Discriminative Self-Paced Group-Metric Adaptation for Online Visual Identification

Jiahuan Zhou, Bing Su, and Ying Wu, Fellow, IEEE

Abstract—Existing solutions to instance-level visual identification usually aim to learn faithful and discriminative feature extractors from offline training data and directly use them for the unseen online testing data. However, their performance is largely limited due to the severe distribution shifting issue between training and testing samples. Therefore, we propose a novel online group-metric adaptation model to adapt the offline learned identification models for the online data by learning a series of metrics for all sharing-subsets. Each sharing-subset is obtained from the proposed novel frequent sharing-subset mining module and contains a group of testing samples that share strong visual similarity relationships to each other. Furthermore, to handle potentially large-scale testing samples, we introduce self-paced learning (SPL) to gradually include samples into adaptation from easy to difficult which elaborately simulates the learning principle of humans. Unlike existing online visual identification methods, our model simultaneously takes both the sample-specific discriminant and the set-based visual similarity among testing samples into consideration. Our method is generally suitable to any off-the-shelf offline learned visual identification baselines for online performance improvement which can be verified by extensive experiments on several widely-used visual identification benchmarks.

Index Terms—Learning from sharing, frequent pattern mining, online adaptation, person re-identification, self-paced learning

1 Introduction

The goal of visual identification is to retrieve the same identity images of a query probe from a gallery set. As an attractive research task in the computer vision community, visual identification has attracted increasing attention owing to its importance as a critical link to practical public camera surveillance applications. Over the past years, a popular solution to visual identification is performing supervised discriminative feature learning [1], [2], [3], [4], [5], [6], [7] from the given offline training data, then directly applying the learned models to the online unlabeled testing data for evaluation. However, due to the severe training-testing data distribution shifting (testing data are drawn from totally different classes against the training data as shown in Fig. 1) caused by large variations in visual appearance, object pose, camera viewpoint, illumination change, and background clutter, the performance of offline learned models is indeed limited. Moreover, this performance degradation is even more critical when an instance-level visual identification problem (e.g., person re-identification (P-RID), vehicle re-identification (V-RID), instance discrimination learning, etc) is considered. Since different instances from the same category in the training and testing sets are considered as different individual classes, extreme divergences between training and testing data caused by large intra-instance variations may result in a significant performance drop of the offline learned models. As demonstrated by Fig. 1, regardless of which visual identification benchmarks or state-of-the-art methods are selected, the critical training-testing distribution shifting issue always exists.

To narrow such a distribution gap between training and testing samples, a straightforward solution is adapting the offline learned models to fit the online testing data. Recently, various online visual identification methods are proposed which can be roughly categorized into two branches. The set-centric re-ranking approaches [9], [10], [11], [12], [13] focus on optimizing the ranking list of queries based on the similarity relationships among testing samples. Their performance relies on the offline models learned from training data, and treating different testing samples equally largely ignores the individual characteristics, hence the improvement is neither significant nor stable. The other category is query-specific feature adaptation [8], [14], [15] which aims to enhance the feature discriminant of each query individually that the generic offline learned feature is adapted to an instance-specific local feature for each query. Compared with the set-centric ones, the individual discriminant of each query is enhanced while the

Jiahuan Zhou is with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China. E-mail: jiahuanzhou@pku.edu.cn.
Bing Su is with the Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China. E-mail: subingats@gmail.com.
Ying Wu is with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208 USA. E-mail: yingwu@northwestern.edu.

Manuscript received 25 April 2021; revised 17 January 2022; accepted 16 August 2022. Date of publication 19 August 2022; date of current version 6 March 2023.
This work was supported in part by National Science Foundation under Grant IIS-1815561 and IIS-2007613, in part by the National Natural Science Foundation of China under Grants 61976206 and 61832017, in part by Beijing Outstanding Young Scientist Program under Grant BJJWZYJH012019100020098, in part by the Beijing Academy of Artificial Intelligence (BAAI), in part by the Fundamental Research Funds for the Central Universities, in part by the Research Funds of Renmin University of China under Grant 21XNLG05, in part by the Public Computing Cloud, Renmin University of China, in part by Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, and in part by the Public Policy and Decision-making Research Lab of Renmin University of China.
(Corresponding author: Bing Su.)
Recommended for acceptance by B. Leibe.
Digital Object Identifier no. 10.1109/TPAMI.2022.3200036

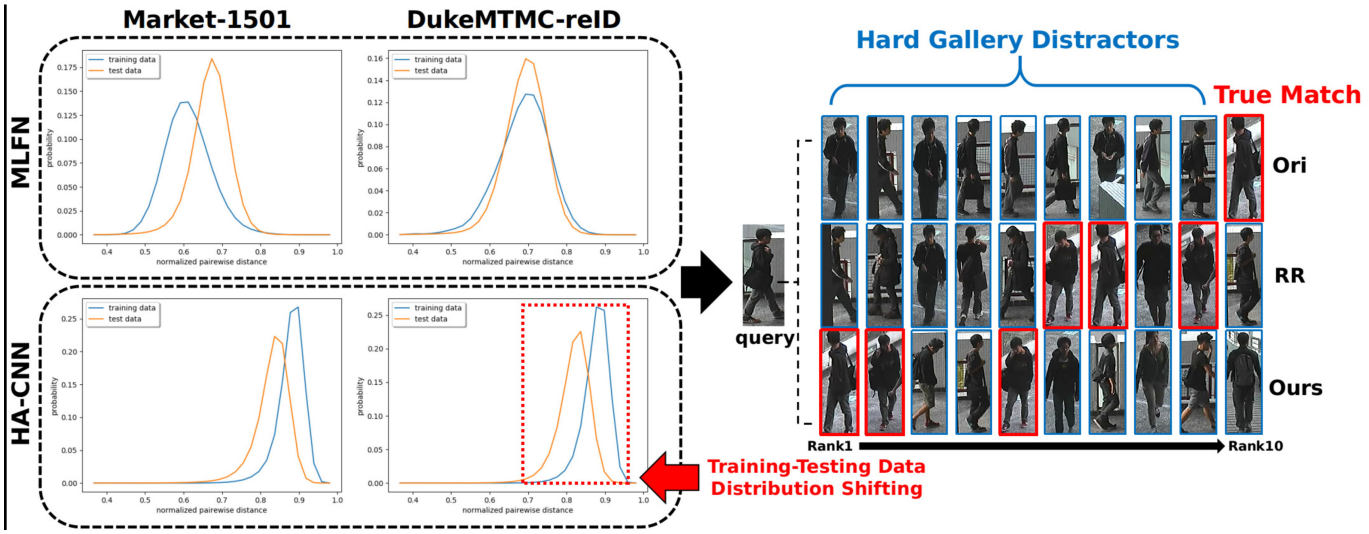


Fig. 1. Taking P-RID task as an example, the normalized pair-wise distance distributions of both training and testing samples based on the well-trained state-of-the-art HA-CNN and MLFN networks on the Market1501 and DukeMTMC-reID datasets are presented. The results demonstrate the severe training-testing data distribution shifting issue, where the extremely challenging hard negative distractors (in blue box) will significantly influence the retrieval accuracy (the Original top-10 retrieval results). Even using the state-of-the-art online re-ranking method [8] (RR), the ground-truth (in red box) still has a lower rank than the distractors. Our method succeeds in handling the distractors so that the true-match is successfully re-ranked to the top position in the list (Ours).

visual similarity relationships among given testing samples are ignored. Moreover, existing query-specific models [8], [14], [15] completely ignore the counterpart gallery data during adaptation. Even a discriminative probe-specific metric can be learned, the "hard" gallery samples with large intra-class and small inter-class variances will tremendously degrade its performance since they are still indistinguishable under the learned query-specific metric. From the efficiency perspective, existing query-specific adaptation methods suffer from heavy online computation costs since they have to repeatedly and individually handle each testing sample for adaptation, and such computational burden is even severe when a large-scale testing set is given.

To mitigate the aforementioned issues, we propose a novel online self-paced group-metric adaptation (SPGMA) algorithm which not only takes individual characteristics of testing samples into consideration but also fully explores the visual similarity relationships among all query and gallery samples. As illustrated by Fig. 2, at the online identification stage, the redundant intrinsic visual similarity relationships among the unlabeled query (gallery) set are utilized by the proposed frequent sharing-subset (SSSet) mining algorithm to automatically mine concise and salient visual sharing associations of samples. Since a sharing-subset contains a group of testing samples that share strong visual similarities, their local distributions can be jointly adjusted by efficient metric adaptations for all of them. Furthermore, to readily handle large-scale testing samples (especially hundreds of thousands of gallery instances), we introduce a self-paced learning strategy [16], [17], [18] to gradually include testing samples into adaptation from easy to difficult. Thus, by iterating between our proposed unsupervised frequent sharing subset mining and online self-paced SSSet selection algorithms, much fewer group-metric adaptations will be learned and the online optimization could be more efficient since fewer testing samples are used in each learning iteration. Once a series of such kinds of SSSet-based metrics are learned, for each query (gallery), its instance-specific local metric is obtained via a multi-metric late fusion of all the group-metrics.

Therefore, our proposed online SPGMA model can significantly refine the ranking performance, and the success of learning from sharing relies on discovering the latent sharing relationships among samples, which cannot be found by treating each instance independently [19]. Learning from sharing is good at handling such conditions that only a limited number of positive learning data are available by taking the sharing relationships as data augmentation. Therefore the sharing strategy is particularly suitable for the learning of online instance-level visual identification in where each testing sample itself is the only positive sample available for learning. To sum up, our contributions are as follows: 1) To handle the severe shifted training-testing data distribution issue in visual identification, we leap from off-line global learning to online instance-specific adaptation. 2) By automatically mining various frequent sharing-subsets, the intrinsic visual similarity relationships among testing samples can be fully explored via a self-paced SSSet selection strategy to gradually adapt sharing-subsets to fit the learned group-metrics. Therefore, both superior online re-ranking performance and efficient learning from sharing merits can be achieved. 3) Our proposed model can be readily applied to any existing offline visual identification baselines for online performance improvement. Therefore, our appealing efficiency-and-effectiveness superiority is not only verified by extensive experiments on various P-RID and image retrieval benchmarks based on various state-of-the-art visual identification models but also guaranteed by several theoretically sound justifications.

