Learning Visual Instance Retrieval from Failure: Efficient Online Local Metric Adaptation from Negative Samples

Jiahuan Zhou, Ying Wu, *Fellow, IEEE,*

Abstract—Existing visual instance retrieval (VIR) approaches attempt to learn a faithful global matching metric or discriminative feature embedding offline to cover enormous visual appearance variations, so as to directly use it online on various unseen probes for retrieval. However, their requirement for a huge set of positive training pairs is very demanding in practice and the performance is largely constrained for the unseen testing samples due to the severe data shifting issue. In contrast, this paper advocates a different paradigm: part of the learning can be performed online but with nominal costs, so as to achieve online metric adaptation for different query probes. By exploiting easily-available negative samples, we propose a novel solution to achieve the optimal local metric adaptation effectively and efficiently. The insight of our method is the local hard negative samples can actually provide tight constraints to fine tune the metric locally. Our local metric adaptation method is generally applicable to be used on top of any offline-learned baselines. In addition, this paper gives in-depth theoretical analyses of the proposed method to guarantee the reduction of the classification error both asymptotically and practically. Extensive experiments on various VIR tasks have confirmed our effectiveness and superiority.

Index Terms—Visual Instance Retrieval, Online Metric Adaptation, Hard Negative Samples.

1 Introduction

Visual Instance Retrieval (VIR) generally refers to retrieving the same-instance images for the query instance image from a large, unordered image collection, gallery set, based on the visual similarities between the query probe and the gallery images. The gallery images may be obtained from different cameras at a different time against the query probe so that the difficulties of VIR are mainly caused by the large and complex visual appearance variations under various views, poses, illumination and occlusion conditions. Owing to these challenges, VIR remains a critical yet very challenging task in computer vision community which plays an important role in various research topics, e.g., image retrieval (Img-R) [1], [2], [3], [4], person re-identification (P-RID) [5], [6], [7], and vehicle re-identification (V-RID) [8], [9] etc.

Most attempts to VIR focus on facilitating the retrieval by learning a discriminative matching metric [5], [6], [10], [11], [12] or feature embedding [3], [4], [8], [9], [13], [14], [15], [16], [17], [18] to better capture the visual similarities. In this paper, we use the same term metric to represent both the matching metric and feature embedding for convenience since they are indeed interchangeable. These offline metric learning methods typically attempt to train a faithful global metric offline, hoping to cover the enormous visual appearance variations so as to directly use it online for all testing probes. The training data for such offline learning are generally sample pairs: a *positive* pair refers to two images of the same identity, and a *negative* pair otherwise. These methods usually demand a huge set of positive/negative training pairs to facilitate learning. In practice, although it is relatively easy to collect negative pairs, it is in general difficult to obtain many positive pairs for a specific instance. Therefore, the metrics learned from insufficient positive training data are likely to be biased. In addition, most methods aim to learn a positive semi-definite (PSD) Mahalanobis metric, but it is computationally intensive to learn such a strictly PSD metric, while ignoring the PSD constraint leads to unstable and noisy metrics [5].

In contrast to the aforementioned methods, this paper advocates a different paradigm: **shifting part of the learning to the online local metric adaptation**. Specifically, for each online probe at the testing time, our new approach learns a dedicated local metric with a nominal computational cost. Combining a global baseline with local metric adaptation achieves an adaptive nonlinear metric. In our approach, its online learning is special, because there are no positive training pairs available at all for the testing probe, as its identity is unknown.

An attractive property of our proposed method is that it only uses negative data from a negative sample database (NDB) for adaptation learning. We call it **OLMANS** for short of Online Local Metric Adaptation from Negative Samples. For a given testing probe, a specific subset of samples from NDB are selected to form informative negative pairs with this testing probe. These utilized samples from NDB are visually similar to the probe, but are guaranteed to have different identities from the probe (at least with a very large probability). These negative samples provide effective local discrimination for further constraining the local metric tuning, by pushing away local false positives (shown in Fig. 1). For each testing probe, our method learns a strictly PSD local metric via solving a max-min optimization problem efficiently. Comparing to offline learning schemes, the computational cost of the proposed online adaptation is negligible. Moreover, our method is generally applicable to be used on top of any offline learned baselines without any modification to them.

Another significant property of our proposed OLMANS is that it is justified and backed up with a theoretical guarantee to improve

- *The authors are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, 60208. E-mail: {jzt011, yingwu}@eecs.northwestern.edu*

Manuscript received xxxx.

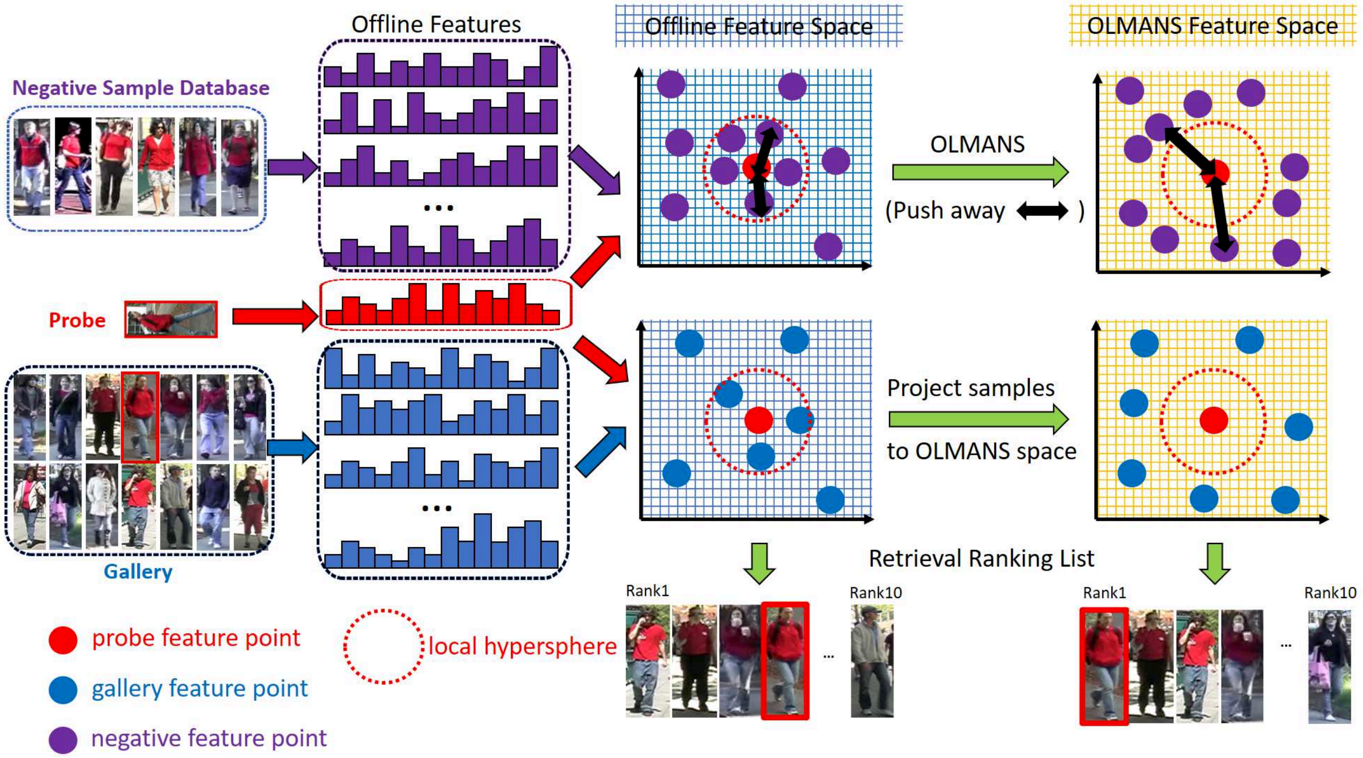


Fig. 1. The overall idea of our proposed online local metric adaptation algorithm illustrated in the context of P-RID. Unlike existing offline learning-based methods that learn a single global metric or feature embedding for all probe and gallery samples, we exploit negative samples to learn a dedicated local metric for each online probe to adapt the offline learned global feature space to an instance-specific discriminative local feature space (called OLMANS feature space). The hard negatives in NDB around the local hypersphere of the query probe are pushed far away so the final retrieval result in OLMANS feature space is improved.

the performance of the underlying VIR baseline. This paper gives in-depth theoretical analyses to well justify our proposed method. We first prove that the novel OLMANS guarantees the reduction of classification error asymptotically when there are an infinite number of learning data. Then we pursue the best approximation of the asymptotic case by using a finite number of learning data, since we can prove that the learning objective of the proposed local metric adaptation is equivalent to the optimal approximation of the asymptotic case. In addition, we also provide consistency and sample complexity analysis to guarantee the generalization ability of our proposed OLMANS. These theoretical analyses indicate that the learned local metric is bound to improve the VIR performance. These properties have been confirmed to be significantly effective and practical by our extensive experiments and comparative studies on different VIR benchmarks: P-RID (VIPeR, GRID, CUHK03, Market1501, DukeMTMC-reID and MSMT17) and Img-R (Oxford, Paris, $\mathcal{R}$Oxford and $\mathcal{R}$Paris).

This paper is an extension of our previous conference paper [19], while we have made a lot of extensions including: 1) We extend our proposed OLMANS model to a more general form to better fit the set-query scenario. The semantic and visual similarity relationships of the given set-based queries from the same instance are fully explored for a robust and discriminative metric adaptation. 2) The theoretical analyses with a thorough proof of our OLMANS are completely presented in Sec. 4, which theoretically guarantee the correctness of our proposed method. 3) We compare our method with the widely-used online re-ranking technique since both our OLMANS and re-ranking methods are applied to the offline learned VIR baselines on online stage for further performance boosting, while our OLMANS outperforms re-ranking in both the performance and efficiency. 4) We evaluate our OLMANS on two generic VIR tasks: person re-identification (P-RID) and image retrieval (Img-R). Compared with [19] which only focuses on the specific P-RID problem, the evaluation on a general image retrieval task verifies the generalization ability and effectiveness of our method. 5) For the P-RID evaluation, more ablation experiments are conducted in Sec. 5 to further investigate our proposed method. In addition, unlike [19] that only uses the handcrafted feature and small-scale P-RID datasets, we explore more state-of-the-art deep learning-based models as our baselines and evaluate three more challenging large-scale P-RID benchmark datasets (CUHK03 [20] with new protocol, DukeMTMC-reID [21] and MSMT17 [22]) to challenge various data conditions.

The rest of our paper is organized as follows: Section. 2 summarizes the previous works on VIR. We describe our proposed OLMANS algorithm in Section. 3, and illustrate its performance on many benchmark datasets in Section. 5. In Section. 4, we theoretically analyze some important properties of our proposed algorithm.

2 Related Work

2.1 Person Re-identification

In this work, we focus more on the local metric learning-based P-RID approaches and convolutional neural network (CNN)-based deep feature embedding P-RID models.

