID, three aggregation variants achieve very similar performance.

Ranking List Length Impact. We evaluate the impact of ranking list length k on the re-id performance. Here, the length $k = 0$ means only using the internal ranking stream. Fig. 6 shows that: (I) As k extends from 0 to 20, the rank-1 accuracies improve until the peaks (94.7% on Market-1501 and 87.0% on DukeMTMC-ReID) at $k = 5$ followed by steady decreases. This indicates that (1) the ranking context information can be effectively encoded into the ranking representation to improve re-id performance, (2) but as more false positive matches are included, the discrimination capability of a best-matched candidate in a gallery can be contaminated, resulting in the slight deterioration of the rank-1 accuracy. (II) For the mAP performance, as k increases from 0 to 5, the rate significantly increases from 78.9% to 86.8% on Market-1501 and from 71.7% to 78.2% on DukeMTMC-ReID. This demonstrates that the learning ranking context cue is an important strategy for finding more true positive matches on the top ranks. From $k = 5$ to $k = 20$, mAP gradually reaches the peaks at $k = 10$ followed by a steady decrease trend. This trend is reasonable and expected, because higher mAP suggests more true positive matches are retrieved; Therefore, more candidates in a ranking list can improve the generalisation over true positive candidates in a gallery. However, when setting a large length (*e.g.* $k = 20$), there will be more false positive matches in a ranking list, which intuitively leads to slight performance deterioration.



(a) Market-1501



(b) DukeMTMC-ReID

Fig. 5. Evaluation on aggregation module variants.

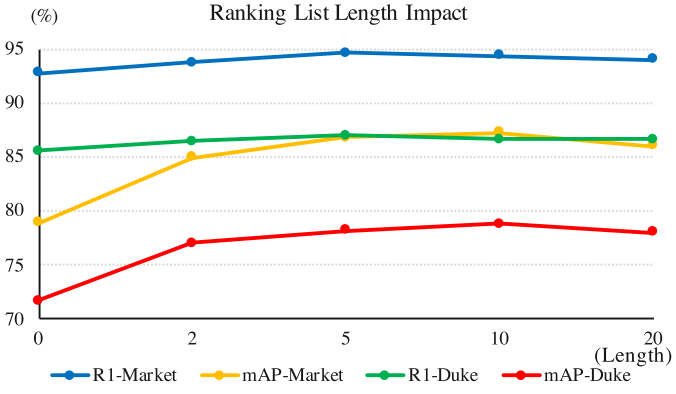


Fig. 6. Evaluation on ranking list length.

Fusion Weight Impact. To evaluate the fusion weight α, which is used to balance the internal and external ranking ensembles for re-id, we conduct experiments with α varies from 0 to 1. Here, $\alpha = 0$ means only the external ranking stream is used, while $\alpha = 1$ means the opposite – only the internal ranking stream is used. We can see from Fig. 7 that: (I) The internal and external ranking streams perform differently in terms of mAP and rank-1 accuracy. Specifically, in mAP, the external stream performs better than the internal counterpart, which indicates that the ranking context information is more important for retrieving more positive candidates in a gallery; In rank-1 accuracy, the internal stream performs better, indicating that part-based fine-grained features are more

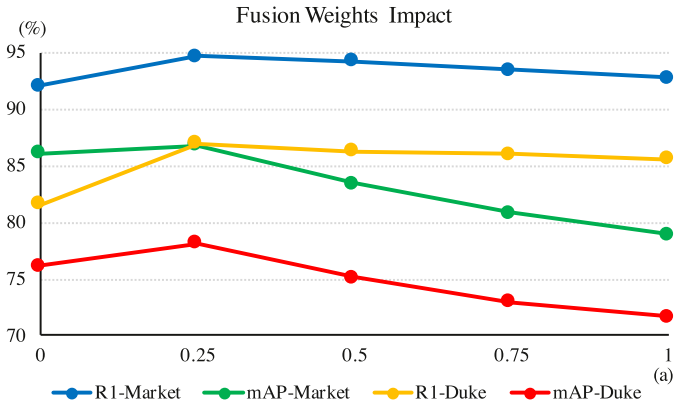


Fig. 7. Evaluation on fusion weight.

favourable to search a best-matched candidate in a gallery. (II) The fusion of internal and external ranking representations achieves better overall performance than using any one alone. When setting $\alpha = 0.25$, RANGEv2 achieves the best performance (in terms of mAP and rank-1 accuracy, 86.8% and 94.7% on Market-1501, and 78.2% and 87.0% on DukeMTMC-ReID, respectively). These results demonstrate that the internal and external ranking representations are well complementary.

5. Conclusion and future work

In this work, we proposed a novel idea of jointly learning the ranking context information and appearance feature representation for person re-identification (re-id). It significantly differs from existing re-id methods either ignoring the ranking context information (*e.g.* PCB [9]) or applying post-rank strategies (*e.g.* K-reciprocal [12]). To realise the proposed idea, we formulated a novel two-stream person re-id architecture (RANGEv2) for learning the hybrid external and internal ranking representations. RANGEv2 exploits the ranking contextual information with the external stream and extracts part-based fine-grained information with the internal stream. Integrating them into a unified architecture enables to maximise their respective advantages in a cohort for improving the re-id performance. Extensive experiments on four large-scale re-id benchmarks have clearly shown the superiority of the proposed method over a wide range of state-of-the-art supervised and unsupervised re-id methods. We also conducted a spectrum of detailed component analysis for giving the model formulation insights.