This manuscript is an extension of our previous conference paper [20], while we have made a lot of extensions including 1) To facilitate the online computation cost and further improve identification performance, a classic self-paced learning algorithm is explored to gradually include testing

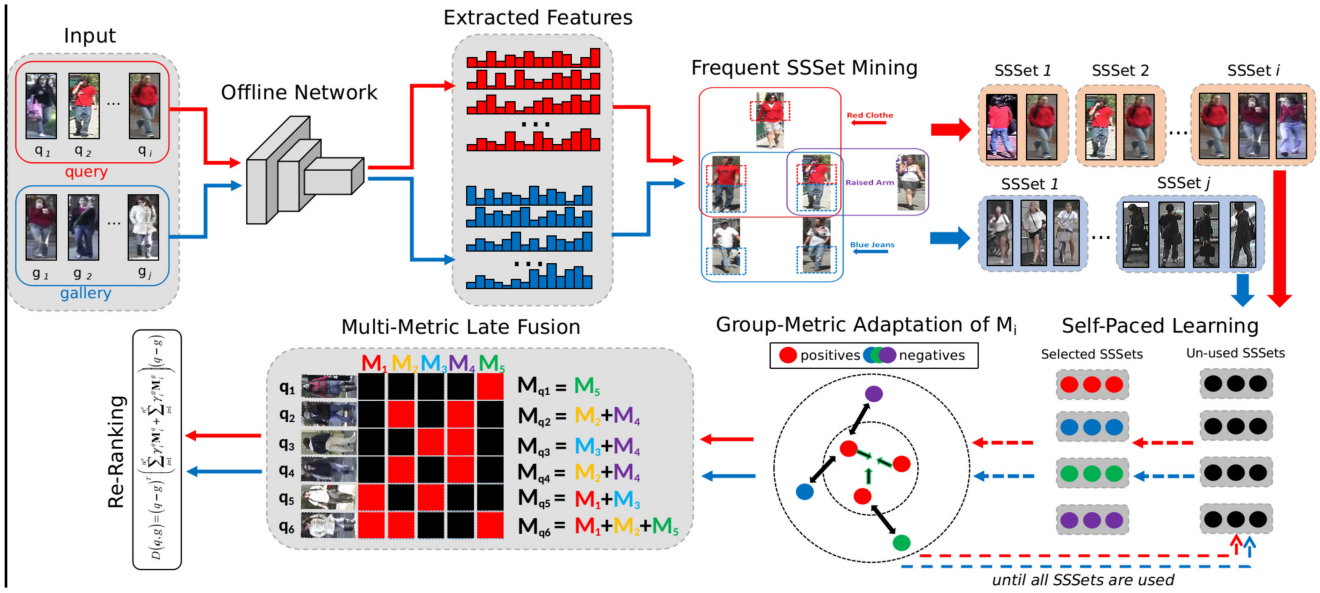


Fig. 2. The online testing query and gallery samples are first fed into the offline learned network to extract feature descriptors. The proposed frequent sharing-subset (SSSet) mining algorithm is performed to generate multiple sharing-subsets which are further utilized by the self-paced SSSet selec-tion algorithm to iteratively determine which SSSets will be involved for each learning round. The "easy" SSSets will be processed first so that the model can be accordingly trained from scratch. When the model can handle these "easy" SSSets well, the "harder" ones will be gradually involved to further improve the effectiveness of the model. Within each learning round, the selected SSSets will be fed into the proposed online group-metric adaptation model for local discriminant enhancement (The same sample can be contained by multiple SSSets since it may share different visual sim-ilarity relationships with different samples). Such learning continues until all the obtained SSSets are processed. Finally, by fusing the learned group-metrics for each query and gallery sample, our final ranking list is obtained by a bi-directional retrieval matching.

samples into adaptation from easy to difficult which elaborately simulates the learning principle of humans. 2) Thorough theoretical analyses are provided, and the solution to special testing sample conditions is discussed to complete the proposed method. 3) Compared with [20] which only conducts experiments under the P-RID setting, we further evaluate our method on a completely different but challenging visual identification task, image retrieval. Extensive experiments are conducted on four widely-used image retrieval benchmarks and promising improvement is obtained compared with the state-of-the-art baselines. 4) More ablation experiments (e.g., affinity matrix refinement visualization, re-ranking improvement results, online computation cost comparison, etc) are conducted to further investigate our proposed method.

2 Related Work

2.1 Local Metric Learning

To facilitate visual identification, several discriminative local metric learning methods are proposed. To tackle the multi-modal distributions of identity appearances, Zhang et al. [21] utilized the local distance comparison in P-RID to obtain an accurate retrieval. A regularized local metric learning (RLML) method was proposed by Liong et al. [22] handle the common over-fitting issue in visual identification via exploring the merits of both the global and local metrics. A sample-specific SVM classifier is learned in Zhang et al. [15] for each training sample, then the weight parameters of a testing sample can be inferred. In order to relax the requirement of a large-number labeled images for learning, a novel one-shot learning approach is proposed by Bak et al. [1] which only requires a single image from each camera for training, thus the learning result is specific to the only sample. However, these local metric learning methods still perform an offline global-learning procedure that heavily relies on labeled training data. Their performance is indeed limited if testing data are from different distributions. Instead, our method adopts an online local adaptation manner to adapt the offline learned baselines to each testing sample specifically.

2.2 CNN-Based Feature Extraction

CNN-based feature extraction has achieved state-of-the-art performance in visual identification. A novel Harmonious Attention CNN (HA-CNN) proposed by Li et al. [3] tries to jointly learn attention selection and feature representation in a CNN by maximizing the complementary information of different levels of visual attention (soft attention and hard attention). Wang et al. [4] proposed a novel deeply supervised fully attentional block that can be plugged into any CNNs to solve visual identification, and a novel deep network called Mancs is designed to learn stable features. Chen et al. [5] proposed an Attentive but Diverse Network (ABD-Net) which integrates attention modules and diversity regularization throughout the entire network to learn features that are representative, robust, and more discriminative for P-RID. Zheng et al. [6] aimed at improving the learned fea-to tures by better leveraging the generated data by designing a joint learning framework that couples feature learning and data generation end-to-end. Li et al. [23] proposed a Feature-Fusing Graph Neural Network (FFGNN) to utilize the relationships among the nearest neighbors of the given training images for feature learning. A self-critical attention learning (SCAL) method is proposed by Chen et al. [24] to generate both spatial-wise and channel-wise attention for discriminative identification. To strengthen discriminative features and

suppressing irrelevant ones in visual identification, Zhang et al. [25] designed an effective Relation-Aware Global Attention (RGA) module to capture the global structural information for better attention learning.

Almost all the aforementioned instance identification methods focus on learning discriminative metrics or features from the offline training data to facilitate the matching. When their models are well trained offline, they will not modify the model any more and directly use them for the unseen testing data. However, the data distribution shifting between training and testing samples largely limits the performance of these models. To tackle this issue, our proposed method is suitable for any CNNs for sample-specific local metric adaptation at the inference stage aiming to well handle the data shifting issue and gain further performance improvement.

2.3 Online Re-Ranking

In recent years, increasing efforts have been paid to online re-ranking in visual identification. Ye et al. [9] revised the ranking list by considering the nearest neighbors of both the global and local features. An unsupervised re-ranking model proposed by Garcia et al. [10] takes advantage of the content and context information in the ranking list. Zhong et al. [11] proposed a k-reciprocal encoding approach for re-ranking, which relies on a hypothesis that if a gallery image is similar to the probe in the k-reciprocal nearest neighbors, it is more likely to be a true-match. Zhou et al. [8] proposed to learn an instance-specific Mahalanobis metric for each query sample by using extra negative learning samples at the online stage. Barman et al. [12] focused on how to make a consensus-based decision for retrieval by aggregating the ranking results from multiple algorithms, only the matching scores are needed. Fan et al. [26] proposed a progressive unsupervised learning method to transfer pre-trained deep representations to unseen domains for unsupervised P-RID. Bai et al. [13] concentrated on re-ranking with the capacity of metric fusion for retrieval by proposing a unified ensemble diffusion framework. However, the aforementioned online re-ranking methods either simply treat different testing samples equally without considering the instance-specific characteristics or completely ignore the intrinsic visual similarity relationships among testing samples. Therefore, their performance improvement is neither stable nor significant.

2.4 Self-Paced Learning

Self-paced learning (SPL), designed through simulating the learning principle of humans/animals, becomes a popular research topic in recent years. To alleviate the heuristic easiness measure requirement, Kumar et al. [16] proposed to re-formulate the key principle of Curriculum Learning as a concise SPL model. Jiang et al. [17] extended SPL by considering the diversity of samples selected in the training steps. Kamran et al. [18] proposed a balanced SPL model in the designed generative adversarial clustering network by considering an unsupervised loss based on the adjacency matrix. Recently, studies [27], [28] focusing on adopting SPL in the interested tasks to avoid getting stuck in bad local minima and improving the generalization of their models are proposed and attracted increasing attention. In recent years, several self-paced learning-based works are studied in the person re-identification area. Xin et al. [29] proposed a semi-supervised P-RID method by utilizing a small portion of labeled training samples to fine-tune a CNN model, and then propagating the labels to the unlabeled portion for further fine-tuning the overall system in a self-paced manner. Zhou et al. [30] proposed a self-paced constraint and symmetric regularization to help the relative distance metric training the deep neural network, so as to learn the stable and discriminative features for identification. Ge et al. [31] proposed a novel self-paced contrastive learning framework with hybrid memory which can generate different level of supervision signals for different domains to facilitate identification.

Although the aforementioned methods explore SPL to facilitate offline learning, the requirement of labeled data prevents their usage in the online testing phase. Also, the pseudo-label noise propagation and training-testing distribution gap may still result in severe performance degradation of these SPL-based methods. In this work, to tackle the potential large-scale testing data, we introduce SPL to gradually involve samples into adaptation from easy to difficult. Hence the testing samples can be better adapted to the learned group-metrics.

3 Online Self-Paced Group-Metric Adaptation From Sharing

3.1 Problem Settings and Notations

At the online testing stage of visual identification, two disjoint datasets, a query set Q and a gallery set G are given as:

$$Q = \{q_i; l_i^q\}_{i=1}^{n_q}, \quad G = \{g_i; l_i^g\}_{i=1}^{n_g},$$

that $q_i; g_i \in R^d$ are the extracted feature representations from an offline baseline model, either handcraft designed or learned deep features. $l_i^q; l_i^g \in \{1; 2; \ldots; c\}$ are the labels from c classes which are totally different from the training sample classes. We aim to rank G for a query q based on the pairwise similarity distance to a gallery g, $d(q; g) = \|q - g\|^2$ and our goal is to re-rank G for q by refining $d(q; g)$ to improve the rank of true-matches for q.

3.2 Unsupervised Frequent Sharing-Subset Mining

Although the identity labels l_i^q, l_i^g are unknown during online testing, the visual similarity relationships of Q and G are intrinsic and verified to be effective in investigating the underlying similarity structure of samples by previous re-ranking methods [10], [11]. However, due to the large-scale testing sample size (especially for G), the redundancy and repeatability of visual similarity relationships significantly limit the performance of previous online re-ranking methods. Inspired by the well-established frequent itemset mining technique [32], we propose an unsupervised frequent sharing-subset (SSSet) mining algorithm to automatically mine various SSSets $\{S_i\}_{i=1}^{n_s}$ from Q and G, that all the samples in S_i share a Strong Association Rule on visual similarity [32]. Therefore, the mined SSSets not only keep the strong and reliable visual similarity sharing information but also significantly alleviate the redundancy issue. Compared with the originally combinatorial problem suffering from exponential complexity $O(2^n)$, the time complexity of our proposed algorithm is $O(n^2)$ which is much more efficient when large-scale testing samples are given. It is worth