Local Metric Learning: [23] formulated the P-RID problem as a local distance comparison problem to handle the multi-modal distributions of the visual appearances. [24] proposed the Locally-Adaptive Decision Functions (LADF) which integrates a traditional distance metric with a local decision rule. [25] employed the Local Fisher Discriminant Analysis (LFDA) which combines the Fisher Discriminant Analysis (FDA) and Local Preserving Projections (LPP) to exploit the local geometrical information of samples. [26] developed a regularized local metric learning (RLML) method to combine global and local metrics, so as to utilize the local data distribution to alleviate over-fitting. [27] proposed LSSCDL to learn a specific SVM classifier for each training sample, then the weight parameters of a new sample can be inferred. A novel multi-task maximally collapsing metric learning (MtMCML) model was proposed by [28]. In order to relax the large-number labeled image pair requirement in P-RID, a novel one-shot learning approach is proposed by [10] which only requires a single image from each camera for training, thus the learning result is specific to the only sample. In contrast to the local metric learning methods, our proposed approach is mainly focused on learning local metrics specifically adaptive to individual testing probes. Different from RLML that requires clustering in advance to obtain the local data distributions, our new approach does not need clustering but is rather instance-based learning, and thus avoiding the risk of inaccurate clustering results. Also note that MtMCML learning still follows the global manner although it learns different metrics for different cameras. In contrast to LADF that needs a large number of positive sample pairs to drive the local decision function learning, our new approach only uses negative sample pairs which are much easier to obtain. LSSCDL also requires a lot of positive training pairs for offline learning, but ours performs online learning per probe without the requirement of positive pairs. Although [10] performs one-shot learning to each sample, but it needs extra camera network information for one-shot learning.

Deep Feature Embedding: The convolutional neural network (CNN)-based P-RID approaches aim to integrate the feature extraction and metric learning into one end-to-end framework, in which a neural network is built to extract from each pedestrian image a feature that satisfies a certain ranking criterion. [20] firstly utilized deep learning method to extract more effective and discriminative features to facilitate P-RID. [29] proposed a scalable deep feature learning model for P-RID via relative distance comparison based on triplet loss. [30] proposed a novel moderate positive mining method to embed a robust deep metric for P-RID. [31] suggested a new loss for learning deep embeddings and demonstrate competitive results of the new loss on a number of P-RID datasets. CNN-based feature extraction has achieved the state-of-the-art performance in P-RID owning to a better spatial alignment of local image parts. A novel Harmonious Attention CNN (HA-CNN) proposed by [13] 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). [32] proposed a network called CAN which combines attention methods with LSTM to obtain discriminative attention feature of the whole image. [33] proposed a novel deeply supervised fully attentional block that can be plugged into any CNNs to solve P-RID problem, and a novel deep network called Mancs is designed to learn stable features for P-RID. Besides the aforementioned methods, the utilization of hard negatives attracts more and more attention in deep metric learning area. [34] proposed a framework of deep adversarial metric learning (DAML) which can be generally applicable to various supervised metric learning approaches. DAML aims to generate synthetic hard negatives from the observed negative samples by exploiting what to generate potential hard negatives adversarial to the learned metric as complements. [35] proposed a novel applicable framework named deep variational metric learning (DVML) to disentangle intra-class variance via variational inference and leverages the distribution to generate discriminative samples to improve robustness. The generated negative samples could be utilized to facilitate the learning and enhance the generalization ability of the learned model. However, these well-trained networks are directly applied to the testing data for deep feature extraction, no local adaptation is in the loop. The data shifting between training and testing samples definitely limits the performance of the learned models. Therefore, our proposed OLMANS is suitable for any CNNs for instance-specific local adaptation in the inference stage, which can address the data shifting issue well and gain further performance improvement.

2.2 Image Retrieval

A thorough survey of image retrieval researches is introduced in [37]. In this work, we mainly focus on two main branches of image retrieval, multiple local feature aggregation-based approaches and deep learning-based models.

Local Descriptor Aggregation: Previous image retrieval methods aim to aggregate a set of local feature descriptors into a global one for robust retrieval. [38] designed a graph-based ranking model to aggregate the retrieval results from multiple features into one, then the retrieval scores are weighted to determine the final retrieval matching. [39] proposed a novel coupled MultiIndex(c-MI) framework to fuse both color feature and SIFT feature in a product manner at indexing level. [40] proposed a semantic-aware co-indexing scheme to fuse the SIFT feature and semantic attributes for image retrieval. In [41], multiple visual features are fused in the similarity score level based on the shapes of ranking scores. By considering these local descriptor aggregation methods as offline baselines, our proposed OLMANS can be readily implemented on the top of the fused feature for further local similarity adaptation.

CNN Fine-tuning: [42] demonstrated that the pre-trained models from ImageNet for object classification is suitable for image retrieval by fine-tuning them on an external set of Landmarks images. [43] also confirmed the importance of fine-tuning the pre-trained models to improve image retrieval, but argued that a good image representation and a ranking loss should be used in learning, instead of the classification loss. [4] addressed the unsupervised fine-tuning of CNNs for image retrieval on a large collection of unordered images in a fully automated manner. By considering the fine-tuned CNN as a global deep feature extractor to the probe and gallery samples, our proposed OLMANS method can be readily applied on top of it to further boost the performance.

2.3 Online Re-Ranking

The online re-ranking technique is widely adopted for further performance improvement in VIR. [44] revised the ranking list by considering the nearest neighbors of both the global and local features. An unsupervised re-ranking model proposed by [45] takes advantage of the content and context information in the ranking list. [46] proposed a k-reciprocal encoding approach for

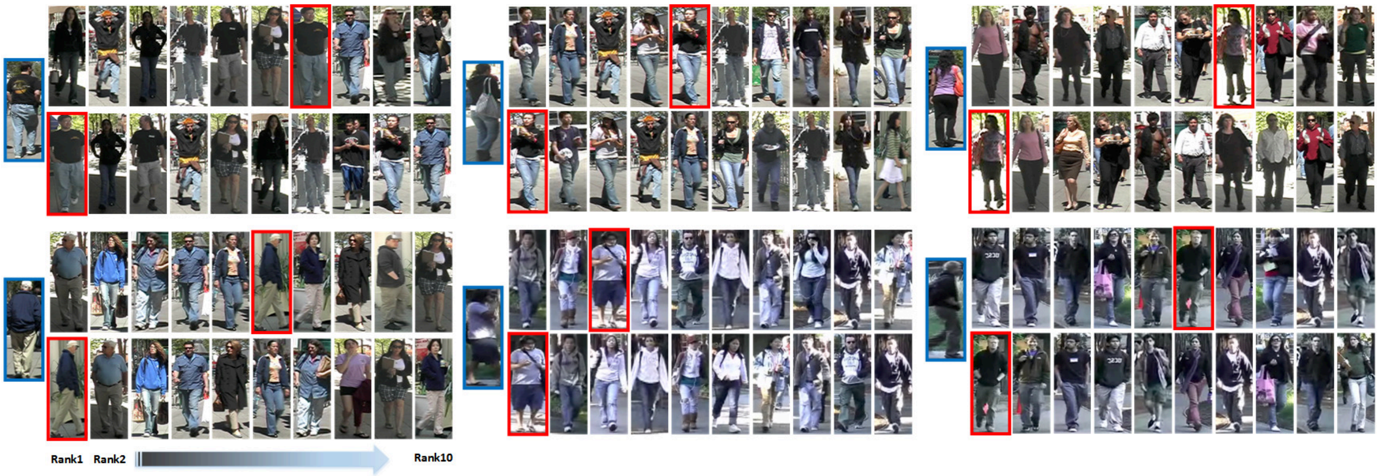


Fig. 2. The improvement of ranking result by our OLMANS on VIPeR [36]. BLUE boxes: input probes, RED: gallery targets. For each case, the top row is the result from the baseline [5], and the bottom row is our result. (Best view in color and enlarged)

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. [47] 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. Both our proposed OLMANS and re-ranking share the same appealing online manner, but our algorithm outperforms re-ranking by several unique merits which will be discussed in Sec. 4.4.

3 Learning from Failure: Online Local Metric Adaptation from Negative Samples

3.1 Problem Settings

On the online testing stage of VIR, two disjoint datasets, a **probe set** $\mathcal{P}$ and a **gallery set** $\mathcal{G}$ are given as:

$$\mathcal{P} = \{(p_i, l_i^p)\}_{i=1}^n \, \mathcal{G} = \{(g_i, l_i^g)\}_{i=1}^m \tag{1}$$

that $p_i, g_i \in \mathbb{R}^d$ are the extracted feature representations from a baseline model, either handcraft features or learned deep features. $l_i^p, l_i^g \in \{1, 2, ..., c\}$ are the labels from c instances which are totally different from the training sample classes. The common-used *closed-set condition* is adopted that both the $\mathcal{P}$ and $\mathcal{G}$ contain samples from all the c instances respectively. VIR aims to rank $\mathcal{G}$ for a query probe p_i based on the pair-wise similarity distance between a gallery image g_j, $D(p_i, g_j) = \|p_i, g_j\|^2$. Our goal is to re-rank $\mathcal{G}$ for p_i by refining $D(p_i, g_j)$ to boost the rank of true-matches for p_i via utilizing an additional negative sample database (NDB), denoted by $\mathcal{Y} = \{y_i\}_{i=1}^k$, the details about $\mathcal{Y}$ will be discussed shortly in Sec. 3.2.

3.2 OLMANS for Single-Instance Query

The performance of VIR depends on the similarity matching between one probe p_i and one gallery image g_j. Different methods adopt different loss functions to learn the feature representations p_i and g_j with the expectation that the similarity structure in the learned feature space should be aligned, so as to pull the samples from the same instance group closer and to make different instances more discriminative. However, the offline learned feature embedding from training samples does not aim to fit the local distributions for all the testing samples specifically, it may lead to large biases and distortions in some places in the feature space. As illustrated in Fig. 1, our proposed approach puts an instance-specific local metric adaptation on top of the global baselines in an online manner.

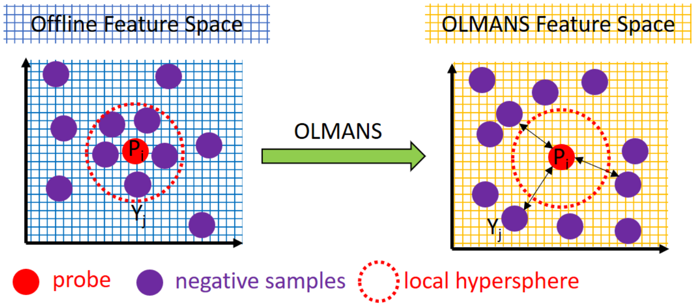


Fig. 3. The local metric $\mathbf{M}_i$ for a single probe p_i can push the closest negative sample y_j of p_i away from its local hypersphere $\Omega(p_i)$

To enhance the local discriminant of query probes, in this paper, we propose OLMANS, an online local metric adaptation algorithm by exploring only negative samples, to adaptively adjust the metric dedicated to a specific query probe with minimum online learning burden. Specifically, for a probe image p_i in the probe set $\mathcal{P}$, we aim to learn a local Mahalanobis distance $\mathbf{M}_i$ only using the samples in a negative sample database $\mathcal{Y}$ as learning data. This negative sample database provides rather faithful negative samples to the probes with a large probability. There are many ways to collect $\mathcal{Y}$, e.g., data from a different benchmark can be used, or false positive matches from images that belong to different instance classes. The insight here is that all such negative samples are "hard negatives" for the probes. In this research, we have investigated how $\mathcal{Y}$ influences the performance in Sec. 5.

