Abstract—Conventional multi-view re-ranking methods usually perform asymmetrical matching between the region of interest (ROI) in the query image and the whole target image for similarity computation. Due to the inconsistency in the visual appearance, this practice tends to degrade the retrieval accuracy particularly when the image ROI, which is usually interpreted as the image objectness, accounts for a smaller region in the image. Since Privileged Information (PI), which can be viewed as the image prior, is able to characterize well the image objectness, we are aiming at leveraging PI for further improving the performance of multi-view re-ranking in this paper. Towards this end, we propose a discriminative multi-view re-ranking approach in which both the original global image visual contents and the local auxiliary PI features are simultaneously integrated into a unified training framework for generating the latent subspaces with sufficient discriminating power. For the on-the-fly re-ranking, since the multi-view PI features are unavailable, we only project the original multi-view image representations onto the latent subspace, and thus the re-ranking can be achieved by computing and sorting the distances from the multi-view embeddings to the separating hyperplane. Extensive experimental evaluations on the two public benchmarks, Oxford5k and Paris6k, reveal that our approach provides further performance boost for accurate image re-ranking, whilst the comparative study demonstrates the advantage of our method against other multi-view re-ranking methods.

Index Terms—Multi-view re-ranking, privileged information (PI), latent subspaces, multi-view embeddings.

I. INTRODUCTION

RECENT years have witnessed massive efforts devoted to advancing the research over image re-ranking, which allows significant improvement the retrieval accuracy by refining the query model. Among all the re-ranking approaches, the subspace based strategy has become a promising line of research due to its desirable property in uncovering the discriminative latent subspace underlying the original high-dimensional feature space. In particular, multi-view re-ranking methods are capable of exploring the visual complementarity among heterogeneous feature spaces, which, therefore, leads to a latent representation with sufficient descriptive power. In order to further improve the separability of the query model, the classification mechanism is usually incorporated into subspace learning based re-ranking for producing a generic and discriminative framework [1], [2].

Despite their success in image re-ranking, conventional subspace based approaches directly leverage the visual features generated from the whole image for training the query model while ignoring the important role of image objectness in similarity matching. In many cases, actually, it is the region of interest (ROI) that characterizes the image objectness and captures the users’ query intention rather than the whole image containing complex background contents. In this sense, training the query model without considering the objectness tends to introduce the query-irrelevant noise, which leads to biased re-ranking results and thus adversely affects the retrieval performance. Therefore, it is crucial to incorporate the objectness into the trained query model for further improving the re-ranking accuracy.

It is well known that Privileged Information (PI) gives the supplementary cues about the training examples [3], [4]. Since PI is typically more informative about the task at hand than the raw data, it is usually combined with the original training examples for further improving the accuracy of the trained model. Recent research has clearly demonstrated the beneficial effect of PI learning in a wide range of vision tasks. Without loss of generality, PI can be defined in terms of four different modalities in the context of object classification, namely attributes, annotator rationales, bounding boxes and textual descriptions [4]. In particular, the PI in terms of bounding boxes can be viewed as an image prior, since it is capable of highlighting the object region and encoding the principal visual cues in the image. Besides, it is also available with easy-to-implement ROI annotation, which is
Fig. 1. The system flowchart of the proposed approach. In our method, two major steps are involved, namely model training and on-the-fly re-ranking. With the original training data and the annotated PI regions derived from user interaction, the former integrates the global multi-view embedding and its PI-based counterpart into a unified framework, which produces a PI-aware latent subspace with sufficient discriminating power. For on-the-fly re-ranking, the latent representations of the target images are obtained by projecting the multi-view features onto the PI-aware subspace, and thus the images are re-ranked by the signed distances from the separating hyperplane accordingly.

tailored for the user interaction in image re-ranking. Therefore, in this paper, we only focus on the PI formulated as the bounding box, whilst aiming to exploit both the original and the supplementary PI features to train the re-ranking model for accurate retrieval. More specifically, inspired by the unified subspace based re-ranking framework proposed in [1], we propose a discriminative PI-aware multi-view re-ranking method in which multi-view local PI features are also integrated into the query model training along with their global counterparts.

Fig. 1 gives the processing pipeline of the proposed method. Analogous to the DMINTIR re-ranking method in [1], our approach comprises two steps, namely query model training and on-the-fly re-ranking. In the model training, we first identify the query-relevant images from the top returned shortlist, and then annotate the PI regions in these positive examples with cropped ROI bounding boxes via user interaction. In addition, the lowly scored images in the original ranking list are automatically recognized as the negative training examples, and their corresponding PI regions can be obtained by the off-the-shelf saliency detector [5]. Thus, the training data consisting of both global contents and additional local PI regions can be handled in the original and the privileged spaces respectively. Then, we compute the multi-view features in both spaces and project them onto the PI-aware latent subspace for uncovering the underlying low-dimensional representations. Meanwhile, a PI-aware latent subspace with sufficient discriminating power can be obtained by jointly optimizing the separating hyperplanes of the dual subspaces. For the on-the-fly re-ranking, due to the unavailability of the PI in the target images, we directly project the multi-view features onto the PI-aware subspace for generating the discriminative representations, and thus the database images can be re-ranked by computing and sorting the distances from the separating hyperplane for performance improvement.
suggests the benefits of PI for accurate re-ranking. In re-ranking, the single-view feature often fails to provide a comprehensive visual description, and thus leads to an image signature with insufficient descriptive power. By contrast, multiple views allow us to take advantage of the complementarity among multiple heterogeneous features, which, consequently, substantially benefits the re-ranking performance. Earlier multi-view re-ranking approaches leverage the low-level features (e.g., bag of features, color histogram and wavelet textures) for characterizing the visual contents in the images [6]–[10]. Then, either a linear transformation [9] or complex hypergraph manifolds [6]–[8], [11] are learned from these multi-view features to uncover the intrinsic structure or a low-dimensional subspace for re-ranking. Besides, robust estimator has also been utilized for multi-view intact space learning [12].