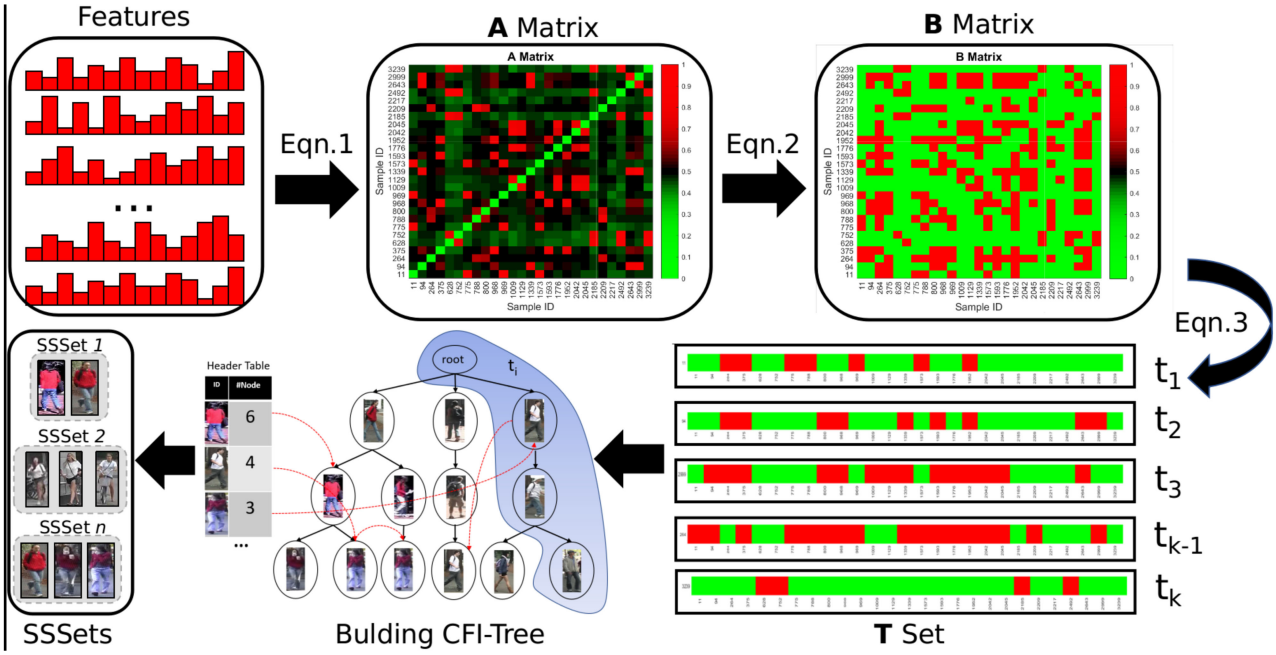


Fig. 3. The pipeline of our proposed unsupervised frequent sharing-subset mining algorithm. Given the extracted features of testing samples, the affinity matrix A can be computed. To keep only the most reliable sharing relationships, a threshold Q is used for filtering so that a binary index map B is obtained. Then each non-zero row B_j of B can be considered as an element t_i in the set T. Moreover, T is utilized to build a CFI-Tree that is the input for the off-the-shelf FP-Close mining algorithm to mine all the frequent SSSets.

mentioning that [32] is a classic method that utilizes frequent itemset mining to handle the classification tasks but it only focuses on mining the low-level local textural features from the patches in one image. Thus [32] is difficult to handle the higher-level visual similarity relationships between different images. Our proposed method focuses on mining the frequent SSSets based on the reliable and strong visual relationships obtained from the whole image globally.

Algorithm 1. Building CFI-Tree from T

Require: The given query set Q and obtained set T
Ensure: A CFI-tree
1: For all the given n_q testing samples in Q, we firstly index them from 1 to n_q;
2: Count how many times the given n_q testing samples are contained by the elements in T and sort all the samples in decreasing order of their count;
3: Create the root node (null);
4: Scan the set T, get the elements of length 1 (e.g., the length of B_i is 1 if there is only one non-zero element in the), and sort these elements in decreasing support count;
5: Load an element in T at a time. Sort the samples in this element according to the last step;
6: For each element in T, insert its samples to the constructed Tree from the root node and increment occurrence record at every inserted node;
7: Create a new child node if reaching the leaf node before the insertion completes;
8: If a new child node is created, link it from the last node consisting of the same item;
9: Return the constructed tree as the final CFI-Tree

The overall pipeline of our proposed unsupervised frequent SSSet mining algorithm is illustrated in Fig. 3. Taking the query set Q as an example, we first prepare a set $T = \{t_i\}_{i=1}^{n_t}$ from Q where each t_i is a subset of Q. The affinity matrix $A \in R^{n_q \times n_q}$ of Q is defined as:

$$A_{i,j} = \begin{cases} \exp\left(\frac{d(q_i,q_j)}{2s}\right) \Big/ \sum_{j \neq i} \exp\left(\frac{d(q_i,q_j)}{2s}\right), & j \neq i \\ 0, & j = i \end{cases} \tag{1}$$

where s is the variance parameter of distance matrix from Q so that $A_{i,j}$ represents the soft-max normalized visual similarity between q_i and q_j. The i-th row of A represents the similarity distribution between q_i and the other samples in Q. To keep only the most reliable sharing relationships, a threshold Q defined as the average affinity of Q is used for outlier filtering $Q = \sum_{i=1}^{n_q} \sum_{j=1}^{n_q} A_{i,j} / n_q^2$. Therefore, a binary index map B is obtained by:

$$B_{i,j} = \begin{cases} 1, & A_{i,j} \geq Q \\ 0, & A_{i,j} < Q \end{cases} \tag{2}$$

The non-zero $B_{i,j}$ implies the strong similarity sharing relationship between q_i and q_j. Therefore each non-zero row B_j of B can be considered as an element t_i in set T:

$$T = \{t_i\} = \{B_j\}, \forall k B_j k_1 > 0 \tag{3}$$

Once set T is obtained, we propose to mine the frequent SSSets from T that each sharing-subset is represented by a mined frequent pattern from a classical FP-Close mining algorithm [33]. To do so, a Closed Frequent Itemset Tree (CFI-Tree) as shown in Fig. 4 is firstly constructed following the Algorithm 1 under a minimum support 5. Finally, the obtained CFI-tree will be fed into the off-the-shelf FP-Close mining algorithm [33] to mine all the closed frequent patterns $\{S_i\}_{i=1}^{n_s}$ as demonstrated by Algorithm 2.

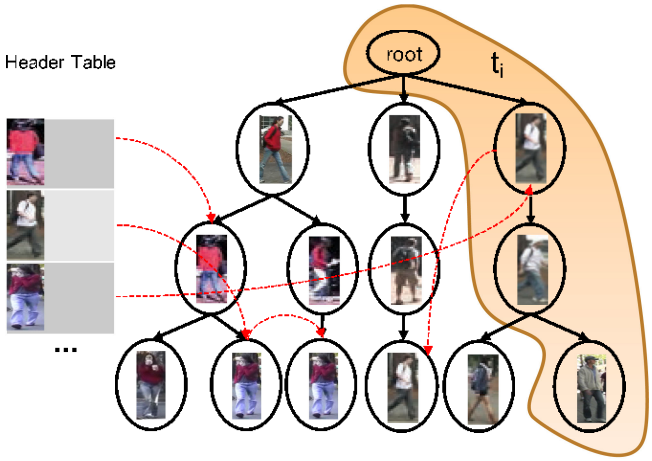


Fig. 4. A CFI-Tree is constructed based on T . The same identity may be contained by multiple t_i so that there may be multiple nodes for the same identity.

Algorithm 2. Frequent Sharing-Subset Mining

Require: The given query set Q
Ensure: n_s sharing-subsets $\{S_i\}_{i=1}^{n_s}$
1: Compute the affinity matrix A of Q by Eq. (1);
2: Compute the threshold Q;
3: Compute the binary index map B by Eq. (2);
4: Obtain the transaction set T by Eq. (3);
5: Build the CFI-Tree following Algorithm 1;
6: Frequent sharing-subset mining via the FP-Close mining algorithm in [33] based on the obtained CFI-Tree;
7: Return $\{S_i\}_{i=1}^{n_s}$

3.3 Efficient Group-Metric Adaptation

Once all the SSSets $\{S_i\}_{i=1}^{n_s}$ are obtained, our goal is to jointly learn n_s SSSets-based local Mahalanobis metrics for $\{S_i\}_{i=1}^{n_s}$ by optimizing Eq. (4) aiming to collapse the same-SSSet samples together meanwhile push the different-SSSet samples far away:

$$\begin{aligned}&\arg\min_{\{M_i\succeq 0\}} \frac{1}{2}\sum_{i=1}^{n_s}\|M_i\|^2 \\ &\text{w.r.t. } (s_u^i - s_v^j)^T M_i (s_u^i - s_v^j) \geq 2, \forall s_u^i \in S_i, s_v^j \in S_j, \\ &\quad (s_u^i - s_v^j)^T M_i (s_u^i - s_v^j) \leq 0, \forall s_u^i \in S_i, s_v^j \in S_i\end{aligned} \tag{4}$$

The learned metric M_i from Eq. (4) is shared by all the samples in S_i. Suppose we have n_s SSSets and $O(n)$ samples in each S_i, there are totally $O(n^2 n_s^2)$ inequality constraints and $O(n_s n^2)$ equality constraints in Eq. (4) which are too many to deal with. Thus we aim to reduce the constraint size in Eq. (4) by revealing that Eq. (4) has an exactly equivalent form by only keeping the constraints related to one anchor sample s^i in S_i (s^i can be any samples in S_i). The equivalent form is shown by Eq. (5):

$$\begin{aligned}&\arg\min_{\{M_i\succeq 0\}} \frac{1}{2}\sum_{i=1}^{n_s}\|M_i\|^2 \\ &\text{w.r.t. } (s^i - s_v^j)^T M_i (s^i - s_v^j) \geq 2, \forall s^i \in S, s_v^j \in S_j \\ &\text{learning of } (s^i - s_v^i)^T M_i (s^i - s_v^i) \leq 0, \forall s^i \in S, s_v^i \in S_i\end{aligned} \tag{5}$$

Revisit Eq. (4), its equality constraints propose to collapse all $s_u^i \in S_i$ together. Therefore keeping only the equality constraints related to the anchor sample s^i achieves the same collapsing performance. So as to the inequality constraints in Eq. (4). Finally, we can reduce the constraint size by only keeping the constraints related to s^i as in Eq. (5). The reformed objective Eq. (5) has only $O(n_s^2 n)$ and $O(n_s n)$ inequality and equality constraints respectively. An important merit of Eq. (5) is that it can be efficiently optimized:

Theorem 1. All the vectors $s^i - s_v^i$ in Eq. (5) form a spanning space $H = \text{span}\{\sum_v (s^i - s_v^i)\}$. Eq. (5) is equivalent to replace $s^i - s_v^j$ by $h_v^?$, the projection of $s^i - s_v^j$ in $H_v^?$, that $H^?$ is the orthogonal space of H.

Proof. Since M_i is positive semi-definite, we could have: $(s^i - s_v^j)^T M_i (s^i - s_v^j) = 0 \Leftrightarrow (s^i - s_v^j) \perp M_i$, $M_i(s^i - s_v^j) = 0 \Leftrightarrow (s^i - s_v^j) \perp M_i$, $M_i h = 0, \forall h \in H$. Projecting $s^i - s_v^j$ to H and $H^?$ generates two orthogonal bases h_v and $h_v^?$ respectively, so $s^i - s_v^j = h_v + h_v^?$. Replacing inequality constraints in Eq. (5) by $h_v + h_v^?$ gives $(s^i - s_v^j)^T M_i (s^i - s_v^j) = (h_v + h_v^?)^T M_i (h_v + h_v^?) = h_v^T M_i h_v$. Now Eq. (5) has an equivalent form as:

$$\begin{aligned}&\arg\min_{\{M_i\}} \frac{1}{2}\sum_{i=1}^{n_s}\|M_i\|^2 \text{ w.r.t. } M_i \succeq 0 \\ &h_v^{?T} M_i + M_j h_v^? \geq 2, \forall s^i \in S_i, s^j \in S_j \\ &M_i h = 0, \forall h \in H\end{aligned} \tag{6}$$

□

Finally, we prove that Eq. (6) has the same solution to Eq. (4) by eliminating its PSD and equality constraints.