We propose to pursue an optimal PSD Mahalanobis metric $\mathbf{M}_i$ for the local adaptation of p_i, by maximizing the distance to the closest (or "hardest" conceptually) negative sample of p_i, as shown in Fig. 3:

$$\mathbf{M}_i = \arg\max_{\mathbf{M}_i \succeq 0} \left(\min_{1 \le j \le k} \left(p_i - y_j\right)^T \mathbf{M}_i \left(p_i - y_j\right) \right) \tag{2}$$

To pursue a stable solution to Eqn. 2, we need to regularize $\mathbf{M}_i$. This can be done via minimizing the norm under a fixed margin constraint, instead of maximizing the margin under a fixed norm constraint [48], so the alternative objective is:

$$\begin{aligned} &\mathbf{M}_i = \arg\min_{\mathbf{M}_i} \frac{1}{2} \|\mathbf{M}_i\|^2 \\ &sub\ \ to: \ \ \mathbf{M}_i \succeq 0 \\ &\left(p_i - y_j\right)^T \mathbf{M}_i \left(p_i - y_j\right) \ge 2, \ \ \forall 1 \le j \le k \end{aligned} \tag{3}$$

where the constant 2 is arbitrary only for manipulation convenience. While this is a convex semi-definite programming problem, it can be very slow for high dimensional data, even for the state-of-the-art PSD solvers.

In the proposed OLMANS approach, we relax the PSD constraint requiring $\mathbf{M}_L^i \succeq 0$, but we prove below that the relaxed objective is equivalent to a kernel SVM problem with a quadratic kernel. And thus the solution is still a PSD metric. In addition, it can be readily solved with off-the-shelf SVM solvers such as LIBSVM [49]. More importantly, we also prove that this learning objective is equivalent to the best approximation to the asymptotic classification error, which is proved to be lower than the global baseline (details see Sec. 4).

Theorem 1. The solution to Eqn. 3 is equivalent to a kernel SVM with $k(x, y) = \langle x, y \rangle^2$ on $\{\tilde{y}_0, \tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_k\}$ where $\tilde{y}_j = p_i - y_j$ (for $j \ge 1$), and $\tilde{y}_0 = p_i - p_i = 0$.

Proof 1. Define auxiliary labels by:

$$\zeta_j = \left\{ \begin{array}{ll} -1, & j = 0 \\ 1, & j \ne 0 \end{array} \right. \tag{4}$$

so the objective Eqn. 3 can be rewritten as:

$$\begin{aligned} &\mathbf{M}_i = \arg\min_{\mathbf{M}_i} \frac{1}{2} \|\mathbf{M}_i\|^2 \\ &sub\ \ to: \ \ \zeta_j \left(\tilde{y}_j^T \mathbf{M}_i \tilde{y}_j - 1 \right) \ge 1, \forall \ \ 0 \le j \le k \end{aligned} \tag{5}$$

Eqn. 5 is exactly an SVM problem with quadratic kernel and with bias fixed to one. Next we prove the solution to objective Eqn. 5 is exactly the same as that to the original objective Eqn. 3. Consider the dual of the SVM, the optimal solution $\mathbf{M}_i$ has the form:

$$\mathbf{M}_i = \sum_{j=0}^{k} \alpha_j \zeta_j \tilde{y}_j \tilde{y}_j^T, \ \ \alpha_j \ge 0 \tag{6}$$

Since $\tilde{y}_j \tilde{y}_j^T$ is PSD for $j \ge 1$ ($\tilde{y}_0 \tilde{y}_0^T = 0$) and $\zeta_j = 1$ for $j \ge 1$, so we have:

$$\mathbf{M}_i = \sum_{j=0}^{k} \alpha_j \zeta_j \tilde{y}_j \tilde{y}_j^T = \sum_{j=1}^{k} \alpha_j \tilde{y}_j \tilde{y}_j^T \succeq 0 \tag{7}$$

It is obvious that the positive semi-definiteness of $\mathbf{M}_i$ is guaranteed even if no PSD constraint is explicitly imposed in our learning objective Eqn. 5.

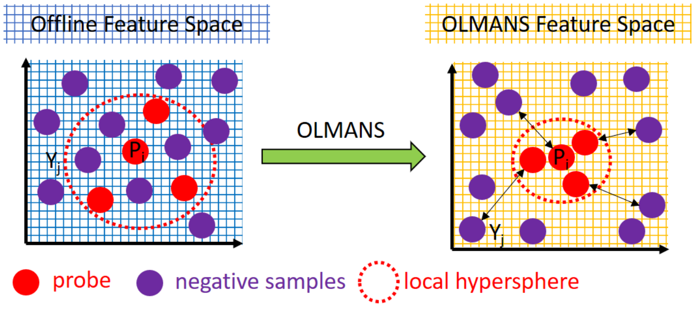


Fig. 4. The local metric $\mathbf{M}_i$ for a set-based probe $\mathcal{P}_i$ can pull the same-instance samples together meanwhile push the closest negative samples y_j away from the local hypersphere $\Omega(\mathcal{P}_i)$

3.3 OLMANS for Instance-Set Query

In Sec. 3.2, we demonstrate our proposed OLMANS algorithm in the context of single-instance query scenario. However, in visual instance retrieval, there will be multiple images of the same instance as the query probe, which is known as the *multi-shot query*. Following our OLMANS algorithm in Sec. 3.2, for each individual image of the same instance, a local metric will be learned which is linear to the query number n. However, such an individual-based learning manner ignores the visual similarity relationships among the given set-based query which is neither effective nor efficient. Therefore, for such an instance-set query, we generalize our OLMANS algorithm to learn a set-specific local Mahalanobis metric in order to collapse the same-instance samples together meanwhile push the negative samples in $\mathcal{Y}$ far away, as shown in Fig. 4. For the i-th instance with query set $\mathcal{P}_i = \{p_r^i\}_{r=1}^{n_i}$, the designed objective for learning its specific Mahalanobis metric $\mathbf{M}_i$ is:

$$\begin{aligned} &\mathbf{M}_i = \arg\min_{\mathbf{M}_i} \frac{1}{2} \|\mathbf{M}_i\|^2 \\ &sub\ \ to: \ \ \mathbf{M}_i \succeq 0 \\ &\left(p_r^i - y_j\right)^T \mathbf{M}_i \left(p_r^i - y_j\right) \ge 2, \ \ \forall 1 \le r \le n_i, \ \ \forall 1 \le j \le k \\ &\left(p_r^i - p_j^i\right)^T \mathbf{M}_i \left(p_r^i - p_j^i\right) = 0, \ \ \forall 1 \le r \le n_i, \ \ \forall 1 \le j \le n_i \end{aligned} \tag{8}$$

Therefore the learned $\mathbf{M}_i$ from Eqn. 8 is shared by all the same-instance samples in $\mathcal{P}_i$. While there are total $O(n^2)$ constraints in Eqn. 8 which is difficult to deal with, so we aim to reduce the constraint size in Eqn. 8 to facilitate optimization.

Theorem 2. Eqn. 8 has an exactly equivalent form by only keeping the constraints related to one anchor sample p^i in the query set $\mathcal{P}_i$, that p^i can be any sample in $\mathcal{P}_i$. Therefore the equivalent form is Eqn. 9:

$$\begin{aligned} &\mathbf{M}_i = \arg\min_{\mathbf{M}_i} \frac{1}{2} \|\mathbf{M}_i\|^2 \\ &sub\ \ to: \ \ \mathbf{M}_i \succeq 0 \\ &\left(p^i - y_j\right)^T \mathbf{M}_i \left(p^i - y_j\right) \ge 2, \ \ \forall 1 \le j \le k \\ &\left(p^i - p_j^i\right)^T \mathbf{M}_i \left(p^i - p_j^i\right) = 0, \ \ \forall 1 \le j \le n_i \end{aligned} \tag{9}$$

Proof 2. Revisit Eqn. 8, its equality constraints propose to collapse all $p_r^i \in \mathcal{P}_i$ together. Therefore keeping only the equality constraints related to the anchor sample p^i achieves the same collapsing performance. So as to the inequality constraints in Eqn. 8. Finally, we can reduce the constraint size by only keeping the constraints related to p^i as in Eqn. 9. The re-written objective Eqn. 9 has only linear-scale $O(n)$ constraints, compared to the original quadratic-scale $O(n^2)$ constraints in Eqn. 8.

An important merit of Eqn. 9 is that it can be efficiently optimized by solving a much easier version [48]:

Theorem 3. All the vectors $p^i - p^i_j$ in Eqn. 9 form a spanning space $\mathbf{S} = span(\sum_j \lambda_j (p^i - p^i_j))$. The Eqn. 9 is equivalent to replace $p^i - y_j$ by t_j, the projection of $p^i - y_j$ in $\mathbf{S}^{\perp}$, that $\mathbf{S}^{\perp}$ is the orthogonal space of $\mathbf{S}$.

Proof 3. Since $\mathbf{M}_i$ is positive semi-definite, the constraint $(p^i - p^i_j)^T \mathbf{M}_i (p^i - p^i_j) = 0$ is equivalent to $\mathbf{M}_i (p^i - p^i_j) = 0$ which means the $\mathbf{M}_i s = 0$ for all $s \in \mathbf{S}$. Projecting $p^i - y_j$ to $\mathbf{S}$ and $\mathbf{S}^{\perp}$ generates two orthogonal bases s_j and t_j respectively, so $p^i - y_j = s_j + t_j$. Replace the inequality constraints in Eqn. 9 by $s_j + t_j$:

$$\begin{aligned} &\left(p^i - y_j\right)^T \mathbf{M}_i \left(p^i - y_j\right) = \left(s_j + t_j\right)^T \mathbf{M}_i \left(s_j + t_j\right) \\ &= {t_j}^T \mathbf{M}_i t_j \end{aligned} \tag{10}$$

Now Eqn. 9 has an equivalent form as:

$$\begin{aligned} &\mathbf{M}_i = \arg\min_{\mathbf{M}_i} \frac{1}{2} \|\mathbf{M}_i\|^2 \\ &sub\ \ to: \ \ \mathbf{M}_i \succeq 0 \\ &{t_j}^T \mathbf{M}_i t_j \geq 2, \ \ \forall 1 \leq j \leq k \\ &\mathbf{M}_i s = 0, \ \ \forall s \in \mathbf{S} \end{aligned} \tag{11}$$

Finally, we prove that Eqn. 11 has the same solution to Eqn. 8 by eliminating its PSD and equality constraints.

Theorem 4. The solution to Eqn. 8 is exactly the same as solving the Eqn. 11 by relaxing its equality and PSD constraints, since they are indeed off-the-shelf.