Low-level features encode the visual patterns intuitively, yet fails to provide higher-level image representation. Recently, deep features have been used as a desirable alternative in multi-view learning, since they encode high-level semantic attributes in the image with preferable descriptive power [1], [13], [14]. An important line of research incorporates deep feature encoding and image re-ranking in a unified framework. A representative method consists in simulating the dynamics of heat diffusion to achieve the deep feature aggregation, while the re-ranking inspired from the diffusion leads to considerable performance gains [15]. In addition, the deep feature encoding can also be achieved via multi-view embedding for generating a latent subspace representation. Particularly, a discriminative multi-view re-ranking approach has been proposed in [1] to integrate the deep CNN code and the top performing hand-crafted feature TE into a generic and unified framework, to produce a latent low-dimensional subspace maintaining sufficient separability. Thus, multi-view features can be projected onto this subspace to generate robust latent representations for accurate re-ranking. Albeit effective, [1] directly exploits the global features for multi-view embedding while downplaying visual cues in the query region. As a result, it exhibits suboptimal performance when there exist complex background contents and severe geometric transformation of query object.

B. Learning Using Privileged Information

In the computer vision community, PI, which is interpreted as the auxiliary information about the training data, can be used for learning better recognition systems. Recently, extensive efforts are devoted to exploring PI cues for enhancing the model training in a variety of vision tasks [4], [16]–[24]. The earliest research over PI learning integrates PI into the classic SVM algorithm, which produces an extended paradigm termed Learning Using Privileged Information (LUPI) [3]. The resulting model is also referred to as SVM+. In [4], four different PI types are explored and handled in a unified LUPI framework in the context of object classification. Besides, a novel rank transfer approach comparable to the conventional SVM+ algorithm is also proposed for solving the LUPI task. While the extensions of SVM+ algorithm to multiclass problem are possible, PI is also incorporated into the framework of generalized matrix learning vector for prototype-based classification [25].

II. RELATED WORK

A. Multi-View Image Re-Ranking

In re-ranking, the single-view feature often fails to provide a comprehensive visual description, and thus leads...
metric learning [16]. Analogously, person re-identification is also addressed in [23] by joint distance metric learning with the help of PI. Besides, PI is also embedded into the deep Convolutional Neural Networks (CNNs) and Recurrent Neural Network (RNNs) for image classification and action recognition [19], [20], [24]. In addition to the aforementioned applications, human-generated captions are used as PI for learning the improved representation in semantic retrieval [21].

In order to further exploit the complementary information among multiple feature sets, a new multi-view privileged SVM model is proposed by incorporating the LUPI paradigm into multi-view learning framework, which satisfies both consensus and complementary principles for multi-view learning [22].

Although great progress has been made in PI learning, how to make use of PI cues in image re-ranking for further performance improvement remains an open problem. In this paper, we propose a generic PI-aware re-ranking framework in which the original global representations and the additional PI cues are simultaneously incorporated into subspace-based multi-view embedding. The resulting PI-aware subspace preserves sufficient discrimination in the image, and thus can be used for generating discriminative objectness-aware latent representation for accurate re-ranking. To the best of our knowledge, this is the first time the PI learning is explored in image re-ranking.

### III. THE PROBLEM FORMULATION

Given the initial ranking results $\mathcal{R}$ and query $\mathcal{Q}$, image re-ranking is formally defined as refining $\mathcal{R}$ by re-evaluating the relevance of target images to $\mathcal{Q}$ with a re-ranking model $\mathcal{M}$ and generating the polished image ranks $\hat{\mathcal{R}}$ accordingly: $(\mathcal{R}, \mathcal{Q}) \xrightarrow{\mathcal{M}} \hat{\mathcal{R}}$. In the classification-based image re-ranking, it is required to label partial images from scratch for training $\mathcal{M}$. To be specific, user relevance feedback is usually imposed on the top returned shortlist $\mathcal{R}_K$ generated from $\mathcal{R}$ to recognize the query-relevant images as the true positive $I^+$, while the low-scored images are automatically identified as the negative instances $I^-$ ($K$ denotes the shortlist size). Thus, the training set $\mathcal{S} = (S^+, S^-)$ is produced where $S^+ = \{i^+ | \forall i = 1, \ldots, |S^+|\}$ and $S^- = \{i^- | \forall j = 1, \ldots, |S^-|\}$ denote the positive and negative set respectively. Since PI learning is also involved in our approach, we encode the privileged data as the ROI bounding boxes in addition to the aforementioned binary labels.