RANGEv2 also has some limitations to be solved in future work. First, using a two-stream architecture, the number of parameters in RANGEv2 is approximately doubled. Distilling the knowledge in RANGEv2 into a smaller model is an effective solution. Second, RANGEv2 is not sufficiently effective for addressing domain discrepancy problem in unsupervised person re-id. How to further refine the false matches due to the domain discrepancy with the aid of additional information (such as attribute [1,5] and camera labels [35]) is worth more investigation. Our future work will focus on addressing these limitations and developing our RANGEv2 for real-world applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by Queen Mary University of London Principal's Scholarship, Vision Semantics Limited, Alan Turing Institute Turing Fellowship, and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

[1] S. Li, H. Yu, R. Hu, Attributes-aided part detection and refinement for person re-identification, Pattern Recognit. 97 (2020) 107016.
[2] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, C. Zhang, Alignedreid++: dynamically matching local information for person re-identification, Pattern Recognit. 94 (2019) 53–61.
[3] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, Q. Tian, Deep representation learning with part loss for person re-identification, IEEE Trans. Image Process. 28 (6) (2019) 2860–2871.
[4] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2138–2147.
[5] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognit. 95 (2019) 151–161.
[6] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Multi-type attributes driven multi-camera person re-identification, Pattern Recognit. 75 (2018) 77–89.
[7] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, J. Sun, Perceive where to focus: learning visibility-aware part-level features for partial person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 393–402.
[8] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019.
[9] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: European Conference on Computer Vision, 2018, pp. 480–496.
[10] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, Deep-person: learning discriminative deep features for person re-identification, Pattern Recognit. 98 (2020) 107036.
[11] C.-X. Ren, X.-L. Xu, Z. Lei, A deep and structured metric learning method for robust person re-identification, Pattern Recognit. 96 (2019) 106995.
[12] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3652–3661.
[13] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 420–429.
[14] R. Yu, Z. Zhou, S. Bai, X. Bai, Divide and fuse: a re-ranking approach for person re-identification, in: British Machine Vision Conference, 2017.
[15] S. Bai, X. Bai, Sparse contextual activation for efficient visual re-ranking, IEEE Trans. Image Process. 25 (3) (2016) 1056–1069.
[16] G. Wu, X. Zhu, S. Gong, Person re-identification by ranking ensemble representations, in: Proceedings of the IEEE International Conference on Image Processing, 2019.
[17] W. Li, X. Zhu, S. Gong, Scalable person re-identification by harmonious attention, Int. J. Comput. Vis. (2019) 1–19.
[18] S. Paisitkriangkrai, C. Shen, A. Van Den Hengel, Learning to rank in person re-identification with metric ensembles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1846–1855.
[19] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: European Conference on Computer Vision, 2014, pp. 688–703.
[20] Y. Liu, L. Shang, A. Song, Adaptive re-ranking of deep feature for person re-identification, arXiv preprint arXiv:1811.08561(2018).
[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[22] A. Graves, S. Fernández, J. Schmidhuber, Bidirectional LSTM networks for improved phoneme classification and recognition, in: International Conference on Artificial Neural Networks, 2005, pp. 799–804.
[23] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1077–1085.
[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[25] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3754–3762.
[26] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.

[27] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer GAN to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 79–88.
[28] X. Chang, T.M. Hospedales, T. Xiang, Multi-level factorisation net for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2109–2118.
[29] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bags of tricks and a strong baseline for deep person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[30] X. Shen, Z. Lin, J. Brandt, S. Avidan, Y. Wu, Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3013–3020.
[31] J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, Attention-aware compositional network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2119–2128.
[32] J. Liu, Z.-J. Zha, H. Xie, Z. Xiong, Y. Zhang, CA3Net: contextual-attentional attribute-appearance network for person re-identification, in: Proceedings of the ACM International Conference on Multimedia, 2018, pp. 737–745.
[33] H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: clustering and fine-tuning, ACM Trans. Multimed. Comput. Commun. Appl. 14 (4) (2018) 83.
[34] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2275–2284.
[35] Z. Zhong, L. Zheng, S. Li, Y. Yang, Generalizing a person retrieval model hetero-and homogeneously, in: European Conference on Computer Vision, 2018, pp. 172–188.
[36] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, J. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 994–1003.
[37] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Proceedings of the AAAI Conference on Artificial Intelligence, 1996, pp. 226–231.
[38] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737(2017).
[39] C.-X. Ren, B. Liang, P. Ge, Y. Zhai, Z. Lei, Domain adaptive person re-identification via camera style generation and label propagation, IEEE Trans. Inf. Forensics Secur. 15 (2019) 1290–1302.
[40] H.-X. Yu, A. Wu, W.-S. Zheng, Unsupervised person re-identification by deep asymmetric metric embedding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 956–973.

Guile Wu is working toward the Ph.D. degree at Queen Mary University of London. His research interests include person re-identification, visual object tracking, and deep learning.

Xiatian Zhu was a researcher of Vision Semantics Limited. He received his Ph.D. from Queen Mary University of London. He won The Sullivan Doctoral Thesis Prize 2016, an annual award representing the best doctoral thesis submitted to a UK University in computer vision. His research interests include computer vision and machine learning.

Shaogang Gong is Professor of Visual Computation at Queen Mary University of London (since 2001), a Fellow of the Institution of Electrical Engineers and a Fellow of the British Computer Society. He received his D.Phil (1989) in computer vision from Keble College, Oxford University. His research interests include computer vision, machine learning and video analysis.