Proof 4. If we get rid of the PSD and equality constraints in Eqn. 11, the new form is:

$$\begin{aligned} &\mathbf{M}_i = \arg\min_{\mathbf{M}_i} \frac{1}{2} \|\mathbf{M}_i\|^2 \\ &sub\ \ to: \ \ {t_j}^T \mathbf{M}_i t_j \geq 2, \ \ \forall 1 \leq j \leq k \end{aligned} \tag{12}$$

Eqn. 12 is exactly the same form of the objective in Eqn. 5 which can be efficiently solved via a kernel-SVM solver. Thus the positive semi-definiteness of $\mathbf{M}_i$ is guaranteed by Theorem. 1. For the equality constraints in Eqn. 11, given a member s of $\mathbf{S}$, we have:

$$\mathbf{M}_i s = \left(\sum \alpha_i t_i \cdot t_i^T\right) s = \sum \alpha_i t_i \cdot (t_i^T s) = 0 \tag{13}$$

which proves that the solution to Eqn. 12 satisfies the equality constraints as well.

3.4 Visual Instance Retrieval via OLMANS

On the online testing stage, for a probe p_i from $\mathcal{P}$ and one gallery image g_j from $\mathcal{G}$, the similarity matching between p_i and g_j is measured by combining the original baseline models (with flexible choices) with our local metric adaptation $\mathbf{M}_i$ to achieve an adaptive nonlinear metric:

$$\begin{aligned} &D_{\mathbf{M}_i}(p_i, g_j) \\ &= \|p_i - g_j\|^2 + \lambda \|p_i - g_j\|^2_{\mathbf{M}_i} \\ &= \left(p_i - g_j\right)^T \left(\mathbf{I} + \lambda \mathbf{M}_i\right) \left(p_i - g_j\right) \end{aligned} \tag{14}$$

where $\mathbf{M}_i$ is the learned local metric specific for p_i and λ is the weighting parameter. In this paper, we set λ by Eqn. 15 in all the experiments which can be explained in Sec. 5.

$$\lambda = \max_{1 \leq j \leq m} \left(\|p_i - g_j\|^2\right) / \max_{1 \leq j \leq m} \left(\|p_i - g_j\|^2_{\mathbf{M}_i}\right) \tag{15}$$

We find that even simply using only the learned local metric for retrieval, the results are still much better than using the original global baselines. Further, when combining the global baseline and our learned local metrics, we are able to obtain much better and more stable performances. The reason behind it can be explained by the idea of boosting [50]. Either the global baseline or the local metric can be considered as a "weak" classifier for retrieval, and their combination forms a "stronger" classifier with better and more robust performance.

4 Theoretical Analysis and Justification

In this section, we first prove that the asymptotic error of VIR by using the proposed OLMANS is bound to be lower than that without. When the negative samples are truly hard negative ones, the asymptotic error by using OLMANS can be very close to the Bayesian error (Sec. 4.1). Besides this theoretically meaningful result, we prove that this strong asymptotic error can actually best approximated by using finite data, which is practically also meaningful. More importantly, we prove that this approximation is actually achieved by OLMANS (Sec. 4.2). We also present its consistency and sample complexity analysis in Sec. 4.3.

4.1 Asymptotic Error is Reduced

The core of VIR is indeed a 2-class (ω_+ and ω_-) 1-Nearest neighbor (NN) classification problem by using the gallery set $\mathcal{D}$. If there is infinite number of data, it is well-known that its asymptotic error $\mathbb{P}(e|x)$ is bounded by 2 times the Bayesian error [51]:

$$\mathbb{P}^* \leq \mathbb{P}(e|x) = 2P(\omega_+|x)P(\omega_-|x) \leq 2\mathbb{P}^* \tag{16}$$

where $\mathbb{P}^*$ is the Bayesian error. In our work, we prove that by adding the hard negative samples x_a to $\mathcal{D}$ to form an augmented dataset $\mathcal{D}^a$, the asymptotic error $\mathbb{P}^a(e|x)$ by using $\mathcal{D}^a$ is always smaller than $\mathbb{P}(e|x)$:

$$\mathbb{P}^a(e|x) \leq \mathbb{P}(e|x) \tag{17}$$

Theorem 5. For an input x, its NN is x' in $\mathcal{D}^a$. Define the probability that x' is an augmented data x_a, i.e., $x' \sim x_a$ as $P(x' \sim x_a) = q$; otherwise, x' is not an augmented data x_a, i.e., $x' \neg x_a$, $P(x' \neg x_a) = 1 - q$, where $0 \leq q \leq 1$. The asymptotic error $\mathbb{P}^a(e|x)$ by using $\mathcal{D}^a$ is:

$$\mathbb{P}^a(e|x) = \frac{(2-q)\mathbb{P}(e|x)}{2 - 2q\mathbb{P}(e|x)} \leq \mathbb{P}(e|x) \tag{18}$$

The proof is provided in Appendix.A. Since q is the probability of $P(x' \sim x_a)$, we have $0 \leq q \leq 1$. If $q = 0$ which indicates that the augmented negative data are useless, then we have $\mathbb{P}^a(e|x) = \mathbb{P}(e|x)$. Another extreme is when $q = 1$ implying the negative data are abundant and effective to constrain the classification, then we have [1]

$$\mathbb{P}^a(e|x) = \frac{\mathbb{P}(e|x)}{2[1 - \mathbb{P}(e|x)]} \leq \mathbb{P}(e|x) \tag{19}$$

[1] $\mathbb{P}(e|x) \leq \frac{1}{2}$ is always true.

In this case, when $\mathbb{P}(e|x)$ is very small, we have

$$\mathbb{P}^a(e|x) \simeq \frac{\mathbb{P}(e|x)}{2} \simeq \mathbb{P}^*(e) \tag{20}$$

The asymptotic error of our negative-augmented approach can be very close to the Bayesian error.

4.2 Finite Approximation to $\mathbb{P}^a(e|x)$

The asymptotic error $\mathbb{P}^a(e|x)$ in Eqn. 18 is only meaningful when the sample size is infinite, $n \to \infty$. However, in practice, only finite number of samples are available. To make it practically meaningful, we prove that it can be best approximated by the practical error rate $\mathbb{P}_n(e|x)$ (n is finite) by finding a local metric $\mathbf{M}_x$. And this local metric turns out to be the one for the proposed OLMANS.

Still consider the 2-class 1-NN rule scenario (on the negative augmented data $\mathcal{D}^a$). To make the notation less cluttered, here we use $\mathbb{P}(e|x)$ to indicate $\mathbb{P}^a(e|x)$ without confusion. Given a sample x and its nearest neighbor x' from the finite dataset containing n samples. The probability of error for x is:

$$\begin{aligned}\mathbb{P}_n(e|x) &= P(\omega_+|x)P(\omega_-|x') + P(\omega_-|x)P(\omega_+|x') \\ &= \mathbb{P}(e|x) + [P(\omega_+|x) - P(\omega_-|x)][P(\omega_+|x) - P(\omega_+|x')]\end{aligned}$$

Our goal is to find a best local metric $\mathbf{M}_x$ for x such that the conditional MSE $\min_{\mathbf{M}_x} \mathbb{E}\{[\mathbb{P}_n(e|x) - \mathbb{P}(e|x)]^2|x\}$ is minimized. Since $[P(\omega_+|x) - P(\omega_-|x)]$ is constant for a given x, so the minimization is equal to:

$$\min_{\mathbf{M}_x} \mathbb{E}\{[P(\omega_+|x) - P(\omega_+|x')]^2|x\} \tag{21}$$

Because $P(\omega_+|x') \simeq P(\omega_+|x) + \nabla P(\omega_+|x)^T(x' - x)$, Eqn. 21 is approximately equivalent to:

$$\min_{\mathbf{M}_x} \mathbb{E}\{\|\nabla P(\omega_+|x)^T(x' - x)\|^2|x\} \tag{22}$$

The core here is to compute the gradient of posterior $\nabla P(\omega_+|x)$. Recall our proposed OLMANS approach, a local linear classifier $\mathbf{w}$ where $\mathbf{M}_x = \mathbf{w}\mathbf{w}^T$ is learned for a sample x. So the posterior of x in a logistic sigmoid function form is:

$$P(\omega_+|x) = \frac{1}{1 + e^{\zeta_x(\mathbf{w}^T x + b) - \gamma}}, P(\omega_-|x) = 1 - P(\omega_+|x) \tag{23}$$

The gradient of $P(\omega_+|x)$ can be easily computed:

$$\nabla P(\omega_+|x) = \zeta_x P(\omega_+|x)P(\omega_-|x)\mathbf{w} \tag{24}$$

Substituting Eqn. 24 for $\nabla P(\omega_+|x)$ in Eqn. 22 gives us:

$$\begin{aligned}&\min_{\mathbf{M}_x} \mathbb{E}\{\|\zeta_x P(\omega_+|x)P(\omega_-|x)\mathbf{w}^T(x' - x)\|^2|x\} \\ &= \min_{\mathbf{M}_x}(x' - x)^T\mathbf{w}\mathbf{w}^T(x' - x)\end{aligned} \tag{25}$$

Recall our optimization objective Eqn. 5, for the positive samples, we have $1 - (x' - x)^T\mathbf{M}_x(x' - x) \geq 1$ which is equal to $(x' - x)^T\mathbf{M}_x(x' - x) \leq 0$. On the other hand, $(x - x')^T\mathbf{M}_x(x - x') \geq 0$ is always true for a PSD $\mathbf{M}_x$, so $(x' - x)^T\mathbf{M}_x(x' - x) \equiv 0$ always holds. It is obvious Eqn. 25 is always optimized by adopting the local metric $\mathbf{M}_x$ learned by our algorithm Eqn. 5.

4.3 Consistency and Sample Complexity Analysis

A set of samples $\{x_0, x_1, ..., x_k\}$ is identically drawn from a D-dimensional space $\mathbb{D} \in \mathbb{R}^D$ where l_i is the label of x_i, then a paired sample set $S_k^{pair} = \{s_i\}_{i=1}^k = \{(x_0, x_i)\}_{i=1}^k$ of size k is formed. For our proposed objective Eqn. 5, the true risk over the whole distribution $\mathbb{D}$ and the empirical error based on S_k^{pair} are defined as:

$$\begin{aligned}Err^\lambda(\mathbf{M}_x, \mathbb{D}) &= \mathbb{E}_{x_i, x_j \sim \mathbb{D}}\phi^\lambda(\mathbf{M}_x, (x_i, x_j)) \\ Err^\lambda(\mathbf{M}_x, S_k^{pair}) &= \frac{1}{k}\sum_{i=1}^{k}\phi^\lambda(\mathbf{M}_x, s_i)\end{aligned}$$

where $\phi^\lambda\,(\mathbf{M}_x, s_i)$ is the hinge loss function:

$$\phi^\lambda(\mathbf{M}_x, s_i) = \lambda[\zeta_i\left((x_i - x_0)^T\mathbf{M}_x(x_i - x_0)\right) - \gamma_{\zeta_i}]_+$$

where $\zeta_i = -1$ if $l_i = l_0$ and 1 otherwise, $[A]_+ = \max(0, A)$ is the hinge loss and γ_{ζ_i} is the desired margin. The empirical risk minimizing metric based on S_k^{pair} can be readily defined as $\mathbf{M}_x^* = \arg\min_{\mathbf{M}_x} Err^\lambda(\mathbf{M}_x, S_k^{pair})$. Our goal is to compare the generalization performance of $\mathbf{M}_x^*$ over the unknown $\mathbb{D}$.