Given $\mathcal{S} = (S^+, S^-)$ with annotated ROI bounding boxes obtained by user interaction, the corresponding multi-view features generated from both the whole image and ROI can be denoted as $\mathcal{Z} = \{Z_v\}_{v=1}^m$ in the original space and $\mathcal{Z}^* = \{Z_v^*\}_{v=1}^m$ in the privileged space, where $m$ is the number of views. Since both spaces share the same feature dimensionality as single-view data, we have $Z_v, Z_v^* \in \mathbb{R}^{D_v \times n}$, where $D_v$ is the feature dimensionality of the $v^{th}$ view while $n = |\mathcal{S}|$ is the size of the training set. Meanwhile, $\mathcal{Y} \in \{+1, -1\}^n \times 1$ is the label vector denoting the class label of the training examples. The PI-aware multi-view re-ranking model training is aiming at learning the dual mapping functions:

$$f : \{\mathcal{Z}, \mathcal{Y}\} \rightarrow \{X, w\}$$  \hspace{1cm} (1)

where $X$ and $X^*$ are the respective multi-view subspace embeddings, whilst $w$ and $w^*$ are the dual separating hyper-planes preserving sufficient discriminative power in both underlying subspaces. For the sake of consistency, we learn the function $h(w, w^*)$ such that the two subspaces are mutually interlinked and a PI-aware low-dimensional subspace can be produced.

For the on-the-fly re-ranking, since the PI data is unavailable, we directly project the multiple features of the target images $\{\tilde{Z}_v\}_{v=1}^m$ on the trained PI-aware subspace for generating the latent representations $\tilde{X}$. Thus, the refined image ranks $\hat{\mathcal{R}}$ can be obtained by computing the signed distance of $\tilde{X}$ from the decision boundary $w$ for accurate re-ranking. All the mathematical notations involved in our formulation are summarized in Table I.

### IV. DMVIPR: DISCRIMINATIVE MULTI-VIEW PI-AWARE RE-RANKING

In the state-of-the-art subspace-based multi-view embedding methods, it is assumed that the image feature of a single view $z_v \in \mathbb{R}^{D_v}$ can be recovered from a shared underlying subspace via a view-specific generation matrix $P_v \in \mathbb{R}^{D_v \times d}$ such that:

$$z_v = P_v \cdot x + \epsilon_v$$  \hspace{1cm} (3)

where $x \in \mathbb{R}^d$ is the low-dimensional subspace representation, whilst $\epsilon_v$ is the view-dependent mapping error. Thus, the latent subspace can be obtained by minimizing the following formulation:

$$J(P_v, x) = \frac{m}{\sum_{v=1}^m \|z_v - P_v \cdot x\|^2 + \lambda \sum_{v=1}^m \|P_v\|_F^2 + \beta \|x\|^2}$$  \hspace{1cm} (4)

where $\lambda$ and $\beta$ are the tradeoff parameters between the two regularization terms which are incorporated into the formulation for penalizing the generation matrices and the latent representations to avoid overfitting.

In our case, we impose the multi-view embedding on all the training examples in both the original and the privileged feature spaces, and thus we have the following formulations.

### TABLE I

**THE NOTATIONS INVOLVED IN OUR MATHEMATICAL FORMULATION**

| $n$ | the size of the training set |
|-----|-----------------------------|
| $m$ | the number of data view |
| $D_v$ | the view-specific feature dimensionality |
| $d$ | the dimension of the latent subspace |
| $z_v^{(i)} \in \mathbb{R}^{D_v}$ | the view-specific feature representation in the original space |
| $z_v^{* (i)} \in \mathbb{R}^{D_v}$ | the view-specific feature representation in the privileged space |
| $P_v \in \mathbb{R}^{D_v \times d}$ | the view-specific generation matrix in the original space |
| $P_v^* \in \mathbb{R}^{D_v \times d}$ | the view-specific generation matrix in the privileged space |
| $x_v \in \mathbb{R}^d$ | the sample-specific latent representation in the original space |
| $x_v^* \in \mathbb{R}^d$ | the sample-specific latent representation in the privileged space |
| $w \in \mathbb{R}^d$ | the separating hyperplane in the original space |
| $w^* \in \mathbb{R}^d$ | the separating hyperplane in the privileged space |
| $\epsilon_v \in \{1, -1\}$ | the sample label |
to minimize:

\[
J(P_v, x_i) = \sum_{i=1}^{n} \sum_{v=1}^{m} \|z_v^{(i)} - P_v \cdot x_i\|^2 + \lambda \sum_{v=1}^{m} \|P_v\|_{F}^2 + \beta \sum_{i=1}^{n} \|x_i\|^2
\]

(5)

\[
J(P_v^*, x_i^*) = \sum_{i=1}^{n} \sum_{v=1}^{m} \|z_v^{(i)} - P_v^* \cdot x_i^*\|^2 + \lambda \sum_{v=1}^{m} \|P_v^*\|_{F}^2 + \beta \sum_{i=1}^{n} \|x_i^*\|^2
\]

(6)