Theorem 2. The solution to Eq. (4) is exactly the same as solving the Eq. (6) by relaxing its equality and PSD constraints since they are indeed off-the-shelf.

Proof. If we get rid of the PSD and equality constraints in Eq. (6), the new form is:

$$\begin{aligned}&\arg\min_{\{M_i\}} \frac{1}{2}\sum_{i=1}^{n_s}\|M_i\|^2 \\ &\text{w.r.t. } h_v^{?T} M_i + M_j h_v^? \geq 2, \forall s \in S_i, s_v \in S_j\end{aligned} \tag{7}$$

Eq. (7) is exactly in the same form of a multi-kernel SVM problem so that it can be efficiently solved. Thus the positive semi-definiteness of M_i is guaranteed since $M_i = \sum_v a_v' h_v^? h_v^{?T} \succeq 0$. For the equality constraints in Eq. (6), given a member s of S, we have:

$$M_i h = \sum_v a_v h_v^? h_v^{?T} h = \sum_v a_v h_v^? (h_v^{?T} h) = 0 \tag{8}$$

which proves that the solution to Eq. (7) satisfies the equality constraints as well. □

Therefore, following the above optimization analyses, we could efficiently and jointly perform online group-metric adaptation for all the mined SSSets $\{S_i\}_{i=1}^{n_s}$.

3.4 Self-Paced Group-Metric Adaptation

Although we already significantly simplify the GMA via Eq. (7), simultaneously adapting all the testing samples still results in a sub-optimal solution since the learning

objectives of all $\{M_i\}_i$ may conflict with each other which causes unstable loss optimization. Especially for these "hard" samples that belong to different SSSets but are visual indistinguishable to each other. Besides, the optimization cost of Eq. (7) is quadratic to the number of involved learning samples so that directly handling all the testing samples at once will result in extreme optimization difficulty. Therefore, we propose to incorporate a self-paced learning strategy [16], [17], [18] into the adaptation to gradually tackle the testing samples in a from-easy-to-difficult manner as Eq. (9):

$$\min_{v_i} \sum_{i=1}^{n_s} v_i \ell_i - \lambda_v \sum_{i=1}^{n_s} v_i, \quad s.t. \; v_i \in [0, 1]^{n_s} \tag{9}$$

where $\ell_i = \sum_{j=i}^{n_s} \sum_{s_v \in S_j} \max(0, 2 - d_{s_u, s_v}^T M_i d_{s_u, s_v})$ is the hinge loss related to the i-th SSSet S_i, v_i is the self-paced learning parameter and λ_v is the weighting hyper-parame-ter for controlling the learning pace. Once the mined SSSets are determined, the closed-form solution to Eq. (9) of all $\{v_i\}_{i=1}^{n_s}$ can be readily obtained as Eq. (10):

$$v(\ell, \lambda_v) = \begin{cases} 1, & \text{if } \ell < \lambda_v \\ 0, & \text{if } \ell \geq \lambda_v \end{cases} \tag{10}$$

If $v_i = 1$, the obtained SSSet S_i will be used in this adaptation learning round and $v_i = 0$ represents S_i will not be involved in the current learning round. By gradually increasing λ_v throughout the learning, more "hard" SSSets will be included into the training process. Finally, by conducting an alternative learning between the optimization of self-paced SSSet selection in Eq. (9) and our group-metric adaptation in Eq. (4), our discriminative group-metrics for all $\{S_i\}_{i=1}^{n_s}$ can be readily obtained. Therefore, the overall algorithm of our self-paced group-metric adaptation method (SPGMA) is shown in Algorithm 3.

Algorithm 3. SPGMA: Self-Paced Group-Metric Adaptation

Require: The mined SSSets $\{S_i\}_{i=1}^{n_s}$ via Algorithm 2
Ensure: n_s adapted group-metrics $\{M_i\}_{i=1}^{n_s}$
1: Initialize all the M_i as identical matrix;
2: **for** the number of involved SSSets $< n_s$ **do**
3: Initialize all the v_i in Eq. (9) by ranking all computed ℓ_i;
4: Optimize all the v_i via closed-from solutions in Eq. (10);
5: Based on the obtained $\{v_i\}_{i=1}^{n_s}$, optimize Eq. (4) using the selected SSSets;
6: **end for**
7: Return the computed $\{M_i\}_{i=1}^{n_s}$

3.5 Multi-Metric Late Fusion for Bi-Directional Discriminant Enhancement

As we mentioned, for a query q, it may be contained by multiple SSSets so that there will be multiple learned metrics M_i associated to q. The final metric M_q for q is obtained via a boosting-form multi-metric late fusion [35], [36]:

$$M_q = \sum_{i=1}^{n_s} g_i M_i \Big/ \sum_{i=1}^{n_s} g_i \tag{11}$$

where $g_i = 1$ if $q \in S_i$. In practice, the gallery set G, the counterpart of query set Q, also plays an important role at the online testing stage. As shown by Fig. 1, the re-ranking performance by using only the query-centric metric adaptation may suffer from ambiguous gallery distractors. The similar gallery images from different identities will significantly degrade the discriminant of M_q since these gallery distractors are still indistinguishable under M_q. Therefore, we aim to handle these indistinguishable gallery samples by performing a gallery-centric local discriminant enhancement method using Algorithms 2 and 3. For a gallery sample g, a similar fused metric M_g can be obtained likewise. Therefore the refined distance between q and g is defined as Eq. (12) based on which the re-ranking list of q_i is obtained.

$$d(q, g) = (q - g)^T M_q (q - g) + (q - g)^T M_g (q - g) \tag{12}$$

3.6 Handle Special Conditions of Testing Samples

Recall our proposed SPGMA algorithm, the visual similarity sharing conveyed by the mined SSSets is the key-point. However, there may exist some special testing sample conditions that we need to specifically deal with: (1) For some testing samples that are not visually similar to all the other samples, they are excluded by all the mined SSSets. Therefore, these testing samples will not be covered by the learning of SPGMA. On the other hand, these special testing samples also indicate that they are originally separable from the other samples which motivates us to tackle them in a straightforward and simple way. For these special testing samples, by considering each of them as an individual SSSet, they can be readily involved in our SPGMA by optimizing the same Eq. (4). (2) Another critical condition is that not enough testing samples (only very few or even only one testing sample) are given at once. Under this condition, it is difficult to utilize our proposed unsupervised frequent SSSet mining algorithm to explore visual similarity relationships among testing samples. Fortunately, our SPGMA can still well handle this situation. Similarly, each of these limited testing samples will be considered as an individual SSSet and our proposed group-metric adaptation will degenerate to the form as in [8] which could also be efficiently and effectively optimized via our simplified objective Eq. (7).

4 Theoretical Analyses and Justifications

As demonstrated by Theorem 2, the solution to our SPGMA can be readily transformed into an equivalent form as [8]. Therefore, the appealing theoretical properties in [8] can be inherited by our learned M_i as presented in Theorem 3. Moreover, our late multi-kernel fusion metric Eq. (11) will guarantee a further reduction of generalization error bound as shown in Theorem 4.

Theorem 3. (The reduction of both asymptotic and practical error bound by the learned M_i): As demonstrated by the Theorem 2 in [8], for an input x, its asymptotic error $P^a(e|x)$ by using extra negative data D^a is:

$$P^a(e|x) = \frac{(2 - q)P(e|x)}{2 - 2qP(e|x)} \tag{13}$$

where q is a probability scalar that $0 \leq q \leq 1$ and $P(e|x)$ is the Bayesian error. Moreover, the asymptotic error $P^a(e|x)$ can be

best approximated by the practical error rate $P_n(ej\mathbf{x})$ (n is finite) by finding a local metric M_x which turns out to be the one for our Eq. (4).

Theorem 4. (The reduction of generalization error bound by using $M_{q=g}$ in Eq. (11)): Our fused multi-kernel metric $M_q = (\sum_{i=1}^{n_s} g_i M_i)/\sum g_i$ is a linear combinations of several base kernels M_i from the family of finite Gaussian kernels: $K_G^d := \{K_M : (x_1, x_2) \rightarrow e^{-(x_1-x_2)^T M (x_1-x_2)/2} | M \in R^{d\times d}, M \succeq 0\}$ which is bounded by B_k. Therefore, for a fixed $d \in (0, 1]$, $n_s < n_k$ is the number of metrics (kernels) involved in our final SPGMA learning solution. With probability at least $1-d$ over the choice of a random training set $X = \{x_i\}_{i=1}^n$ of size n we have:

$$E_{est}(M_i) < O\left(\sqrt{\frac{RI(P) + B_k}{n}}\right) \quad (14)$$

$$E_{est}(M_q) < O\left(\sqrt{\frac{\log n_k + B_k + 2n_s}{n}}\right) \quad (15)$$

In our work, we have $n_s \ll n_k$, that the selected number of kernels is much fewer than the total kernel number so that $E_{est}(M_q) < O\left(\sqrt{\frac{\log n_k + B_k}{n}}\right) \ll E_{est}(M_i) < O\left(\sqrt{\frac{n_k + B_k}{n}}\right)$.

The generalization error by using M_q is much smaller than using only any M. The same conclusion can be obtained for M_g likewise.

Proof. The classification rule of our learned M_i can be defined as $z_j(\theta_q^T M_i \tilde{x}_j - 1) > 1$ so that the margin is 1. Motivated by [37], the generalization error $E_{est}(M_i)$ of using kernel M_i is bounded by $O\left(\sqrt{\frac{n_k + B_k}{n}}\right)$. While by using M_q, which is a linear combination of all M_i from the family of finite Gaussian kernel K_G^d, its generalization error $E_{est}(M_q)$ is bounded by $O\left(\sqrt{\frac{\log n_k + B_k + 2n_s}{n}}\right)$ which is guaranteed by the Theorem.2 in [38]. For the kernel family K_G^d, $n_k = O(d^2)$ and in our work, $d = 10^3$ so that $n_k = 10^6$. The selected kernels for combination is about 20 in average so that $n_s \ll n_k$ which means $E_{est}(M_q) < E_{est}(M_i)$. □

5 Experiments

5.1 Experiments on P-RID

5.1.1 Settings

Datasets. We evaluate our proposed SPGMA algorithm on CUHK03 [14], Market1501 [58], DukeMTMC-reID [55], and MSMT17 [59] benchmarks. The statistic details of the above datasets are summarized in Table 1. For CUHK03 [1], the new splitting protocol proposed by [11] is adopted in our experiments so that 767 identities are used for training as well as the left 700 identities are used for testing. As for the other three benchmarks, the pre-determined query and gallery sets are directly utilized with no modification.

Baselines. Our proposed SPGMA method refers to the alternative optimization between the self-paced SSSet selection in Eq. (9) and our group-metric adaptation in Eq. (4). If we only utilize the group-metric adaptation in Eq. (4) to simultaneously optimize all the given testing samples, the obtained GMA method is our original version in [20]. In this work, our proposed GMA/SPGMA method is evaluated based on several state-of-the-art CNN-based P-models: ResNet50 [39], DenseNet121 [34], HA-CNN [3], MLFN [40] and ABDNet [5]. The general CNN models, ResNet50 and DenseNet121, are well trained on each benchmark for feature extraction. HA-CNN, MLFN and ABDNet are re-identification specific networks so that the original works are directly utilized in our experiments. Besides, other state-of-the-art P-RID methods [5], [6], [40], [41], [42], [43], [44], [45], [50], [60] are further compared. Moreover, related online P-RID methods including OL [8] and RR [11] are compared with our SPGMA.