Theorem 6. Let $\phi^\lambda(\mathbf{M}_x, s_i)$ be a distance-based loss function that is λ-Lipschitz in the first argument. Then with probability at least $1 - \delta$ over $\{s_1, ..., s_k\}$ from an unknown B-bounded-support (each $(x, l) \sim \mathbb{D}$, $||x|| \leq B$) distribution $\mathbb{D}$, we have:

$$\begin{aligned}&\sup_{\mathbf{M}_x \in \mathcal{M}}\left[Err^\lambda(\mathbf{M}_x, \mathbb{D}) - Err^\lambda(\mathbf{M}_x, S_k^{pair})\right] \\ &\leq O\left(\lambda B^2\sqrt{D\ln(1/\delta)/k}\right)\end{aligned} \tag{26}$$

Theorem. 6 proves that to achieve an estimation error rate ϵ, $k = \Omega\left((\lambda B^2/\epsilon)^2 D\ln(1/\delta)\right)$ samples are sufficient. The brief proof is shown in Appendix.B.

Theorem 7. Let $\mathbf{M}_x$ be any class of weighting metrics on the feature space $X = \mathbb{R}^D$, and define $d := \sup_{\mathbf{M}_x \in \mathcal{M}} \|\mathbf{M}_x\|_F^2$. Following the same parameter setting in Theorem. 6, we have:

$$\begin{aligned}&\sup_{\mathbf{M}_x \in \mathcal{M}}\left[Err^\lambda(\mathbf{M}_x, \mathbb{D}) - Err^\lambda(\mathbf{M}_x, S_k^{pair})\right] \\ &\leq O\left(\lambda B^2\sqrt{d\ln(1/\delta)/k}\right)\end{aligned} \tag{27}$$

Let $\mathbb{P}$ be the probability measure induced by the random variable $(X; \mathcal{L})$, where $X := (x, x')$, $\mathcal{L} := 1[l = l']$. Define function class:

$$\mathcal{F} := \{X \mapsto \|x - x'\|_{\mathbf{M}_x}\}$$

Following the same steps in the proof of Theorem. 6, we can conclude that the Rademacher complexity of $\mathcal{F}$ is bounded. In particular,

$$\mathcal{R}_k(\mathcal{F}) \leq 4B^2\sqrt{\frac{\sup_{\mathbf{M}_x \in \mathcal{M}}\|\mathbf{M}_x\|_F^2}{k}}$$

Finally, we note that ϕ^λ is λ-Lipschitz in the first argument, so that we can readily apply Theorem.8 in [52].

From Theorem. 7, we observe that if the learned metric $\mathbf{M}_x$ has a low metric learning complexity $d \ll D$, it can help sharpen the sample complexity result, yielding a dataset-dependent bound. Recall our objective Eqn. 5, $d := \sup_{\mathbf{M}_x \in \mathcal{M}} \|\mathbf{M}_x\|_F^2$ is already optimized via our proposed learning objective. Therefore, the bound is further tighter under the same number of samples.

4.4 OLMANS vs Re-ranking

Both our proposed OLMANS algorithm and the widely-used re-ranking technique can be readily combined with any offline learned retrieval models in the online phase for further performance improvement. But our OLMANS owns more unique merits than re-ranking in both the efficiency and effectiveness facets which has been both theoretically proved in Sec. 4 and empirically verified by extensive experiments in Sec. 5.

Data Requirement: Most re-ranking methods require no additional learning samples, but utilize the given query probe and gallery samples to help refine the ranking. In contrast, our OLMANS takes advantage of a set of easily-available negative samples, based on which it finds online adaptation for the optimal local metric.

Effectiveness: The effectiveness of re-ranking depends heavily on the quality of the initial ranking list (if the true match is not in the top-k ranks). It may hurt the initial rank result, because the true match may have a lower rank after re-ranking if the false matches are included in the top-k list. Thus re-ranking may degrade the performance. The performance of our OLMANS model relies on the quality of the set of negative data, as illustrated by Theorem. 5, even if the quality of the given NDB is pretty bad (no hard negatives are provided), OLMANS still won't degrade the original performance. Comparing to re-ranking, our OLMANS has a unique and plausible advantage: it does not degrade the performance of the original methods (the original global metric) in theory. As indicated in the objective Eqn. 3, when the negative samples are not good (i.e., they are already far away from the positive point in the original feature space), the learned local metric $\mathbf{M}_x$ will be the same as the original baseline, since the constraints in Eqn. 3 have already been fulfilled. So OLMANS won't give a worse performance than the original method. As described in Sec. 4, our theoretical analysis has shown that asymptotically our negative-augmented approach always improves the identification performance, and can be very close to the Bayesian error.

Efficiency: Another merit of our OLMANS compared with re-ranking is its high efficiency. OLMANS is very efficient even if there are a lot of negative samples available for local adaptation. Because the learned local metric $\mathbf{M}_x$ is only related to a handful set of hard negatives, not all the negatives. In contrast, other methods, such as re-ranking (depend on data number and nearest neighbor number k), transfer learning, domain adaptation techniques, are usually time-consuming because the affinity relationships among probes and gallery samples have to be computed.

5 Experiments

In this section, to verify the efficiency and effectiveness of our proposed OLMANS method, we evaluate our method on two generic VIR tasks: person re-identification (P-RID) and image retrieval (Img-R).

5.1 Experiment on Person Re-identification

5.1.1 Experiment Settings

Data. We perform thorough experiments and comparative studies to evaluate our method on most widely-used P-RID benchmark datasets: VIPeR [36], GRID [53], CUHK03 [20], Market1501 [54], DukeMTMC-reID [21] and MSMT17 [22]. The statistic details of the above datasets are summarized in Table. 1. For VIPeR and GRID datasets, all the identity pairs are randomly divided into half for training and the other half for testing so that the average results of 10 random trials are reported. For CUHK03, the newly proposed protocol [46] (767 identities are used for training as well as the left 700 identities are used for testing) is adopted in our experiments. As for the other three benchmarks, Market1501, DukeMTMC-reID and MSMT17, the pre-determined probe and gallery sets are directly utilized with no modification.

Evaluation. For a fair comparison, the training data of each dataset are used as the negative training samples for itself, so no more extra information is utilized in the experiment. For all the experiments, the single-shot evaluation setting is adopted and results are shown in the form of Cumulated Matching Characteristic (CMC) curves. Besides, the mean average precision (mAP) results of the latter four benchmarks are also reported.

Feature. Both handcrafted features and learned deep features are explored in our experiments. The high-dimensional handcrafted P-RID feature called LOMO [6] is adopted. Since it is not practical to directly use such a high dimensional feature (26960-dim for the original LOMO feature) in metric learning, we employ principal component analysis (PCA) to reduce the feature dimension to a reasonable scale (1000-dim after PCA). Besides, our proposed algorithm is directly applied to various CNN features presented below for evaluation.

Baseline. Since the global metric learning-based methods perform much better than deep learning-based ones on the small-scale datasets VIPeR and GRID, due to the lack of sufficient training data, we mainly focus on the state-of-the-art global metric learning approaches [5], [6], [11] as our baseline models. As for the other large-scale datasets with plenty of training samples, the state-of-the-art CNN-based P-RID models are selected as our baselines to implement our method on including CaffeNet [55], VGG16 [56], ResNet50 [57], DenseNet121 [14] and HA-CNN [13]. Besides, the other state-of-the-art P-RID methods [15], [16], [17], [18], [58], [59], [60] are further compared for a complete evaluation. Finally, a recently proposed state-of-the-art re-ranking approach [46] is compared with our algorithm. Various ablation studies of our proposed model are explored in Sec. 5.1.4.

5.1.2 Comparisons with State-of-the-art

Experiments on VIPeR: The small-size VIPeR dataset is a widely-used benchmark for P-RID which contains 632 pedestrian image pairs taken from 2 different cameras in an outdoor environment. We conduct the comparison experiment under the same experiment setting and using the same LOMO feature, while the global metric learner MLAPG [5] is selected as our baseline. The results are reported in Table. 2. Our method achieves the best performances on all the ranks. For the important Rank@1 evaluation, our performance 44.97% outperforms the second best approach LSSCDL by 2.31% and the baseline model MLAPG by 4.24%. This promising performance indicates that the proposed local metric adaptation method is consistently effective, several representative examples are shown in Fig. 2. One interesting observation is our improvement performance at Rank@20 is a little bit lower than its performance at Rank@1. This is expected as our local metric becomes less effective when the true positive gallery image is far from the probe in the feature space. Nevertheless, our method still beats all the other approaches at Rank@20.

Experiments on GRID: The GRID dataset [53] contains 250 pedestrian image pairs taken from 8 disjoint camera views and 775 additional images that do not belong to the 250 persons. GRID is a pretty tough dataset because of the large viewpoint variations, the

TABLE 1
The statistics of different P-RID benchmarks.

Dataset	VIPeR	GRID	CUHK03	Market1501	DukeMTMC	MSMT17
#Train-IDs	316	125	767	751	702	1040
#Probe-IDs	316	125	700	750	702	3060
#Gallery-IDs	316	775	700	751	1110	3060
#cam	2	8	2	6	8	15
#images	1264	1025	28192	32668	36411	126441

TABLE 2
Comparison results with the global metric learning methods on VIPeR using the same LOMO feature. RED is the best result and BLUE is the second best one.

Method	R@1	R@5	R@10	R@20
Ours(MLAPG)	44.97	74.43	84.97	93.64
LSSCDL [27]	42.66	-	84.27	91.93
DNSL [11]	42.28	71.46	82.94	92.06
MLAPG [5]	40.73	69.94	82.34	92.37
XQDA [6]	40.00	68.13	80.51	91.08
TMA [61]	39.88	-	81.33	91.46
KISSME [62]	34.81	60.44	77.22	86.71
ITML [63]	24.64	49.78	63.04	78.39
LMNN [64]	29.43	59.78	73.51	84.91
kCCA [65]	30.16	62.69	76.04	86.80
MFA [66]	38.67	69.18	80.47	89.02
kLFDA [66]	38.58	69.15	80.44	89.15

TABLE 3
Comparison with the global metric learning methods on GRID using the same LOMO feature.

Method	R@1	R@5	R@10	R@20
Ours(MLAPG)	30.16	42.64	49.20	59.36
LSSCDL [27]	22.40	-	51.28	61.20
DNSL [11]	15.12	31.92	40.72	53.12
MLAPG [5]	17.60	33.52	43.36	56.08
XQDA [6]	12.96	26.80	34.56	43.52
EPKFM [67]	16.30	35.80	46.00	57.60
MtMCML [28]	14.08	34.64	45.84	59.84
PRDC [7]	9.68	22.00	32.96	44.32

low-resolution image quality and the quantitative distractors. The average performance of 10 random trials is provided in Table. 3. It can be clearly observed that our proposed algorithm outperforms all the existing algorithms at Rank@1 by a very significant 7.8% improvement on the identification rate. From the results we can see that the GRID dataset is more challenging than VIPeR, but our proposed algorithm can still handle it well by adapting the local similarity structure of each probe.