To ensure the model discrimination capability, learning separating hyperplane \(w\) and \(w^*\) in the dual subspaces should also be also encoded in the formulation to distinguish between query-relevant and irrelevant examples. Besides, \(w^*\) learning should play a dominant and leading role, since the privileged features are more informative and confident in discriminatively separating the examples. As a result, simultaneous learning of \(w\) and \(w^*\) can be formulated as:

\[
\min_{w, w^*, b, b^*} \frac{1}{2} \left( \|w\|^2 + \gamma \|w^*\|^2 \right) + C \sum_{i=1}^{n} w^T x_i^* + b^* \\
\text{s.t. } y_i(w^T x_i + b) \geq 1 - (w^T x_i^* + b^*), \\
w^T x_i^* + b^* \geq 0, \quad \forall i = 1, \ldots, n
\]

(7)

It is well-known that the sample-specific slack variable \(\xi_i\) in the standard SVM formulation encodes the degree of confidence with which the sample is correctly classified and facilitates the SVM training with desirable generalization capability. In SVM+ formulated in (7), \(\xi_i\) is replaced by \(w^T x_i^* + b^*\) which can be interpreted as the confidence of correctly classifying the sample in the privileged space. Thus, the privileged data about the learning problem helps to guide the training of a better classifier and produce a more accurate predictor when the privileged data is unavailable at test time.

To summarize, the goal of SVM+ aims to utilize the privileged data \(\{x_i^*\}_{i=1}^{n}\) along with the original data \(\{x_i\}_{i=1}^{n}\) to learn a better separating hyperplane than with only original data involved. Thus, we have the mathematical formulation of our Discriminative Multi-View PI aware Re-ranking (DMVPIR) model by integrating (5), (6) and (7) into a unified framework as follows:

\[
\mathcal{L}(x_i, x_i^*, P_v, P_v^*, w, w^*) = \min J(P_v, x_i) + J(P_v^*, x_i^*) \\
+ \frac{1}{2} \left( \|w\|^2 + \gamma \|w^*\|^2 \right) + C \sum_{i=1}^{n} w^T x_i^* + b^* \\
\text{s.t. } y_i(w^T x_i + b) \geq 1 - (w^T x_i^* + b^*), \\
w^T x_i^* + b^* \geq 0, \quad \forall i = 1, \ldots, n
\]

(8)

where

\[
J(P_v, x_i) = \sum_{i=1}^{n} \sum_{v=1}^{m} \|z_v^{(i)} - P_v \cdot x_i\|^2 + \lambda \sum_{v=1}^{m} \|P_v\|_{F}^2 + \beta \sum_{i=1}^{n} \|x_i\|^2
\]

(9)

\[
J(P_v^*, x_i^*) = \sum_{i=1}^{n} \sum_{v=1}^{m} \|z_v^{(i)} - P_v^* \cdot x_i^*\|^2 + \lambda \sum_{v=1}^{m} \|P_v^*\|_{F}^2 + \beta \sum_{i=1}^{n} \|x_i^*\|^2
\]

(10)

As shown in Eq. (8), our DMVPIR re-ranking model aims to learn a PI-aware subspace with sufficient discriminative power encoded by decision boundary \(w\). For the on-the-fly re-ranking, we project the multi-view feature representations of the target images \(\{\tilde{Z}_i\}_{i=1}^{m}\) onto the PI-aware latent subspace via the optimal learned view-dependent generation matrix \(\{\tilde{P}_v\}_{v=1}^{V}\), which results in the low-dimensional subspace representations \(\tilde{X}\) for the subsequent similarity measure and re-ranking. Mathematically, \(\tilde{X}\) can be obtained by solved from the following minimization problem:

\[
\min \mathcal{L}(\tilde{X}) = \min \sum_{i=1}^{m} \|\tilde{Z}_i - \tilde{P}_v \cdot \tilde{X}\|^2_2 + \beta \|\tilde{X}\|^2_2
\]

(11)

After giving the DMVPIR formulation, we employ a finite system to illustrate the potential advantages of the proposed framework. Our proposed multi-view learning algorithm attempts to learn a finite set \(\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_m\) of hypotheses \((f_1, \ldots, f_m) : (Z^1, \ldots, Z^m) \rightarrow \mathcal{X}\) projecting multi-view examples for a unified representation. Suppose classification as the subsequent task, we need to learn a finite set \(\mathcal{H}\) of hypotheses \(h : \mathcal{X} \rightarrow \{0, 1\}\). We define the overall process as \(h \circ f\), and let \(\text{err}(h \circ f)\) be the expected error and \(\mathbb{E}\text{err}(h \circ f)\) be the empirical error on a training set of \(n\) i.i.d. multi-view examples. Combining Hoeffding’s inequality with a union bound, we obtain that with probability at least \(1 - \delta\), for every \(h \in \mathcal{H}\) and \((f_1, \ldots, f_m) \in \mathcal{F}\),

\[
\text{err}(h \circ f) \leq \mathbb{E}\text{err}(h \circ f) + \mathcal{O}\left(\frac{1}{\sqrt{n}} \sqrt{\ln |\mathcal{H}| + \ln |\mathcal{F}| + \ln \frac{1}{\delta}}\right)
\]

(12)