TABLE 1
The Statistics of P-RID Benchmarks

Dataset	cuhk03	market	duke	msmt17
#T-IDs	767	751	702	1040
#Q-IDs	700	750	702	3060
#G-IDs	700	751	1110	3060
#cam	2	6	8	15
#images	28,192	32,668	36,411	126,441

#T/Q/G-IDs denote the number of training/query/gallery ids.

Evaluation. We follow the same official evaluation protocols in [55], [58], the single-shot evaluation setting is adopted and all the results are shown in the form of Cumulative Matching Characteristic (CMC) at several selected ranks and mean Average Precision (mAP).

5.1.2 Comparison With State-of-the-Arts

Evaluation on CUHK03: The comparison results on CUHK03 (767/700 splitting protocol) are presented in Table 2. Our proposed GMA model significantly boosts the baseline Rank@1 (mAP) performance of ResNet50, DenseNet12, HA-CNN and MLFN to 66.9% (60.7%), 61.6% (54.4%), 69.8% (63.5%) and 73.4% (71.2%) with a 40.0 %(29.7%), 50.2 %(35.7%), 45.4 %(33.4%) and 34.2 %(44.7%) relative improvement respectively. Even compared with the state-of-the-art method MGN [56], our results outperform it by 5% at Rank@1. The reason for such a large improvement is that the "hard" gallery distractors which are still indistinguishable under M_q are well handled by our method, so the ranking of true-match gallery targets is significantly improved. Moreover, by taking advantage of self-paced learning, the obtained SPGMA can further improve the state-of-the-art performance on all the datasets based on all baselines. The adaptation is performed from easily-handled samples to hard samples so that the obtained adaptation metrics can be gradually optimized which results in more discriminative identification performance.

Evaluation on Market1501: The superiority of our method is further verified by the experiments on Market1501. Table 2 demonstrates that although the state-of-the-art approach ABDNet [5] has achieved a pretty high performance (~94%) on Market1501, the improvement of our SPGMA is still over 4% (10%) on Rank@1 (mAP) based on ABDNet (visualization results are shown in Fig. 5).

Evaluation on DukeMTMC-reID: DukeMTMC-reID is a recent benchmark proposed for P-RID, but the latest methods

1. In our experiment, the CUHK03 detected dataset is utilized.

TABLE 2
Compared with the State-of-the-Art P-RID Methods on CUHK03, Market1501, and DukeMTMC-reID Datasets

CUHK03(767/700)			Market1501			DukeMTMC-reID		
Method	**R@1**	**mAP**	**Method**	**R@1**	**mAP**	**Method**	**R@1**	**mAP**
ResNet50 [39]	47.9	46.8	ResNet50 [39]	88.5	71.3	ResNet50 [39]	77.7	58.8
DenseNet121 [34]	41.0	40.1	DenseNet121 [34]	88.2	69.2	DenseNet121 [34]	78.6	58.5
HA-CNN [3]	48.0	47.6	HA-CNN [3]	90.6	75.3	HA-CNN [3]	80.7	64.4
MLFN [40]	54.7	49.2	MLFN [40]	90.1	74.3	MLFN [40]	81.0	62.8
ABDNet [5]	N/A	N/A	ABDNet [5]	93.7	85.5	ABDNet [5]	84.1	67.7
OSNet [41]	N/A	N/A	OSNet [41]	94.2	82.6	OSNet [41]	87.0	70.2
PCB [42]	63.7	67.5	PCB [42]	83.3	69.2	PCB [42]	83.3	69.2
SVDNet [43]	41.5	37.3	SVDNet [43]	82.3	62.1	SVDNet [43]	76.7	56.8
MobileNetv2 [44]	41.0	40.3	MobileNetv2 [44]	84.2	65.8	MobileNetv2 [44]	73.2	52.5
SuffleNet [45]	31.9	31.7	SuffleNet [45]	80.0	58.4	SuffleNet [45]	69.3	46.8
DPFL [46]	40.7	37.0	DNSL [47]	61.0	35.6	DuATM [48]	81.8	64.6
PAN [49]	36.3	34.0	Part-aligned [50]	91.7	79.6	Part-aligned [50]	84.4	69.3
ResNeXt [51]	43.8	38.7	PN-GAN [52]	77.1	63.6	PAN [49]	71.6	51.5
DaRe [53]	55.1	51.3	DeepCC [54]	89.5	75.7	GAN [55]	67.7	47.1
MGN [56]	68.0	67.4	Mancs [4]	93.1	82.3	SPreID [57]	85.9	73.3
GMA+ResNet50	**66.9**	**60.7**	**GMA+ResNet50**	**95.4**	**82.6**	**GMA+ResNet50**	**84.7**	**68.5**
GMA+DenseNet121	**61.6**	**54.4**	**GMA+DenseNet121**	**95.3**	**81.2**	**GMA+DenseNet121**	**84.9**	**68.0**
GMA+HA-CNN	**69.8**	**63.5**	**GMA+HA-CNN**	**96.5**	**85.2**	**GMA+HA-CNN**	**87.1**	**72.2**
GMA+MLFN	**73.4**	**71.2**	**GMA+MLFN**	**96.4**	**85.0**	**GMA+MLFN**	**86.5**	**71.5**
GMA+ABDNet	**N/A**	**N/A**	**GMA+ABDNet**	**97.9**	**92.6**	**GMA+ABDNet**	**87.5**	**73.3**
SPGMA+ResNet50	**67.3**	**61.0**	**SPGMA+ResNet50**	**95.8**	**82.9**	**SPGMA+ResNet50**	**85.2**	**69.0**
SPGMA+DenseNet121	**62.2**	**54.9**	**SPGMA+DenseNet121**	**95.8**	**81.9**	**SPGMA+DenseNet121**	**85.6**	**68.7**
SPGMA+HA-CNN	**70.5**	**64.1**	**SPGMA+HA-CNN**	**97.1**	**85.5**	**SPGMA+HA-CNN**	**87.8**	**72.7**
SPGMA+MLFN	**73.9**	**71.8**	**SPGMA+MLFN**	**96.9**	**85.4**	**SPGMA+MLFN**	**86.8**	**71.9**
SPGMA+ABDNet	**N/A**	**N/A**	**SPGMA+ABDNet**	**98.2**	**92.8**	**SPGMA+ABDNet**	**87.8**	**73.7**

(a) Rank Improvement on Market1501

(b) Rank Improvement on DukeMTMC-reID

Fig. 5. The visualization of rank improvement on Market1501 (a) and DukeMTMC-reID (b) based on the baseline [34]. For each case, its top-5 (left to right) matches are presented and the true-match is labeled by the red box.

have obtained promising performance. As shown in Table 2, the recently published OSNet [41] has raised the state-of-the-art to 87.0% (70.2%). Our ABDNet+SPGMA improves the Rank@1(mAP) result to 87.8%(73.7%), which beats OSNet by a large margin on mAP.

Evaluation on MSMT17: MSMT17 is the latest and largest benchmark so far which is pretty challenging due to the extreme large-scale identities and distractors. We evaluate the performance of selected baselines (a self-paced learning-based baseline Self [31] is also compared) on the MSMT17 dataset with(w/) and without(w/o) our GMA/SPGMA models in Table 3. For all the baselines, both of our models significantly improve their Rank@1 (mAP) performance. The performance of ABDNet is boosted from 82.3%(60.8%) to a state-of-the-art level of 86.0%(64.5%). Table 3 verifies the scalability of our proposed GMA/SPGMA models, even for the extremely large-scale query/gallery sets, our methods are still able to consistently improve the baseline performance.

TABLE 3
Compared with the State-of-the-Arts on MSMT17

MSMT17	Baseline		GMA		SPGMA	
	R@1	mAP	R@1	mAP	R@1	mAP
Self[31]	42.3	19.1	54.7	25.1	55.4	25.7
ResNet50[39]	63.4	34.2	72.8	55.0	73.4	55.5
DenseNet121[34]	66.0	34.6	75.5	43.1	76.2	43.9
HA-CNN[3]	64.7	37.2	74.3	43.8	74.9	44.5
MLFN[40]	66.4	37.2	72.8	43.4	73.3	44.1
ABDNet[5]	82.3	60.8	85.7	64.2	86.0	64.5

*Self is learned in an unsupervised manner.

TABLE 4
The Influence of Each Component in Our Algorithm

Method	CUHK03			Market1501			DukeMTMC-reID			MSMT17		
	R@1	R@20	mAP	R@1	R@20	mAP	R@1	R@20	mAP	R@1	R@20	mAP
HA-CNN [3]	48.0	85.4	47.6	90.6	98.3	75.3	80.7	94.3	64.4	64.7	87.1	37.2
GMA only w/ $\mathbf{M}_q$	63.4	87.6	63.5	93.8	98.8	81.2	83.9	95.3	69.0	68.7	88.7	40.6
GMA only w/ $\mathbf{M}_g$	65.4	86.2	57.3	94.2	98.4	79.1	83.6	94.4	65.7	66.3	86.4	37.5
GMA-Full	**69.8**	**88.8**	**63.5**	**96.5**	**98.9**	**85.2**	**87.1**	**95.8**	**72.2**	**74.3**	**90.0**	**43.8**

5.2 Experiments on Image Retrieval

5.2.1 Settings

Our proposed SPGMA method is a general online adaptation algorithm which can be readily used for any visual identification tasks (e.g., person re-identification, vehicle re-identification, image retrieval, face recognition, etc.). As we all know, vehicle re-identification shares the same protocol with P-RID as well as face recognition shares the same person object with P-RID. Therefore, we selected a challenging but general image retrieval task as our additional experiment application. The general image retrieval contains various scene images which are taken from different viewpoints. Compared with P-RID, there are more object categories in image retrieval tasks, and the variations within the same category and across different categories are more severe. Therefore, to thoroughly evaluate our proposed SPGMA method, we further conduct extensive experiments on this general image retrieval task.

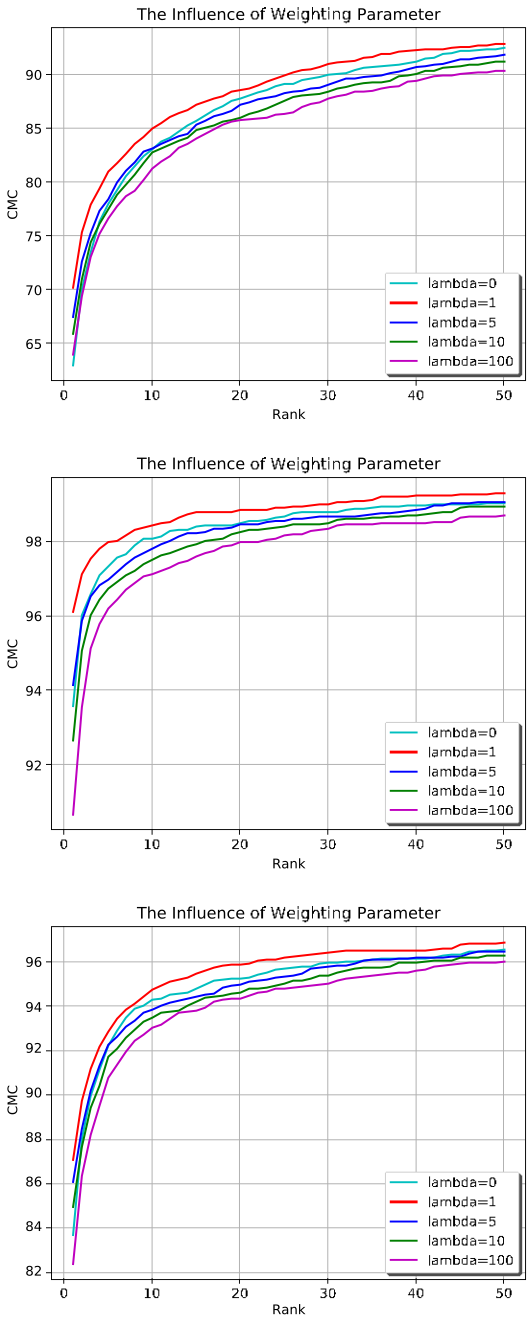


Fig. 6. The influence of on (top) CUHK03, (mid) Market1501 and (bottom) DukeMTMC-reID based on HA-CNN.