Experiments on CUHK03: The CUHK03 is a large-scale dataset which contains 13164 images of 1360 pedestrians. All the images are captured by six surveillance cameras over months. Each person is observed by two disjoint camera views with an average of 4.8 images in each view. In our experiments, three state-of-the-art CNNs including ResNet50, DenseNet121 and HA-CNN are selected as our baselines to extract features of testing data and our proposed OLMANS is directly applied to them. The comparison results under the newly proposed splitting protocol is shown in Table. 4. For all the three baselines, our method further improves the Rank@1 and mAP performances by a large margin (over 14% on Rank@1 and 11% on mAP) to a state-of-the-art level. The results verify that our proposed OLMANS is not only suitable to the handcrafted features, but also works well for the state-of-the-art deep features.

Evaluation on Market1501: Market1501 is a large-scale P-RID benchmark which contains 32668 bboxes of 1501 identities. Each person is recorded by six cameras at most, and two at least. Table. 4 shows the comparison results of our OLMANS on the baselines and against the state-of-the-art results. Although the most recent approaches have achieved a pretty high performance ($\geq$ 90%) on Market1501, the improvement of our method is over 4% and 6% on Rank@1 and mAP for all the three baselines by handling the "hard" probe samples well.

Evaluation on DukeMTMC-reID: DukeMTMC-reID dataset is a recent large-scale benchmark to date proposed for P-RID, but the lasted methods have obtained promising performances. As show in Table. 4, the recently published methods, SPreID [70], PCB [18] and Part-aligned [60], boost the state-of-the-art to 85.9%(73.3%) on Rank@1(mAP). By implementing our OLMANS on HA-CNN, the Rank@1(mAP) result is boosted from 80.7%(64.4%) to 83.9%(69.0%), which approaches the state-of-the-art performance.

Evaluation on MSMT17: MSMT17 [22] is the latest and largest P-RID benchmark so far. The extreme large-scale identities and a large number of distractors make this dataset pretty challenging. We evaluate the performance of the baselines on MSMT17 dataset with(w/) and without(w/o) our algorithm in Table. 5. Our method improves the Rank@1(mAP) performance of DenseNet121 from 66.0%(34.6%) to a state-of-the-art result 75.5%(43.1%). Such results demonstrate that our proposed OLMANS is scalable to the size of dataset, even a large number of testing probes are given, the efficient optimization scheme and theoretical analyses guarantee the performance of our proposed OLMANS.

5.1.3 *Comparison with Re-ranking*

As we discussed in Sec. 4.4, both our proposed OLMANS and the re-ranking technique can be applied to any offline learned P-RID baselines for further online performance improvement. In this part, we evaluate our proposed OLMANS and a state-of-the-art re-ranking method (RR) [46] on the CUHK03, Market1501 and DukeMTMC-reID datasets by selecting two CNN-based P-RID models, HA-CNN [13] and Dense121 [14] as baselines. The comparison results in Table. 6 show that our method can improve the baseline performance significantly at both Rank@1 and mAP evaluations. Compared with [46], our OLMANS works better on improving Rank@1 performance and has comparative improvement on the mAP evaluation since [46] considers the k-reciprocal nearest neighbors of both probe and extra 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. However, our method only utilizes the given query probes and a

TABLE 4
Comparison results on CUHK03, Market1501, and DukeMTMC-reID. **All the results are the best performances reported in their literatures**.

	CUHK03			Market1501			DukeMTMC-reID	
Method	**R@1**	**mAP**	**Method**	**R@1**	**mAP**	**Method**	**R@1**	**mAP**
Ours(ResNet50)	**59.4**	**54.8**	**Ours(ResNet50)**	**91.1**	**76.8**	**Ours(ResNet50)**	**79.1**	**63.5**
Ours(DenseNet121)	**53.1**	**49.3**	**Ours(DenseNet121)**	**90.9**	**75.4**	**Ours(DenseNet121)**	**80.2**	**64.1**
Ours(HA-CNN)	**62.6**	**58.3**	**Ours(HA-CNN)**	**93.8**	**81.1**	**Ours(HA-CNN)**	**83.9**	**69.0**
ResNet50 [57]	47.9	46.8	ResNet50 [57]	88.5	71.3	ResNet50 [57]	77.7	58.8
Dense121 [14]	41.0	40.1	Dense121 [14]	88.2	69.2	Dense121 [14]	78.6	58.5
HA-CNN [13]	48.0	47.6	HA-CNN [13]	90.6	75.3	HA-CNN [13]	80.7	64.4
PCB [18]	63.7	67.5	PCB [18]	83.3	69.2	PCB [18]	83.3	69.2
SVDNet [58]	41.5	37.3	SVDNet [58]	82.3	62.1	SVDNet [58]	76.7	56.8
DPFL [68]	40.7	37.0	DNSL [11]	61.0	35.6	DuATM [69]	81.8	64.6
Mancs [33]	69.0	63.9	Mancs [33]	93.1	82.3	SPreID [70]	85.9	73.3
PAN [71]	36.3	34.0	Part-aligned [60]	91.7	79.6	Part-aligned [60]	84.4	69.3
MLFN [59]	52.8	47.8	PN-GAN [72]	77.1	63.6	PAN [71]	71.6	51.5
DaRe [73]	55.1	51.3	DeepCC [74]	89.5	75.7	GAN [21]	67.7	47.1

TABLE 5
State-of-the-art comparison results on on MSMT17. **All the results are the best performances reported in their literatures**.

Method	**MSMT17**		
	R@1	**R@20**	**mAP**
Ours(ResNet50)	**72.8**	**88.6**	**55.0**
Ours(DenseNet121)	**75.5**	**89.9**	**43.1**
Ours(HA-CNN)	**68.0**	**87.8**	**37.8**
SqueezeNet [15]	30.6	N/A	13.0
MobileNetv2 [16]	44.9	N/A	21.1
SuffleNet [17]	39.6	N/A	17.8
ResNet50 [57]	63.4	86.1	34.2
DenseNet121 [14]	66.0	86.6	34.6
HA-CNN [13]	61.8	85.8	34.6

TABLE 7
The influence of baseline metric choice. **+Ours** means implementing our OLMANS on the baselines. Red represents the better results.

Baselines	**GRID**		**VIPeR**	
	R@1	**R@20**	**R@1**	**R@20**
Euc	9.12	29.76	15.32	50.66
Euc+Ours	20.88	45.12	21.99	56.11
XQDA	12.96	43.52	38.99	91.94
XQDA+Ours	29.20	50.96	43.54	92.15
MLAPG	17.60	56.08	40.28	93.39
MLAPG+Ours	30.16	59.36	44.97	93.64
DNSL	15.12	53.12	40.19	93.54
DNSL+Ours	28.96	56.96	43.67	93.61

TABLE 6
Comparison with the state-of-the-art re-ranking method. **Rank@1(mAP)** result is reported. Red represents the best result.

Method	**CUHK03**	**Market1501**	**DukeMTMC**
HA-CNN	48.0(47.6)	90.6(75.3)	80.7(64.4)
HA-CNN+RR	54.8(55.7)	91.4(79.0)	82.5(69.9)
HA-CNN+Ours	62.3(56.5)	92.7(79.0)	83.7(67.8)
Dense121	41.0(40.1)	88.2(69.2)	78.6(58.5)
Dense121+RR	48.1(51.5)	90.2(85.0)	83.7(76.9)
Dense121+Ours	53.1(49.3)	90.4(74.0)	84.2(67.1)

set of negative samples to gain a large improvement of the baseline performance.

5.1.4 Ablation Study

(1) Influence of Baseline Quality: Our proposed OLMANS algorithm is applied on top of an offline-learned baseline, thus its overall performance may depend on the learning quality of adopted baseline. In order to verify whether our OLMANS can always be helpful, baseline models obtained at various learning stages of a global metric learner [5] are tested, as in general the performance of the baseline learner improves with more training (e.g., more training iterations). As shown in Fig. 5, even the learned global metric does perform poorly (in its early training stages), our online local metric adaptation is able to consistently and significantly improve the performances by a large margin. This is because the local discriminative information introduced by hard negative samples is able to capture the specific crux of one identity which is quite helpful for identification.

(2) Influence of Baseline Metric Choice: An interesting question is whether our OLMANS can always work for any baselines as promised. To verify it, we conduct the following experiment that different kinds of global metric learners, Euclidean distance, XQDA [6], MLAPG [5] and DNSL [11] are adopted for the LOMO feature as the underlying baselines that our OLMANS algorithm is readily applied on. The results on VIPeR and GRID datasets are reported in Table. 7, as well as the complete CMC curves in Fig. 6. We observe that for all the learners, our proposed OLMANS algorithm is able to boost the identification performance with a significantly improvement, even double the Rank@1 performance (on GRID).

(3) Influence of Baseline Feature Choice: We evaluate various feature descriptors for P-RID to verify that the performance of our OLMANS is independent of the choice of feature. Both the hand-crafted features, LOMO [6] and deep features, CaffeNet [55], VGG16 [56] and ResNet50 [57] are examined. The above pre-trained CNN models from which we have removed the final fully-connected (FC) layer are further fine-tuned by the large-scale Market1501 datasets [2], then they are directly used to extract the features for VIPeR and GRID datasets. As can be seen from Table. 8, the performance improvement by our OLMANS method is independent of the used feature descriptors.

(4) Influence of the Weighting Parameter λ: The parameter λ in Eqn. 14 is used to balance the underlying baseline and the

[2] The Rank@1(mAP) performances are: CaffeNet = 44.31(24.0), VGG16 = 63.93(42.5) and ResNet50 = 77.22(56.1)

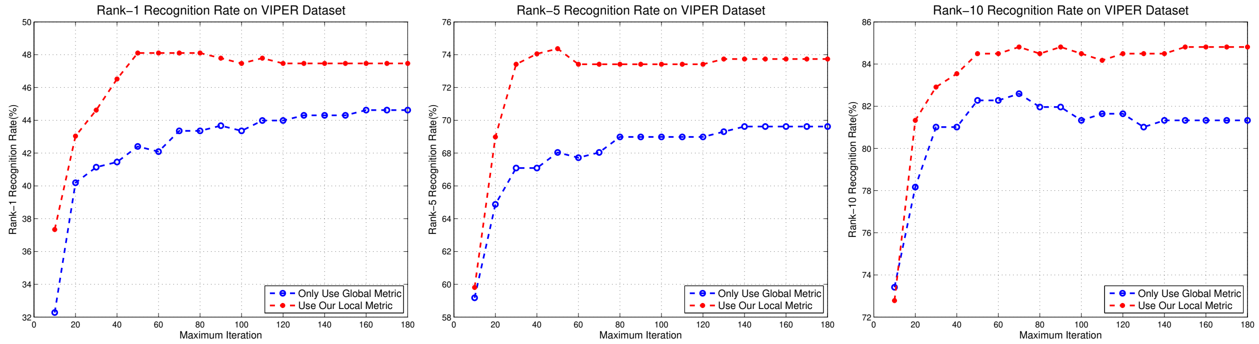


Fig. 5. The influence of baseline quality. The x-axis means the maximum iteration time for offline learning and the y-axis is the identification rate (Rank@1, Rank@5 and Rank@10 on VIPeR).