On the other hand, if we independently treat each view, and learn a finite set \(\mathcal{F}' = F_1' \times \cdots \times F_m'\) of hypotheses \((f_1', \ldots, f_m') : (Z^1, \ldots, Z^m) \rightarrow (\mathcal{X}^1, \ldots, \mathcal{X}^m)\) projecting each view into a different subspace. Since the embedded multiple views do not share the same representations ever, we have to learn different classifiers for the embeddings on different views, that is \((h_1', \ldots, h_m') \in (\mathcal{H}'^m)\), where \(h_v : \mathcal{X}^v \rightarrow \{0, 1\}\). Again Hoeffding’s inequality and a union bound imply that with probability at least \(1 - \delta\), we have for all \((h_1', \ldots, h_m') \in (\mathcal{H}'^m)\) and \((f_1', \ldots, f_m') \in \mathcal{F}'\)

\[
\text{err}(h' \circ f') \leq \mathbb{E}\text{err}(h' \circ f') \\
+ \mathcal{O}\left(\frac{1}{\sqrt{n}} \sqrt{m \ln |\mathcal{H}'^m| + \ln |\mathcal{F}'| + \ln \frac{1}{\delta}}\right)
\]

(13)

Since both of the first terms in Eqs. (12) and (13) involve the training error, we cannot directly compare them. By analyzing their second terms, we suggest that Eq. (12) could imply a tighter bound than that of Eq. (13) in practice for two reasons:

a) Given \(|\mathcal{H}'| \approx |\mathcal{H}|\), if we independently learn embedding for each individual view, we have to learn multiple classifiers on different views, and thus the overall cardinality would
increase with the number of views. Most importantly, with the help of privileged information (see constraints in Eq. (8)), the complexity of space $\mathcal{H}$ could be reduced further, compared with $\mathcal{H}'$ that has no investigation of privileged information. b) Though both $\mathcal{F}$ and $\mathcal{F}'$ can be decomposed as the operation on each view, the cardinality of $\mathcal{F}$ should be smaller than that of $\mathcal{F}'$, due to the constraints between the operations on different views.

Different from DMINTIR [1] and SVM+ algorithms [3], [4], we integrate them into a unified framework in which both multi-view embedding and PI learning are simultaneously achieved for accurate image re-ranking. The novelty of our method consists in mainly exploring how to address the multi-view re-ranking problem with auxiliary privileged data at hand for the first time. We argue the problem we aim to address in this paper is significant in multi-view re-ranking, since it is critical to explore how to make full use of the additional training data available for further improving the re-ranking performance. Besides, we leverage PI learning for conventional multi-view re-ranking, yielding a unified framework which allows fast optimization for efficient model training.

V. OPTIMIZATION

To solve the problem in Eq. (8), we develop an efficient iterative alternating optimization algorithm in which the following five alternating optimization steps iteratively minimize the empirical loss.

**First, we update $x_i$ by fixing the other parameters,** and thus the problem is reduced to the following formulation:

$$\min \mathcal{L} = \min_{x_i} \frac{1}{2} \sum_{v=1}^{m} \|z_d^{(v)} - P_v x_i\|^2 + \beta \|x_i\|^2$$

s.t. $y_i w^T x_i \geq c$ (14)

where $c = 1 - (y_i^T w^* + b^*) - y_i b$.

Furthermore, the objective function in Eq. (14) can be simplified as:

$$\min \mathcal{L} = \min_{x} \frac{1}{2} \sum_{v=1}^{m} \|z_d - P_v x\|^2 + \beta \|x\|^2$$

$$= \min_{x} \frac{1}{2} \sum_{v=1}^{m} (x^T P_v^T P_v x - 2 z_d^T P_v x) + \beta x^T x$$

$$= \min_{x} \frac{1}{2} \sum_{v=1}^{m} (x^T P_v^T P_v x - 2 c_d^T P_v x) + \beta x^T x$$

$$= \min_{x} \frac{1}{2} \sum_{v=1}^{m} (x^T P_v^T P_v x - 2 c_d^T P_v x) + \beta x^T x$$

Thus, the problem is formulated as:

$$\min \frac{1}{2} x^T (2 \sum_{v=1}^{m} P_v^T P_v + 2 \beta \mathbf{I}) x + ( - \sum_{v=1}^{m} z_d^T P_v x) x$$

s.t. $y_i w^T x \geq c$ (16)

where $c = 1 - (x_i^T w^* + b^*) - y_i b$.

Note that Eq. (16) is the classic quadratic programming (QP) problem:

$$\min \mathcal{L} = \min_{x} \frac{1}{2} \sum_{v=1}^{m} \|z_d^{(v)} - P_v x_i\|^2 + \beta \|x_i\|^2 + C w^T x$$

s.t. $x_i^T w \geq c$

$$x_i^T w \geq -b$$ (19)

where $c = 1 - y_i (w^T x_i + b) - b^*$. For the sake of simplicity, the objective function in Eq. (19) can be expressed as:

Thus, the problem can be formulated as:

$$\min \mathcal{L} = \min_{x} \frac{1}{2} \sum_{v=1}^{m} \|z_d^{(v)} - P_v x_i\|^2 + \beta \|x_i\|^2 + C w^T x$$

s.t. $x^T w \geq c$, $w^T x \geq -b$ (21)

Applying the problem in Eq. (21) can be also interpreted as a QP problem formulated as

$$\min \frac{1}{2} x^T U x + V^T x$$

s.t. $Ax \leq g$ (22)
where:

\[ U = 2 \sum_{v=1}^{m} P_v^T P_v^* + 2b^* I, \quad V = C w^* - \sum_{v=1}^{m} 2 P_v^T z_v^* \]

\[ A = -w^T, \quad g = -\max(c, -b^*) \]  

(23)

Analogously, the problem in Eq. (22) can also be solved by an off-the-shelf QP solver.