Data. We evaluate our proposed SPGMA method on four widely-used image retrieval benchmarks: the original Oxford [61], Paris [62] and their corresponding revisited datasets ROxford and RParis from [63]. The annotation mistakes in the original two datasets are corrected, new query images and new evaluation protocols are added. There are 5063 and 6392 images in the Oxford and Paris datasets which are collected from Flickr associated with Oxford and Paris landmarks respectively. There are 55 queries coming from 11 landmarks for each dataset. As for the revisited ROxford and RPari datasets, another 15 queries from 5 out of the original 11 landmarks are along with the original 55 queries for evaluation.

Evaluation. For all the datasets, the mean average precision (mAP) results over all the query images are reported for evaluation. For ROxford and RPari, two new evaluation difficulties, Medium(M) and Hard(H), are both evaluated.

Baseline. A CNN-based image retrieval model, GeM [64] is utilized as the baseline in our experiments to implement our proposed SPGMA on. Two different CNN backbones, VGG16 [65] and ResNet101 [39], are utilized. Besides, whitening is adopted as post-processing for GeM. Therefore, four different baselines, GeM-VGG16, GeM-VGG16-Whiten, GeM-Res101, and GeM-Res101-Whiten, are examined in our experiments. The pre-trained model from a PyTorch implementation [2] is adopted.

5.2.2 Comparison With State-of-the-Arts

As demonstrated by the comparison results reported in Table 7, our proposed SPGMA can improve the mAP performance of the GeM-VGG16 baseline model from (82.5%, 82.2%, 55.5%, 26.6%, 63.0%, 37.2%) to (84.1%, 83.6%, 56.3%, 27.5%, 64.2%, 37.9%) on (Oxford, Paris, ROxford-M, ROxford-H, RParis-M, RParis-H) respectively. A similar improvement is also observed for the GeM-VGG16-Whiten baseline. As for another more powerful GeM-Res101

2. https://github.com/filipradenovic/cnnimageretrieval-pytorch

TABLE 5
Compared with State-of-the-Art Online P-RID Re-Ranking Methods

Method	CUHK03	Market	Duke
HA-CNN [3]	48.0(47.6)	90.6(75.3)	80.7(64.4)
HA-CNN+RR [11]	54.8(55.7)	91.4(79.0)	82.5(69.9)
HA-CNN+OL [8]	62.3(56.5)	92.7(78.9)	83.7(67.8)
HA-CNN+GMA	**69.8(63.5)**	**96.5(85.2)**	**87.1(72.2)**
HA-CNN+SPGMA	**70.5(64.1)**	**97.1(85.5)**	**87.8(72.7)**
Dense121 [34]	41.0(40.1)	88.2(69.2)	78.6(58.5)
Dense121+RR [11]	48.1(51.5)	90.2(**85.0**)	83.7(**76.9**)
Dense121+OL [8]	53.1(49.3)	90.4(74.0)	80.2(64.1)
Dense121+GMA	**61.6(54.4)**	**95.3**(81.2)	**84.9**(68.0)
Dense121+SPGMA	**62.2(54.9)**	**95.8**(81.9)	**85.6**(68.7)

baseline, our SPGMA method further boosts the mAP performance from (81.0%, 87.7%, 55.5%, 27.5%, 70.0%, 44.7%) to (82.2%, 88.6%, 56.8%, 28.4%, 71.0%, 45.7%) on (Oxford, Paris, ROxford-M, ROxford-H, RParis-M, RParis-H) respectively. Compared with another state-of-the-art online re-ranking method, OL [66], the performance improvement by our SPGMA is much larger since our SPGMA can fully explore all the query and gallery samples and the fused local adaptation metrics are more discriminative as demonstrated by Theorem 4.

5.3 Ablation Study

5.3.1 The Effectiveness Influence of Model Components

The final retrieval performance of Eq. (12) relies on a bi-directional retrieval matching, so the influence of each component is shown in Table 4. As we can see, by only keeping the query-specific metric adaptation M_q or the gallery-centric one M_g, we still can achieve a significant improvement. While performing a full-model bi-directional matching, the performance is further boosted by a large margin which demonstrates the necessity of bi-directional local discriminant enhancement.

5.3.2 The Effectiveness Influence of in Eq. (12)

The weighting parameter in Eq. (12) aims to balance the importance of M_q and M_g. The full CMC curves w.r.t of HA-CNN on CUHK03, Market1501 and DukeMTMC-reID are plotted in Fig. 6 respectively. As demonstrated, setting ¼ 1 gives the best performance since we perform a max-normalization to both M_q and M_g, over-weighting either side is prone to suppress the other side's impact.

TABLE 7
Comparison Results on Oxford, Paris, ROxford and RParis

Method	Ox	Paris	$\mathcal{R}$Ox		$\mathcal{R}$Paris	
			M	H	M	H
MAC [15]	56.4	72.3	37.8	14.6	59.2	35.9
SPoC [47]	68.1	78.2	38.0	11.4	59.8	32.4
CroW [67]	70.8	79.7	41.4	13.9	62.9	36.9
R-MAC [68]	66.9	83.0	42.5	12.0	66.2	40.9
NetVLAD [69]	67.6	74.9	37.1	13.8	59.8	35.0
GeM-VGG16 [64]	82.5	82.2	55.5	26.6	63.0	37.2
GeM-VGG16-Whiten [64]	87.2	87.8	60.5	32.4	69.3	44.3
GeM-Res101 [64]	81.0	87.7	55.5	27.5	70.0	44.7
GeM-Res101-Whiten [64]	88.2	92.5	65.3	40.0	76.6	55.2
OL+VGG16 [66]	83.5	82.9	55.9	26.8	63.5	37.3
OL+VGG16-Whiten [66]	88.1	87.9	60.7	32.6	69.7	44.5
OL+Res101 [66]	81.7	87.6	56.1	27.8	70.3	44.9
OL+Res101-Whiten [66]	89.3	92.6	65.7	40.4	76.9	55.4
SPGMA+VGG16	**84.1**	**83.6**	**56.3**	**27.5**	**64.2**	**37.9**
SPGMA+VGG16-Whiten	**89.6**	**89.1**	**61.2**	**33.8**	**70.6**	**45.3**
SPGMA+Res101	**82.2**	**88.6**	**56.8**	**28.4**	**71.0**	**45.7**
SPGMA+Res101-Whiten	**90.3**	**93.3**	**66.8**	**41.2**	**77.5**	**56.2**

The mAP results are reported.

5.3.3 The Effectiveness Comparison Against Online Re-Ranking Methods

Two state-of-the-art online P-RID re-ranking methods, OL [8] and RR [11], are compared with our GMA and SPGMA methods. All these methods can be readily utilized at the online testing stage for further performance improvement. The comparison results in Table 5 show that the query-specific method OL [8] works better on improving Rank@1 performance but has little improvement on mAP due to the lack of gallery-specific local discriminant enhancement. In contrast, since RR [11] considers the k-reciprocal nearest neighbors of both query and gallery data, it achieves a large improvement on mAP but with limited improvement on Rank@1 owing to the lack of instance-specific local adaptation. Our methods outperform the other two approaches significantly at both Rank@1 and mAP due to the full utilization of both the group-level visual similarity sharing information and instance-specific local discriminant enhancement.

5.3.4 The Computation Cost Comparison Against Online Re-Ranking Methods

To thoroughly evaluate the performance of online re-ranking methods, besides the effectiveness comparison, the computation cost is another important factor. Therefore, we have accordingly compared the online inference cost of our proposed methods with the other state-of-the-art re-ranking

TABLE 6
Cross-Dataset Validation Results with Our Model on Market1501 and DukeMTMC-reID

Method	Market1501 → DukeMTMC					DukeMTMC → Market1501				
	R@1	R@5	R@10	R@20	mAP	R@1	R@5	R@10	R@20	mAP
MLFN [40]	45.8	63.9	71.6	78.1	20.3	30.4	47.5	53.9	59.5	17.1
MLFN+GMA	**67.6**	**78.8**	**83.0**	**86.6**	**32.7**	**43.7**	**57.0**	**62.6**	**68.2**	**24.7**
MLFN+SPGMA	**68.2**	**79.3**	**83.5**	**87.0**	**33.3**	**44.1**	**57.4**	**63.0**	**68.7**	**25.1**
DenseNet121 [34]	41.0	56.6	62.8	68.5	23.2	55.0	71.3	78.5	84.3	25.3
DenseNet121+GMA	**53.1**	**67.1**	**72.1**	**75.7**	**32.7**	**76.9**	**85.6**	**89.1**	**91.9**	**40.4**
DenseNet121+SPGMA	**53.6**	**67.5**	**72.5**	**76.1**	**33.2**	**77.4**	**86.1**	**89.6**	**92.4**	**40.8**
HA-CNN [3]	43.3	59.7	66.7	74.6	18.9	24.0	39.0	45.1	51.6	13.5
HA-CNN+GMA	**61.6**	**73.6**	**78.7**	**82.9**	**25.8**	**37.6**	**51.9**	**56.8**	**62.8**	**20.5**
HA-CNN+SPGMA	**62.2**	**74.0**	**79.3**	**83.5**	**26.4**	**38.0**	**52.5**	**57.3**	**63.4**	**20.9**

Market1501 → DukeMTMC mean using the model trained on Market1501 to evaluate DukeMTMC-reID.

methods. Based on a 1024-dimension CNN feature extractor, the overall online inference time for RR [11] and OL [8], GMA and SPGMA on Market1501 dataset is 78.5s, 242.5s, 107.7s, and 96.8s respectively. Compared with RR [11], our inference time is comparable to it (our running time is a little slower than RR [11] but still acceptable for online evaluation, our running time includes the SSSet mining time and self-paced learning time), while our improvement on Rank@1 and mAP is much more stable and significant than RR [11] as shown in Table 5. It is worth noting that RR [11] just learns a single global matching metric for online re-ranking, while our proposed GMA/SPGMA will learn multiple group-metrics instead for instance-level online adaptation.

Compared with another state-of-the-art instance-level online adaptation method OL [8], our proposed methods could obtain both better re-ranking performance and faster online inference speed. If n query samples are given, OL [8] has to learn n separate local metrics for all the samples so the learning complexity is $O(dn^2)$. For our method, we propose to learn only one metric for one SSSet, instead of for one sample. So the number of mined SSSets is much smaller than n which makes our overall inference time is much shorter than OL [8]. The conclusion is also verified by the online efficiency comparison experiments in Fig. 7. As we can see, the total number of learned online adaptation metrics of OL [8] and our SPGMA method on the Market1501, DukeMTMC, and MSMT17 datasets based on HA-CNN, MLFN and DenseNet121 are demonstrated. For all the feature extractors and benchmarks, the number of learned metrics of our proposed SPGMA is the only 40% of OL's so that our online computation cost (96.8s) if largely reduced compared with OL [8] (242.5s).