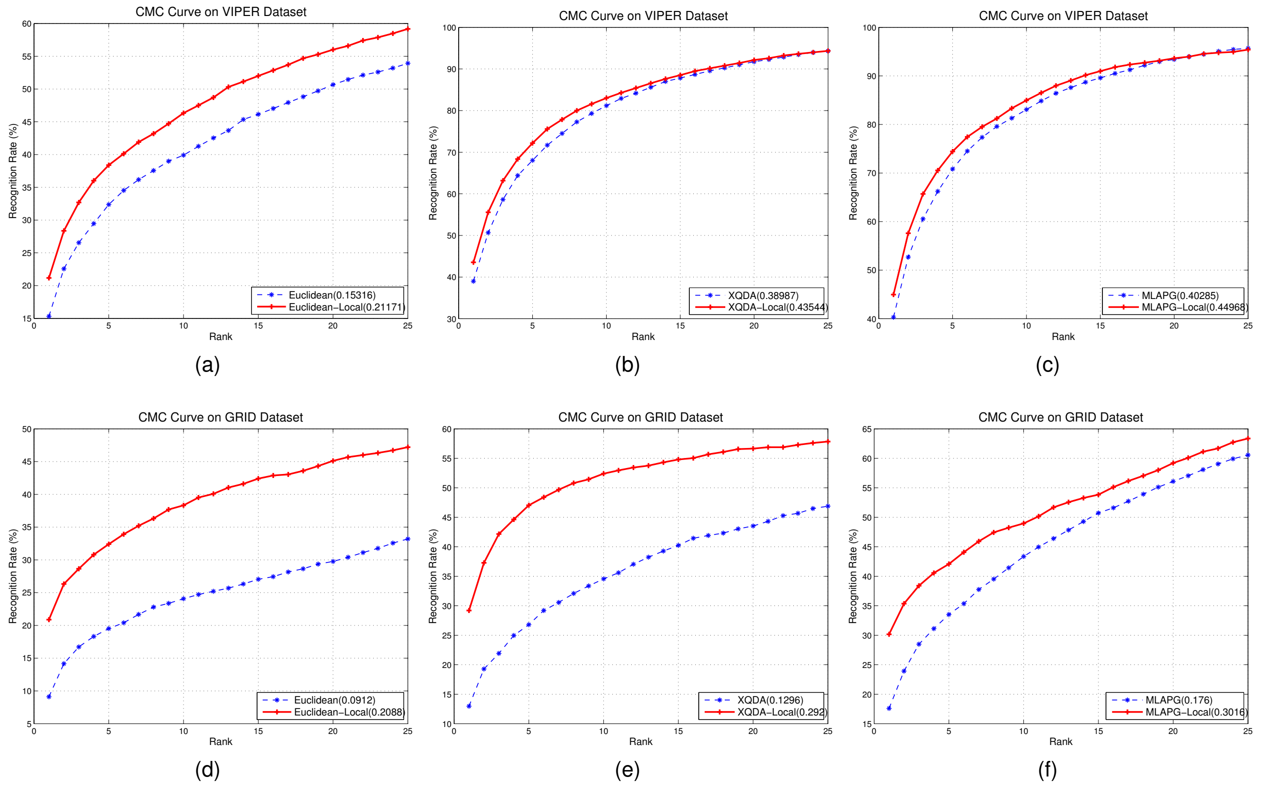


Fig. 6. The influence of baseline metric choice. (a) and (d) are the results on VIPeR and GRID directly using the Euclidean distance; (b) and (e) are XQDA [6] results; (c) and (f) are MLAPG [5] results.

TABLE 8
The influence of baseline feature choices on VIPeR and GRID under different metrics (10-folds average Rank@1 performance is reported). For each result, the former one is the baseline result **without** our OLMANS, and the latter is our OLMANS result.

Dataset	Features	Euclidean	MLAPG	XQDA	DNSL
VIPeR	LOMO	15.32/21.99	40.28/44.97	38.99/43.54	40.19/43.67
	CaffeNet	17.72/21.84	18.35/19.30	20.41/28.16	20.38/23.26
	VGG16	20.25/26.27	20.25/23.73	23.45/29.02	23.86/26.52
	ResNet50	22.78/27.22	23.42/26.58	31.93/40.47	33.70/38.01
GRID	LOMO	9.12/20.88	17.60/30.16	12.96/29.20	15.12/28.96
	CaffeNet	2.40/13.60	5.60/10.42	10.24/21.92	7.28/16.72
	VGG16	6.40/18.44	7.20/16.84	12.72/21.52	10.24/17.36
	ResNet50	12.84/23.22	12.40/19.12	21.44/34.96	17.36/29.44

learned local metric. Different λ will have different influences to the identification performance. We conducted an experiment on VIPeR dataset to determine the value of λ, the results of which are shown in Fig.7. We need to point out some special λ values: The $\lambda = 0$ is the baseline result from [5] without our local metric adaptation and $\lambda = max$ represents that λ is set as Eqn. 15. So setting $\lambda = \max_{1\leq j\leq m}\left(\|p_i - g_j\|^2\right)/\max_{1\leq j\leq m}\left(\|p_i - g_j\|^2_{\mathbf{M}_i}\right)$ achieves the best result because it normalizes the norm scales of the baseline and locally adapted distances.

(5) Influence of Negative Sample Database: For our OLMANS, a negative sample database (NDB) is used to provide the negative training data. Because there are various strategies to collect NDB, we conduct the following experiments to investigate

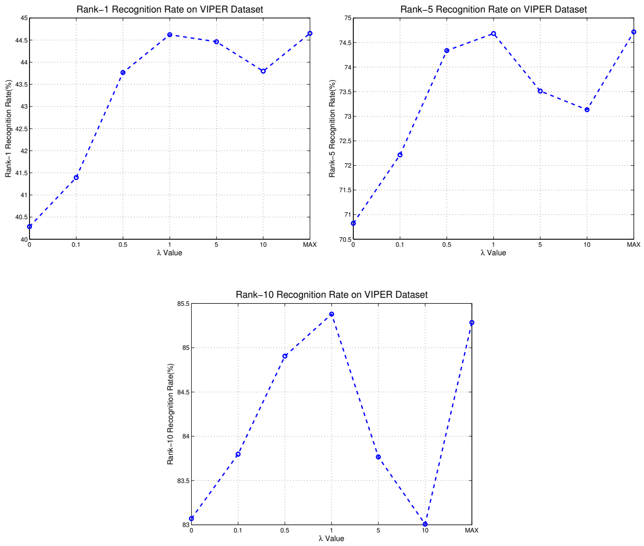


Fig. 7. The influence of parameter λ. The x-axis means the value of λ and the y-axis is the identification rate. The results at Rank@1, Rank@5 and Rank@10 on VIPeR are shown.

TABLE 9
The influence of different NDBs on VIPeR.

Method	**R@1**	**R@5**	**R@10**	**R@20**
Baseline	40.73	69.94	82.34	92.37
Our-D-RAM	39.87	70.51	82.28	91.77
Our-SAME	44.97	74.43	84.97	93.64
Our-D-050	42.63	73.63	84.81	93.54
Our-D-100	43.04	73.86	84.30	93.42
Our-D-500	42.53	73.89	84.15	93.35

TABLE 10
Average Learning time (seconds) on VIPeR.

Method	ITML	MLAPG	LADF
Ave Time	20.5	25.8	31.7
Method	LMNN	PRDC	**OLMANS**
Ave Time	152.9	394.6	4.8

the influences of different NDB choices. The experiments are conducted on VIPeR dataset. Moreover, the global metric learning method MLAPG is adopted as the baseline model.

Using the training data from the same benchmark as the NDB: Here the training samples in VIPeR which have different identities from $\mathcal{P}$(the training data for global metric learning) are used as negative samples. It guarantees that the obtained NDB is clearly meaningful. The P-RID result is given in Table.9 as **Our-SAME**.

Using different benchmark datasets as the NDB: Here we utilize the other benchmark, the GRID dataset as the NDB in our experiment, so that we can guarantee that the identities of all the negative samples in the NDB are different from $\mathcal{P}$. For each probe p_i, the k nearest negative samples are found in the NDB (under the baseline feature) and used for our OLMANS. Different values of k (50, 100, 500) are chosen for further comparisons. The experiment results **Our-D-50/100/500** are shown in Table.9. Moreover, an additional experiment **Our-D-RAM** that uses 50 random negative samples from the NDB for OLMANS is compared. This experiment validates the insight of our method that the effective negative samples are those that are close to the probe in the feature space (e.g., strong false positives).

From Table. 9, it can be observed that **Our-SAME** performs the best because the negative data from the same benchmark dataset are most discriminative. Results on **Our-D-50/100/500** also largely outperform the baseline by consistent improvements. **Our-D-RAM** can not improve the baseline performance since this randomly selected small-size NDB provides no useful hard negatives for OLMANS.

(6) Learning Cost Analysis and Comparison: Although each query probe (or probe set) needs to learn a local Mahalanobis metric on the testing stage, the proposed optimization solution to our OLMANS objective makes the learning efficient and largely reduces the learning time. Table. 10 [3] provides a thorough comparison of average learning time of various state-of-the-art metric learning-based methods on VIPeR dataset. Besides, Table. 11 shows the learning time of different advanced global metric learners on a large-scale dataset, Market1501. All the experiments are conducted on a remote server with an Intel i7-5930K @3.50GHz CPU and 32G memory. The total average learning time of our method on VIPeR is only 4.81 seconds for the adaptation of all the 316 probes, much shorter than learning a single global metric in 25.82 seconds. For the large-scale dataset Market1501, the efficiency advantage of ours is much more pronounced. Our local metric adaptation time is 10 $\sim$ 100 times less than the other global metric learners. So the extra time spent in our OLMANS is indeed nominal compared with learning a global metric.

5.2 Experiment on Image Retrieval

5.2.1 *Experiment Settings*

Data. We evaluate our proposed OLMANS on four widely-used image retrieval benchmarks: the original Oxford [1], Paris [2] and their corresponding revisited datasets $\mathcal{R}$Oxford and $\mathcal{R}$Paris from [3] by correcting annotation mistakes, adding new query images and introducing new evaluation protocols. The Oxford and Paris datasets contain 5063 and 6392 images collected from Flickr associated with Oxford and Paris landmarks respectively. Each dataset contains 55 queries coming from 11 landmarks. For the revisited versions, $\mathcal{R}$Oxford and $\mathcal{R}$Pari, 15 queries from 5 out of the original 11 landmarks are along with the original 55 queries for evaluation.

Evaluation. The training dataset in [4] is used as the NDB. For all the benchmarks, the mean average precision (mAP) results over the query images are reported in our experiments. For $\mathcal{R}$Oxford and $\mathcal{R}$Pari, three new evaluation difficulties, Easy(E), Medium(M) and Hard(H), are evaluated. Since the old setup of Oxford and

TABLE 11
Learning time (seconds) on Market1501.

Method	XQDA	MLAPG	MFA
Train Time	3233.8	2732.8	437.8
Method	kLFDA	DNSL	**OLMANS**
Train Time	995.2	3149.7	19.60

[3] The total learning time of OLMANS includes the local metric adaptation time and retrieval time for all probes.