**Third, we update \( P_v \) by fixing the other parameters**, and thus the problem is reduced to the following formulation:

\[ \min_{P_v} \mathcal{L} = \min_{P_v} \sum_{v=1}^{m} \|Z_v - P_v X\|^2_F + \lambda \|P_v\|^2_F \]  

(24)

Eq. (24) is an unconstrained ridge regression optimization, which could be transformed into:

\[ \min_{P_v} \mathcal{L} = \min_{P_v} \sum_{v=1}^{m} \|Z_v - P_v X\|^2_F + \lambda \|P_v\|^2_F \]

\[ = \min_{P_v} \sum_{v=1}^{m} -P_v tr \left( (Z_v - P_v X)^T Z_v - P_v X \right) + \lambda tr(P_v^T P_v) \]

\[ = \min_{P_v} -tr \left( -X^T P_v^T Z_v - Z_v^T P_v X + \lambda X^T P_v^T P_v X \right) \]

\[ + \lambda tr(P_v^T P_v) \]  

(25)

Thus, we take the derivatives of \( \mathcal{L} \) w.r.t. \( P_v \) and have:

\[ \nabla_{P_v} \mathcal{L} = -X^T P_v^T Z_v - \nabla_{P_v} tr Z_v^T P_v X \]

\[ + \nabla_{P_v} tr X^T P_v^T P_v X + \lambda \nabla_{P_v} tr(P_v^T P_v) \]

\[ = -(X Z_v^T)^T - (X Z_v^T)^T + 2 P_v X X^T + 2\lambda P_v \]

\[ = -2 Z_v X^T + 2 P_v X X^T + 2\lambda P_v \]

\[ = 0 \]  

(26)

Therefore, we obtain the close-form of \( P_v \) as follows:

\[ P_v = Z_v X^T (X X^T + \lambda I)^{-1} \]  

(27)

**Next, we update \( P_v^* \) by fixing the other parameters**, and thus the problem is reduced to the following formulation:

\[ \min_{P_v^*} \mathcal{L} = \min_{P_v^*} \sum_{v=1}^{m} \|Z_v^* - P_v^* X^*\|^2_F + \lambda^* \|P_v^*\|^2_F \]  

(28)

Resembling solving for \( P_v \), we derive the close-form solution of \( P_v^* \) as follows:

\[ P_v^* = Z_v^* X^* (X^* X^*)^{-1} + \lambda^* I \]  

(29)

Finally, we update \( w, w^*, b, b^* \) by fixing the other parameters, and thus the problem is reduced to solving for a classic SVM+ problem:

\[ \min_{w, w^*, b, b^*} \mathcal{L} = \min_{w, w^*, b, b^*} \frac{1}{2} (\|w\|^2 + \gamma \|w^*\|^2) + C \sum_{i=1}^{n} \max(0, 1 - y_i (w^T x_i + b^*)) \]  

s.t. \( y_i (w^T x_i + b) \geq 1 - (w^T x_i^* + b^*) \), \( w^T x_i^* + b^* \geq 0, \quad \forall i = 1, \ldots, n \)  

(30)

which can be solved by a fast algorithm in [26].

---

**Algorithm 1: Summary of Our Optimization Procedure**

**Input:** \( \{z_v^{(i)}\}_{v=1}^{m}, \{z_v^{(i)}\}_{v=1}^{m}, y_1, \lambda, \lambda^*, \beta, \beta^*, \gamma, C, \) \( \forall i = 1, \ldots, n \)

**Output:** \( \{z_i\}_{i=1}^{n}, \{z_i\}_{i=1}^{n}, \{P_v\}_{v=1}^{m}, \{P_v^*\}_{v=1}^{m}, w, w^*, b, b^* \)

1. **Initialize:** \( \{z_i\}_{i=1}^{n}, \{P_v\}_{v=1}^{m}, \{P_v^*\}_{v=1}^{m}, w, w^*, b, b^* \)

2. **Repeat**

3. **\( z_i \) update** through solving Eq. (17) by QP algorithm

4. **\( z_i^* \) update** through solving Eq. (22) by QP algorithm

5. **\( \{P_v^*\}_{v=1}^{m} \) update** by Eq. (27)

6. **\( \{P_v\}_{v=1}^{m} \) update** by Eq. (29)

7. **\( w, w^*, b, b^* \) update** by SVM+ algorithm [26]

8. **Until Convergence**

---

**VI. Analysis of the Computational Complexity**

We now discuss the computational complexity of our DMVIPR algorithm for separate phases.