Finally, we compare the online learning time of our proposed GMA and the extended version of SPGMA. Although there is an extra iterative self-paced SSSet selection component in SPGMA, the overall online adaptation time is indeed shorter than GMA. The reasons are three-fold: (1) The extra time cost of SPGMA comes from the computation of the hinge loss of the obtained SSSets. Usually, the mined SSSets only contain several highly similar samples (less than 10) thus the hinge loss computation is pretty efficient. (2) The closed-form solution to Eq. (9) can be easily obtained via Eq. (10). Therefore, the overall time cost of the self-paced learning strategy is pretty low which could be ignored during our group-metric adaptation algorithm. (3) SPGMA proposes to gradually involve the obtained SSSets into adaptation learning. Thus, for the learning of one group-metric, the involved testing samples of SPGMA are much fewer than GMA which results in a much faster optimization processing. Although the number of adapted group-metrics is the same for both GMA and SPGMA methods, the overall online adaptation time of SPGMA is shorter than GMA.

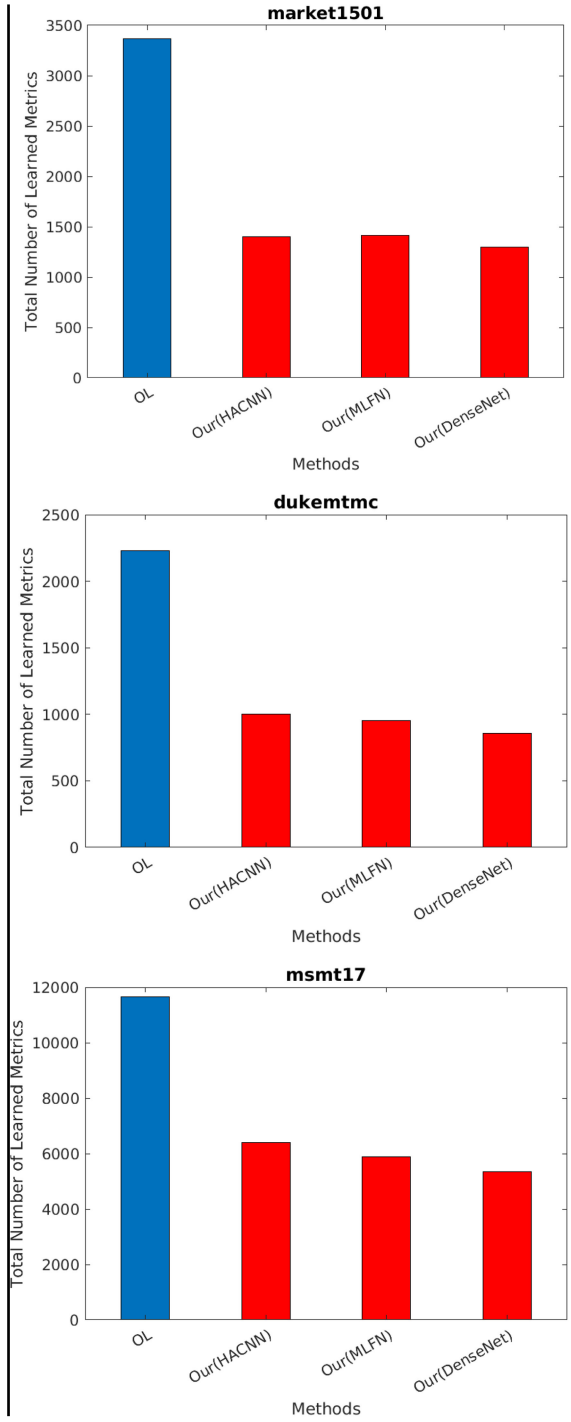


Fig. 7. The comparison of computational cost of OL and our SPGMA method on the Market1501, DukeMTMC-reID, and MSMT17 datasets based on the HA-CNN, MLFN and DenseNet121 baselines. The total number of learned online adaptation metrics are demonstrated.

5.3.5 Cross-Set Generalization Ability Validation

For instance-level identification tasks (e.g., person re-identification), data from non-overlapping identities are provided for training and testing. However, due to large visual appearance variations among training and testing data, there is always a significant performance drop of existing learning-based methods: although identification models are already well-trained on the training dataset, its factual performance on unseen testing data is limited. Such a phenomenon can be observed and verified through our conducted cross-dataset validation experiments in Table 6. Even the state-of-the-art P-RID networks have already been well-trained on the source training dataset (appealing performance can be obtained on the source testing dataset as reported in Table 2), when an unseen target testing dataset from another benchmark is given, their factual identification performance degrades badly. This

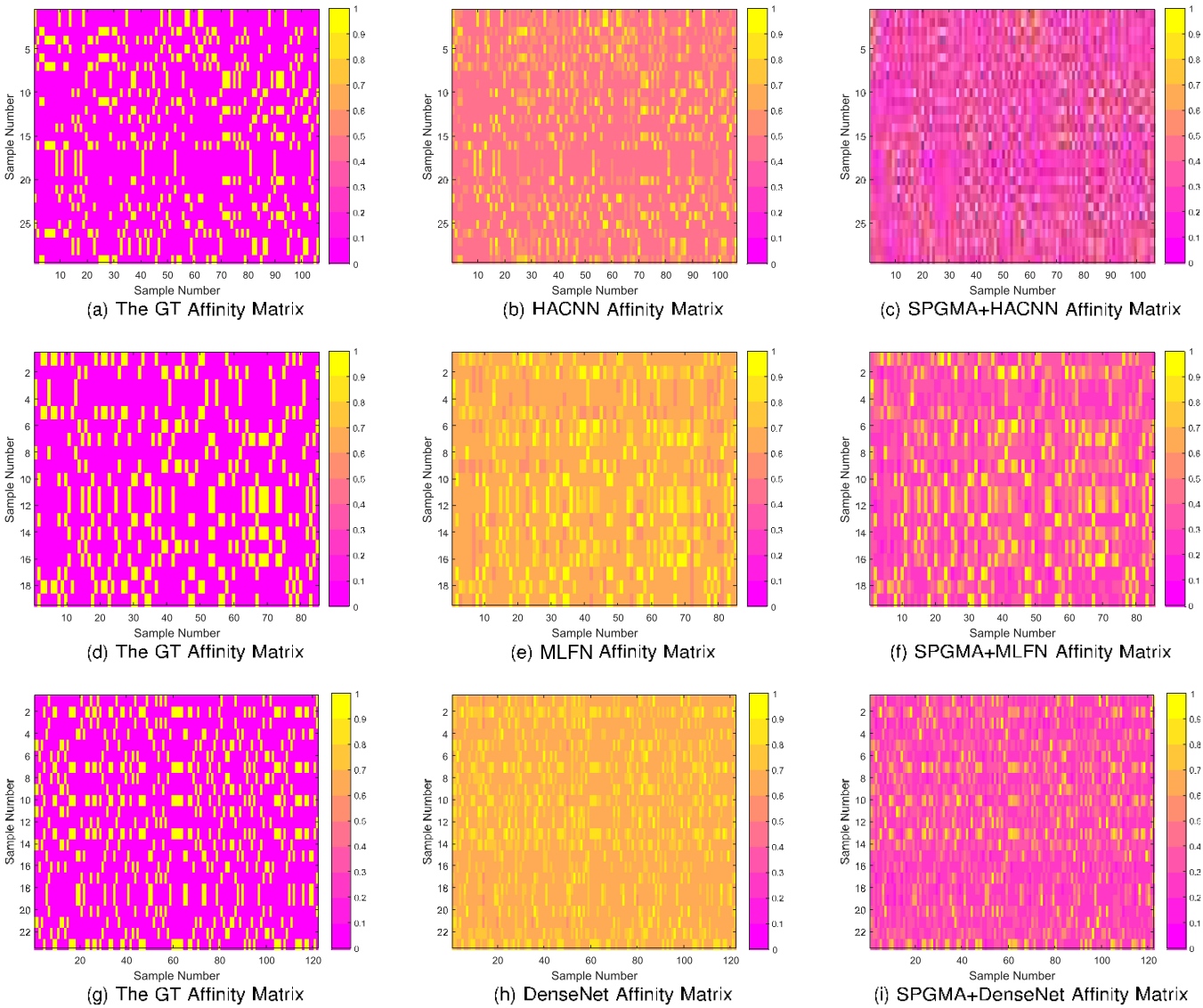


Fig. 8. The affinity matrix refinement by our SPGMA method on the Market1501 dataset based on the HA-CNN, MLFN and DenseNet121 baselines.

result demonstrates these networks are highly over-fitted to the used source training data and their generalization ability to unseen target testing data is pretty poor.

Therefore, we explore the generalization ability of our proposed GMA and SPGMA methods. We claim our improvement is achieved from the testing sample itself which is independent of how the baseline models are trained. Thus we conduct a cross-set generalization ability validation experiment as shown in Table 6. Following the setting in [70], the baseline model trained on Market1501 with our method is evaluated on DukeMTMC-reID and vice versa. The results show our models can consistently and significantly improve the baseline performance regardless of whether the baseline is trained by the same-source data or not.

5.3.6 The Visualization of Affinity Matrix Refinement

The core idea of our proposed SPGMA method is to adapt the local similarity of each online testing sample to better fit its inherent affinity relationships. Therefore, to further verify that our method is able to refine the local similarity of samples and largely alleviate the data shifting problem, we visualize the affinity matrix of testing samples with/without our SPGMA and compare them with the ground-truth results. Extensive experimental results on Market1501 (Fig. 8) and on DukeMTMC-reID (Fig. 9) demonstrate the effectiveness of our method to refine the affinity matrix of samples. As shown in Figs. 8 and 9, without our proposed SPGMA model, the offline learned SOTA baselines can not obtain the correct affinity matrix of the testing samples (the middle column) compared with the ground-truth results (the left column), their affinity matrix is indistinguishable due to the severe data shifting variations. Our SPGMA model can successfully address the data shift problem by adjusting the original affinity matrix to be more coherent with the ground-truth. This is the main reason why our proposed method can significantly improve the identification performance.