TABLE 12
Comparison results on Oxford, Paris, $\mathcal{R}$Oxford and $\mathcal{R}$Paris. The mAP results are reported.

Method	**Datasets**					
	Oxford	**Paris**	**$\mathcal{R}$Oxford**		**$\mathcal{R}$Paris**	
			M	**H**	**M**	**H**
Ours(VGG16)	**83.5**	**82.9**	**55.9**	**26.8**	**63.5**	**37.3**
Ours(VGG16-Whiten)	**88.1**	**87.9**	**60.7**	**32.6**	**69.7**	**44.5**
Ours(Res101)	**81.7**	**87.6**	**56.1**	**27.8**	**70.3**	**44.9**
Ours(Res101-Whiten)	**89.3**	**92.6**	**65.7**	**40.4**	**76.9**	**55.4**
MAC [27]	56.4	72.3	37.8	14.6	59.2	35.9
SPoC [11]	68.1	78.2	38.0	11.4	59.8	32.4
CroW [5]	70.8	79.7	41.4	13.9	62.9	36.9
R-MAC [6]	66.9	83.0	42.5	12.0	66.2	40.9
NetVLAD [67]	67.6	74.9	37.1	13.8	59.8	35.0
GeM-VGG16 [4]	82.5	82.2	55.5	26.6	63.0	37.2
GeM-VGG16-Whiten [4]	87.2	87.8	60.5	32.4	69.3	44.3
GeM-Res101 [4]	81.0	87.7	55.5	27.5	70.0	44.7
GeM-Res101-Whiten [4]	88.2	92.5	65.3	40.0	76.6	55.2

Paris appears to be close to the new Easy setup, so we report only the M and H results in our experiments.

Baseline. A CNN-based image retrieval model, GeM [4] is adopted as baseline in our experiment to implement our proposed OLMANS on. Two different CNN backbones, VGG16 [56] and ResNet101 [57], are evaluated. Besides, the whitening is adopted as a 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 [4] is utilized in our work.

5.2.2 Comparison with State-of-the-art

The comparison experiment results are shown in Table. 12. Compared with the baseline models, GeM-VGG16, by implementing our proposed OLMANS to them, the mAP performance of GeM-VGG16 is improved from (82.5%, 82.2%, 55.5%, 26.6%, 63.0%, 37.2%) to (83.5%, 82.9%, 55.9%, 26.8%, 63.5%, 37.3%) on (Oxford, Paris, $\mathcal{R}$Oxford-M, $\mathcal{R}$Oxford-H, $\mathcal{R}$Paris-M, $\mathcal{R}$Paris-H) respectively. The similar improvement is also observed for the GeM-VGG16-Whiten baseline. As for another more powerful baseline with a different backbone network, GeM-Res101, our method further boosts the mAP performance from (81.0%, 87.7%, 55.5%, 27.5%, 70.0%, 44.7%) to (81.7%, 87.6%, 56.1%, 27.8%, 70.3%, 44.9%) on (Oxford, Paris, $\mathcal{R}$Oxford-M, $\mathcal{R}$Oxford-H, $\mathcal{R}$Paris-M, $\mathcal{R}$Paris-H) respectively. The improvement of OLMANS can be further boosted by selecting NDB elaborately. While OLMANS still shows promising performance on image retrieval task under different baseline models, which verifies the generalization ability of proposed method.

6 Conclusions

In this paper, we proposed a novel online local metric adaptation algorithm to learn a dedicated Mahalanobis metric for each query probe on the online testing stage of visual instance retrieval (VIR). This new approach only uses negative samples for metric adaptation, which is practical in real situation. It largely reduces the demand for a large number of positive training data as in existing offline learning-based VIR methods, and it only incurs minimum computational costs to perform online learning. In-depth theoretical analyses well justify our algorithm and extensive experiments on different tasks demonstrate that our new approach consistently and significantly outperforms the state-of-the-arts. In this work, our proposed method is considered as a general and independent module for any offline metric learning or feature extraction baselines for further online local adaptation. In the future, it is interesting to extend our proposed approach into a deep metric learning approach since it could be directly involved in the model learning.

[4] https://github.com/filipradenovic/cnnimageretrieval-pytorch

References

[1] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[2] ——, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[3] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," in *CVPR*, 2018.

[4] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE T-PAMI*, 2018.

[5] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *ICCV*, 2015.

[6] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[7] W. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE T-PAMI*, 2013.

[8] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *ICCV*, 2017.

[9] Y. Zhou and L. Shao, "Aware attentive multi-view inference for vehicle re-identification," in *CVPR*, 2018.

[10] S. Bak and K. Carr, "One-shot metric learning for person re-identification," in *CVPR*, 2017.

[11] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016.

[12] M. F. T Ali and S. Chaudhuri, "Maximum margin metric learning over discriminative nullspace for person re-identification," in *ECCV*, 2018.

[13] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, 2018.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, 2017.

[15] 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," *arXiv*, 2016.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

[17] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *CVPR*, 2018.

[18] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *ECCV*, 2018.

[19] J. Zhou, P. Yu, W. Tang, and Y. Wu, "Efficient online local metric adaptation via negative samples for person re-identification," in *ICCV*, 2017.

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

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

[22] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person trasfer gan to bridge domain gap for person re-identification," in *CVPR*, 2018.

[23] G. Zhang, Y. Wang, J. Kato, T. Marutani, and K. Mase, "Local distance comparison for multiple-shot people re-identification," in *Asian Conference on Computer Vision*, 2013.

[24] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*, 2013.

[25] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013.

[26] V. E. Liong, J. Lu, and Y. Ge, "Regularized local metric learning for person re-identification," *PR Letters*, 2015.

[27] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in *CVPR*, 2016.

[28] L. Ma, X. Yang, and D. Tao, “Person re-identification over camera networks using multi-task distance metric learning,” *IEEE TIP*, 2014.

[29] S. Ding, L. Lin, G. Wang, and H. Chao, “Deep feature learning with relative distance comparison for person re-identification,” *PR*, 2015.

[30] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, “Embedding deep metric for person re-identification: A study against large variations,” in *ECCV*, 2016.

[31] E. Ustinova and V. Lempitsky, “Learning deep embeddings with histogram loss,” in *NIPS*, 2016.

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

[33] 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 *ECCV*, 2018.

[34] Y. Duan, W. Zheng, X. Lin, J. Lu, and J. Zhou, “Deep adversarial metric learning,” in *CVPR*, 2018.

[35] X. Lin, Y. Duan, Q. Dong, J. Lu, and J. Zhou, “Deep variational metric learning,” in *ECCV*, 2018.

[36] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” in *IEEE PETS Workshop*, 2007.

[37] L. Zheng, Y. Yang, and Q. Tian, “Sift meets cnn: A decade survey of instance retrieval,” *IEEE T-PAMI*, 2018.

[38] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, “Query specific fusion for image retrieval,” in *ECCV*, 2012.

[39] L. Zheng, S. Wang, Z. Liu, and Q. Tian, “Packing and padding: Coupled multi-index for accurate image retrieval,” in *CVPR*, 2014.

[40] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian, “Semantic-aware co-indexing for image retrieval,” in *ICCV*, 2013.

[41] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, “Query-adaptive late fusion for image search and person re-identification,” in *CVPR*, 2015.

[42] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, “Neural codes for image retrieval,” in *ECCV*, 2014.

[43] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, “Deep image retrieval: Learning global representations for image search,” in *ECCV*, 2016.

[44] M. Ye, J. Chen, Q. Leng, C. Liang, Z. Wang, and K. Sun, “Coupled-view based ranking optimization for person re-identification,” in *ICMM*, 2015.

[45] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel, “Person re-identification ranking optimisation by discriminant context information analysis,” in *ICCV*, 2015.

[46] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” 2017.

[47] A. Barman and S. K. Shah, “Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems,” in *ICCV*, 2017.

[48] E. Fetaya and S. Ullman, “Learning local invariant mahalanobis distances,” *ICML*, 2015.

[49] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, 2011.

[50] R. E. Schapire and Y. Freund, *Boosting: Foundations and algorithms*. MIT press, 2012.

[51] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE TIT*, 1967.

[52] P. L. Bartlett and S. Mendelson, “Rademacher and gaussian complexities: Risk bounds and structural results,” *JMLR*, 2002.

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

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

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012.

[56] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv*, 2014.

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

[58] Y. Sun, L. Zheng, W. Deng, and S. Wang, “Svdnet for pedestrian retrieval,” in *ICCV*, 2017.

[59] X. Chang, T. M. Hospedales, and T. Xiang, “Multi-level factorisation net for person re-identification,” in *CVPR*, 2018.

[60] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, “Part-aligned bilinear representations for person re-identification,” in *ECCV*, 2018.

[61] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, “Temporal model adaptation for person re-identification,” in *ECCV*, 2016.

[62] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in *CVPR*, 2012.

[63] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *ICML*, 2007.

[64] K. Q. Weinberger, J. Blitzer, and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” in *NIPS*, 2005.

[65] G. Lisanti, I. Masi, and A. Del Bimbo, “Matching people across camera views using kernel canonical correlation analysis,” in *Proceedings of the International Conference on Distributed Smart Cameras*, 2014.

[66] F. Xiong, M. Gou, O. Camps, and M. Sznaier, “Person re-identification using kernel-based metric learning methods,” in *ECCV*, 2014.

[67] 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 *CVPR*, 2015.

[68] Y. Chen, X. Zhu, and S. Gong, “Person re-identification by deep learning multi-scale representations,” in *ICCV*, 2017.

[69] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, “Dual attention matching network for context-aware feature sequence based person re-identification,” in *CVPR*, 2018.

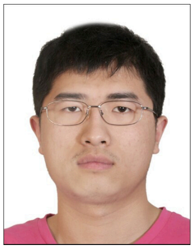
[70] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, “Human semantic parsing for person re-identification,” in *CVPR*, 2018.

[71] Z. Zheng, L. Zheng, and Y. Yang, “Pedestrian alignment network for large-scale person re-identification,” *TCSVT*, 2018.

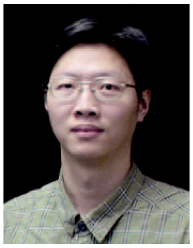
[72] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, “Pose-normalized image generation for person re-identification,” in *ECCV*, 2018.

[73] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, “Resource aware person re-identification across multiple resolutions,” in *CVPR*, 2018.

[74] E. Ristani and C. Tomasi, “Features for multi-target multi-camera tracking and re-identification,” in *CVPR*, 2018.



Jiahuan Zhou received his B.E. (2013) from Tsinghua University, the Ph.D. degree (2018) in the Department of Electrical Engineering & Computer Science, Northwestern University. Currently, he is a Postdoctoral Fellow in Northwestern University. His current research interests include computer vision, pattern recognition, visual identification and machine learning.



Ying Wu received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 1994, 1997, and 2001, respectively. In 2001, he joined the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA, as an Assistant Professor. He was promoted as an Associate Professor in 2007 and a Full Professor in 2012. His current research interests include computer vision, image and video analysis, pattern recognition and machine learning. He is a Fellow of the IEEE, for his “fundamental contributions to visual motion analysis and visual pattern discovery in computer vision”.