In the model training, the overall computational overhead consists of three main parts, i.e., solving for \( x \) and \( x^* \) in Eqs. (17) and (22), computing \( P_v \) and \( P_v^* \) in Eqs. (27) and (29) as well as updating \( w, w^* \), \( b, b^* \) with SVM+ algorithm. Since both \( x \) and \( x^* \) are estimated by an off-the-shelf QP solver in practice, the corresponding time complexity can be computed in \( O(d^3) \), and thus the total cost for the whole training set in dual spaces amounts to \( 2n \cdot O(d^3) \). In Eq. (25) and (27), computing \( P_v \) and \( P_v^* \) requires \( O(D_v \cdot n \cdot d) + O(d^2 \cdot n) + O(d^3) + O(D_v \cdot d^2) \) time. It can be approximated by
\[ O(D_o \cdot n \cdot d) + O(D_o \cdot d^2) \], since \( D_o \gg d \) in our case. Thus, updating all the view-specific generation matrices in dual spaces takes \( 2 \sum_{i=1}^{m} O(D_o \cdot n \cdot d) + O(D_o \cdot d^2) \). As for the \( \theta, \theta^*, b, b^* \) update, we directly use the fast linear SVM+ algorithm implemented in [26], and the time complexity is roughly \( O(2d) + O(d) \) [26]. Therefore, the total cost amounts to \( 2n \cdot O(d^3) + 2 \sum_{i=1}^{m} O(D_o \cdot n \cdot d) + O(D_o \cdot d^2) + O(2d) + O(d) \), which is thus reduced to \( 2n \cdot O(d^3) + 2 \sum_{i=1}^{m} O(D_o \cdot n \cdot d) + O(D_o \cdot d^2) \) approximately.

During the re-ranking stage, the computational cost comprises the multi-view embedding for generating the latent representations shown in Eq. (30) and the subsequent cosine similarity. The former is calculated in \( \sum_{i=1}^{m} O(d^2 \cdot D_o) + O(d^3) + 2 \sum_{i=1}^{m} O(d \cdot D_o \cdot n) + O(d^2 \cdot n) \) which can be approximated by \( \sum_{i=1}^{m} O(d^2 \cdot D_o) + O(d \cdot D_o \cdot n) \), while the latter has a time complexity of \( O(n \cdot d) \) for efficient similarity measure.

VII. Experiments

In this section, we will evaluate our DMVPIR method for image re-ranking. First, we will introduce the public benchmark datasets as well as the experimental setup and the performance measure. Subsequently, thorough qualitative and quantitative evaluations will be carried out to demonstrate the performance of our approach. Besides, we also conduct a comparative study for showing the superiority of our method to the state-of-the-arts.

A. Benchmark Datasets and Performance Measures

We evaluate our DMVPIR re-ranking approach on two public datasets, Oxford5k [27] and Paris6k [28], both of which are usually used as benchmarks for instance-level image retrieval. The two datasets include 5,063 and 6,392 images of 11 famous landmarks in Oxford and Paris respectively, and each landmark is represented by five query instances. This results in a total of 55 query groups for querying the whole dataset. All the images in the dataset fall into four groups according to the query-specific relevance. Besides, Flickr100k dataset [27] consisting of 100,071 images is also involved in our experiments to evaluate the performance of our proposed method in the large-scale scenario. Average Precision (AP) score is computed to evaluate single query experiments, and mean average precision (mAP) obtained by averaging all the AP scores is used as the overall performance measure. Besides, we also adopt the Normalized Discounted Cumulative Gain (NDCG) for evaluation [29]. The NDCG score at position \( P \) for a specific query can be computed as:

\[
\text{NDCG}@P = Z_P \sum_{i=1}^{P} \frac{2^{l(i)} - 1}{\log(i + 1)}
\]

where \( P \) is the ranking depth, \( l(i) \) denotes the relevance of the \( i^{th} \) ranked image to the specific query, and \( Z_P \) is the normalization constant that makes the optimal NDCG@P equal 1. Similar to mAP, mean NDCG (mNDCG) score is also used to evaluate the overall performance.

B. Multi-View Features

Following [1], we leverage three complementary image signatures for multi-view feature representations in our approach, namely CNN, TE and VLAD+. CNN feature is a 4,096-dimensional vector which consists in the activations of the upper layer of the deep neural network. In addition to the pre-trained VGG-16 architecture [30], we also leverage a fine-tuned deep model trained completely on Pittsburgh [31] for generating our CNN codes, since Pittsburgh is a dataset which consists of the street-view images resembling those used in our experiments. In implementation, we use the MatConvNet toolbox [32] for VGG-16, while utilize the fine-tuned model publicly available in [33]. Known as the best shallow image signature thus far, TE referred to as triangulation embedding is viewed as a promising alternative to FV vector [34], whilst VLAD+ developed from Root-SIFT descriptor is more computationally efficient for fast retrieval [35]. In implementation, we use the same vocabulary sizes for TE and VLAD+ as in [1], which lead to 8,064 and 16,384-dimensional vectors for respective representations. The complementarity among the three heterogeneous features can be fully exploited for multiple feature embedding, since deep CNN feature enables high-level image description, whilst TE and VLAD+ inherit desirable invariant property from robust local descriptors.