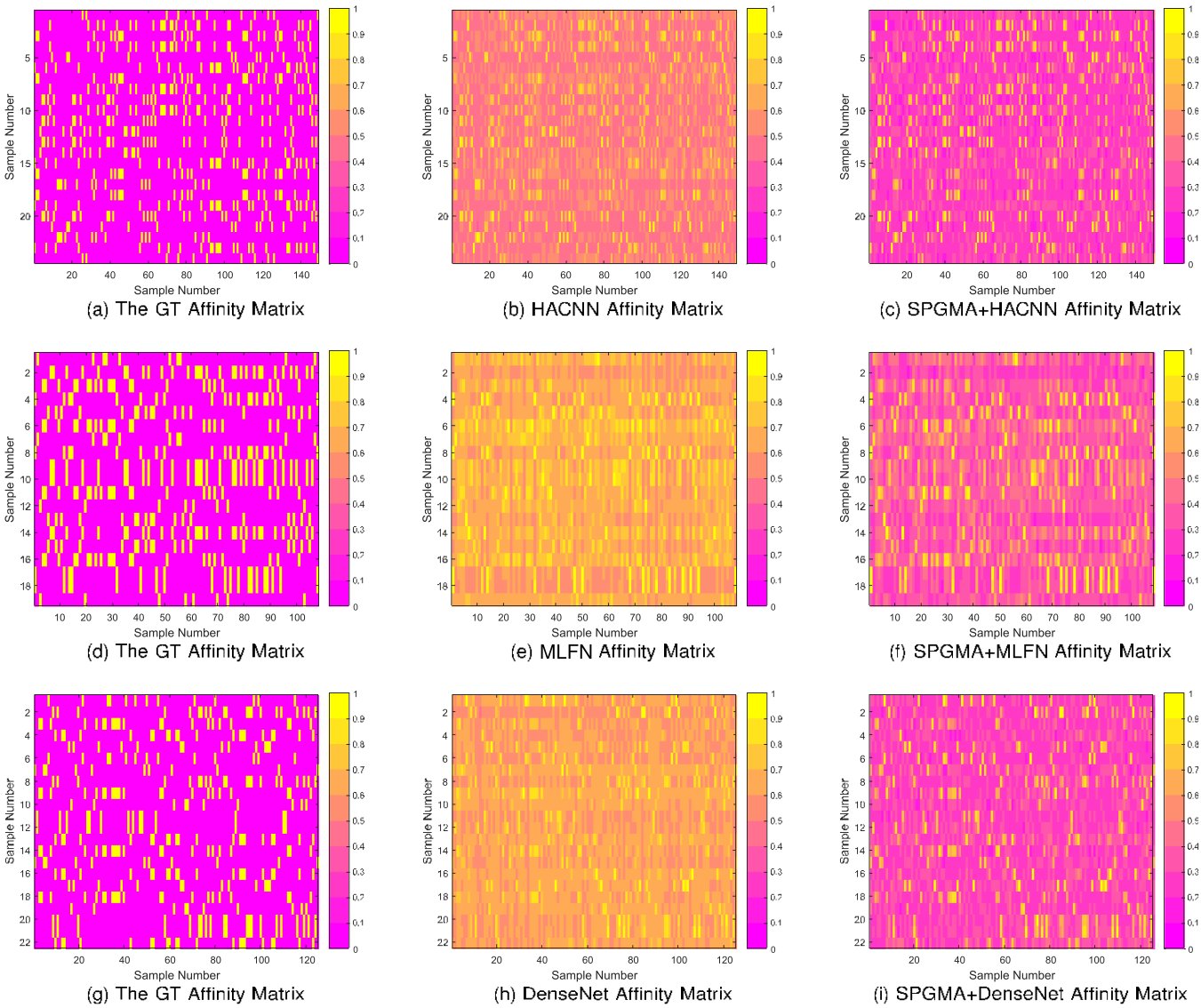


Fig. 9. The affinity matrix refinement by our SPGMA method on the DukeMTMC-reID datasets based on the HA-CNN, MLFN and DenseNet121 baselines.

6 Conclusion

Unlike previous online re-ranking works for visual identification, in this article, we propose a novel online self-paced group-metric adaptation algorithm which not only takes individual characteristics of testing samples into consideration but also fully utilizes the visual similarity relationships among both query and gallery samples. To handle a large number of testing samples, we introduce self-paced learning to gradually include samples into adaptation from easy to difficult which elaborately simulates the learning principle of humans. Our proposed SPGMA method can be readily applied to any existing visual identification baselines with the guarantee of performance improvement, and a theoretically sound optimization solution to SPGMA keeps a low online computational burden. Compared with the other state-of-the-art online rank refinement approaches, the proposed SPGMA model achieves a significant improvement on Rank@1 (mAP) performance. Moreover, by implementing our SPGMA method to the state-of-the-art baselines, their performance is further boosted by a large margin on both the person re-identification and image retrieval tasks.

Acknowledgments

The authors would like to thank the associate editor and anonymous reviewers for their valuable comments.

References

[1] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2990–2999.

[2] T. Ali and S. Chaudhuri, "Maximum margin metric learning over discriminative nullspace for person re-identification," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 122–138.

[3] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2285–2294.

[4] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 365–381.

[5] T. Chen et al., "ABD-Net: Attentive but diverse person re-identification," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 8351–8361.

[6] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2138–2147.

[7] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9317–9326.

[8] J. Zhou, P. Yu, W. Tang, and Y. Wu, "Efficient online local metric adaptation via negative samples for person re-identification," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2420–2428.

[9] M. Ye, J. Chen, Q. Leng, C. Liang, Z. Wang, and K. Sun, "Coupled-view based ranking optimization for person re-identification," in Proc. Int. Conf. Multimedia Model., 2015, pp. 105–117.

[10] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel, "Person re-identification ranking optimisation by discriminant context information analysis," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1305–1313.

[11] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1318–1327.

[12] A. Barman and S. K. Shah, "Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1115–1124.

[13] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 740–749.

[14] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in Proc. Asian Conf. Comput. Vis., 2012, pp. 31–44.

[15] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific SVM learning for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1278–1287.

[16] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in Proc. Int. Conf. Neural Inf. Process. Syst., 2010, Art. no. 2.

[17] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2078–2086.

[18] K. Ghasedi, X. Wang, C. Deng, and H. Huang, "Balanced self-paced learning for generative adversarial clustering network," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4391–4400.

[19] R. Caruana, "Multitask learning," Mach. Learn., vol. 28, no. 1, pp. 41–75, 1997.

[20] J. Zhou, B. Su, and Y. Wu, "Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2909–2918.

[21] G. Zhang, Y. Wang, J. Kato, T. Marutani, and K. Mase, "Local distance comparison for multiple-shot people re-identification," in Proc. Asian Conf. Comput. Vis., 2012, pp. 677–690.

[22] V. E. Liong, J. Lu, and Y. Ge, "Regularized local metric learning for person re-identification," Pattern Recognit. Lett., vol. 68, pp. 288–296, 2015.

[23] Y. Li, H. Yao, L. Duan, H. Yao, and C. Xu, "Adaptive feature fusion via graph neural network for person re-identification," in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 2115–2123.

[24] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 9637–9646.

[25] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3186–3195.

[26] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," ACM Trans. Multimedia Comput., Commun., Appl., vol. 14, no. 4, pp. 1–18, 2018.

[27] H. Li, M. Gong, D. Meng, and Q. Miao, "Multi-objective self-paced learning," in Proc. AAAI Conf. Artif. Intell., 2016, pp. 1802–1808.

[28] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 594–602.

[29] X. Xin, X. Wu, Y. Wang, and J. Wang, "Deep self-paced learning for semi-supervised person re-identification using multi-view self-paced clustering," in Proc. IEEE Int. Conf. Image Process., 2019, pp. 2631–2635.

[30] S. Zhou et al., "Deep self-paced learning for person re-identification," Pattern Recognit., vol. 76, pp. 739–751, 2018.

[31] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," 2020, arXiv:2006.02713.

[32] B. Fernando, E. Fromont, and T. Tuytelaars, "Effective use of frequent itemset mining for image classification," in Proc. Eur. Conf. Comput. Vis., 2012, pp. 214–227.

[33] G. Grahne and J. Zhu, "Efficiently using prefix-trees in mining frequent itemsets," in Proc. Workship Frequent Itemset Mining Implementations, 2003, Art. no. 65.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4700–4708.

[35] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," J. Mach. Learn. Res., vol. 7, pp. 1531–1565, 2006.

[36] S. Sonnenburg, G. Rätsch, and C. Schäfer, "A general and efficient multiple kernel learning algorithm," in Proc. Adv. Neural Inf. Process. Syst., 2005, pp. 1273–1280.

[37] N. Srebro and S. Ben-David, "Learning bounds for support vector machines with learned kernels," in Proc. Int. Conf. Comput. Learn. Theory, 2006, pp. 169–183.

[38] Z. Hussain and J. Shawe-Taylor, "Improved loss bounds for multiple kernel learning," in Proc. 14th Int. Conf. Artif. Intell. Statist., 2011, pp. 370–377.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[40] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2109–2118.

[41] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3702–3712.

[42] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 480–496.

[43] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 3800–3808.

[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

[45] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6848–6856.

[46] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in Proc. IEEE Int. Conf. Comput. Vis. Workshops, 2017, pp. 2590–2600.

[47] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1239–1248.

[48] J. Si et al., "Dual attention matching network for context-aware feature sequence based person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5363–5372.

[49] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[50] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 402–419.

[51] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1492–1500.

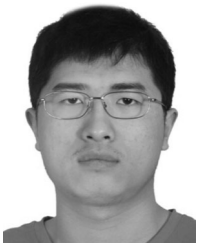
[52] X. Qian et al., "Pose-normalized image generation for person re-identification," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 650–667.

[53] Y. Wang et al., "Resource aware person re-identification across multiple resolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8042–8051.

[54] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6036–6046.

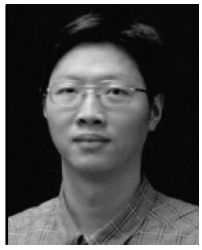
[55] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 3754–3762.

[56] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in Proc. 26th ACM Int. Conf. Multimedia, 2018, pp. 274–282.
[57] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1062–1071.
[58] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1116–1124.
[59] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 79–88.
[60] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size," 2016, arXiv:1602.07360.
[61] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.
[62] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2008, pp. 1–8.
[63] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5706–5715.
[64] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1655–1668, Jul. 2018.
[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv:1409.1556.
[66] J. Zhou and Y. Wu, "Learning visual instance retrieval from failure: Efficient online local metric adaptation from negative samples," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2858–2873, Nov. 2020.
[67] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 3685–3693.
[68] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 2197–2206.
[69] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1565–1573.
[70] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 172–188.



Jiahuan Zhou received the BE degree from Tsinghua University, in 2013, and the PhD degree from the Department of Electrical Engineering & Computer Science, Northwestern University, in 2018. Currently, he is an Assistant Professor with the Wangxuan Institute of Computer Technology, Peking University. Before joining Peking University, he was a postdoctoral fellow and research assistant professor in Northwestern University. His research interests include computer vision, pattern recognition, visual identification and machine learning. He served as an area chairs for CVPR, ICME and ICPR.

Bing Su received the BS degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2010, and the PhD degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. From 2016 to 2020, he worked with the Institute of Software, Chinese Academy of Sciences, Beijing. Currently, he is an Associate Professor with the Gaoling School of Artificial Intelligence, Renmin University of China. His research interests include pattern recognition, computer vision, and machine learning.



Ying Wu (Fellow, IEEE) received the BS degree from the Huazhong University of Science and Technology, Wuhan, China, in 1994, the MS degree from Tsinghua University, Beijing, China, in 1997, and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 2001. From 1997 to 2001, he was a Research Assistant with the Beckman Institute for Advanced Science and Technology, UIUC. During summer 1999 and 2000, he was a research intern with Microsoft Research, Redmond, Washington. In 2001, he joined the Department of Electrical and Computer Engineering, Northwestern University, Evanston, Illinois, as an assistant professor. He was promoted to associate professor in 2007 and full professor in 2012. He is currently full professor of electrical engineering and computer science with Northwestern University. His current research interests include computer vision, robotics, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human-computer interaction. He serves as the associate editor-in-chief for APR Journal of Machine Vision and Applications, and associate editors for IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, SPIE Journal of Electronic Imaging. He served as program chair and area chairs for CVPR, ICCV and ECCV. He received the Robert T. Chien Award at UIUC in 2001, and the NSF CAREER award in 2003. He is a Fellow of the IAPR.