C. Interactive Relevance Feedback With PI Annotation

Analogous to [1], given the ranking images obtained in the first place, we utilize the user relevance feedback (URF) performed once for assembling the positive query-relevant images while automatically recognizing the lowly scored examples as the negative distractors for training our re-ranking model. Different from the conventional URF methods [1], [36], [37], however, both a click indicating the query-relevance of an image and the object ROI capturing the user query are required for obtaining the auxiliary PI data our scenario. To be specific, we annotate the image ROIs in the positive examples while adopting the off-the-shelf saliency detector [5] for generating the PI regions in the negative images. Thus, the original set of training images alongside the corresponding supplementary PI data are delivered to the subsequent module for extracting multi-view features. Since the user interaction with PI annotation is performed on the shortlisted images relatively accounting for a small proportion of the top returned results, this practice incurs affordable overhead on the system.

D. The Labeled Images in the Training Process

With the help of the above-mentioned user interaction, one image in the shortlist is manually labeled as the positive (+1) through one user click indicating it is query-relevant, whilst the negative labels (−1) are automatically generated from the low-ranked images in the initial retrieval results. Thus, the binary training labels are produced. Since each query landmark includes five query instances in our experiments, we calculate the average number of manually labeled images for each query landmark on both datasets as illustrated.
TABLE II

| Query           | # of manual labels | Query           | # of manual labels |
|-----------------|-------------------|-----------------|-------------------|
| all_souls       | 29.4              | defense         | 27.8              |
| ashmolean       | 12.8              | eiffel          | 33                |
| balliol         | 8.2               | invalides       | 38.2              |
| bodleian        | 10.6              | louvre          | 29.8              |
| christ_church   | 26.2              | moulinrouge     | 29.4              |
| cornmarket      | 6                 | museedorsay     | 24.2              |
| hertford        | 29                | notredame       | 38.8              |
| keble           | 6                 | pantheon        | 38.8              |
| magdalen        | 7.4               | pompidou        | 34                |
| pitt_rivers     | 4.8               | sacrecoeur      | 36.4              |
| radcliffe_camera| 38.6              | triomphe        | 38.6              |
| Mean            | 16.3              | Mean            | 33.5              |

TABLE III

| Performance Measure | Paris6k | Oxford5k |
|---------------------|---------|----------|
| mAP                 | 61.76   | 62.04    |
| mNDCG@50            | 87.70   | 82.07    |
| mNDCG@100           | 77.96   | 70.16    |

The Statistics of the Manually Labeled Images on Average for Different Query Landmarks on Both Datasets

in Table II. It is shown that the number of manually labeled images varies from 5 to 39 approximately, leading to an average of 16 and 34 positive instances per query group on both datasets. With the size of negative set empirically set as 100, a total of approximately 116 and 134 instances are respectively used for the optimization per query.

E. Model Selection

In DMVPIR, six hyperparameters in Eq. (8) need to be carefully tuned, i.e., \( \lambda, \lambda^*, \beta, \beta^*, \gamma, C \). To this end, we perform model selection on a single query, and the optimal parameters obtained accordingly are used for evaluating the other query groups on the two benchmark datasets. In implementation, we select the query “all_souls” for model training with varying parameters.

F. Experimental Results

1) Comparison of Baseline Methods: In our baseline retrieval systems, a global image signature is combined with cosine similarity for generating a set of ranking images in the first place. In our case, we evaluate three image representations introduced in section VII-B, which leads to different baseline methods respectively denoted as TE_cos, CNN_cos and VLAD+_cos. Table III gives the performance of different baselines. It is clearly shown in Table III that TE_cos consistently outperforms the other two approaches by achieving the highest mAP at 61.76% and 62.04% on the respective datasets as well as higher mNDCG scores. Surprisingly, CNN_cos exhibits performance inferior to TE_cos, which can be attributed to the pre-trained deep model with insufficient descriptive power. Table IV gives the performance of baseline CNN_cos using the fine-tuned model. It can be observed that fine-tuning brings tremendous performance boosts by reporting the significantly increased mAP scores at 75.96% and 68.09% on both datasets. Although fine-tuning allows further improving the retrieval performance of CNN_cos, we still use the TE_cos as the baseline for the subsequent re-ranking, since in our work we only focus on the image re-ranking which operates independently of the baseline method.

2) The Performance of Our DMVPIR Method: We impose our DMVPIR method on the baseline TE_cos for accurate re-ranking. Fig. 3 presents the comparison of the baseline and our re-ranking approach in terms of the AP score. It is observed that DMVPIR provides significant performance gains ranging from 1.9% on “invalides” to 56% on “bodleian” for different query groups. In particular, DMVPIR reports respective mAP scores at 81.51% and 77.83% on two datasets and outperforms the baseline system by approximately 20% and 16%. This strongly suggests the beneficial effect of the proposed re-ranking approach. The only exceptions come from the queries “notredame” and “sacrecoeur” when slight performance drop occurs. This implies the generalization capability of DMVPIR is somewhat prone to the high nonlinearity of our model and the redundancy occasionally present in the training examples.

In addition, we compare the baseline and DMVPIR methods by computing NDCG scores in Table V and VI. As shown in the two tables, DMVPIR dramatically boosts the baseline results from 70.27% to 79.30% on Oxford5k while the performance gain also reaches 7% on Paris6k in terms of mNDCG@50. Similar trend can also be observed for mNDCG@100 score on both datasets, demonstrating that our re-ranking method considerably benefits the performance.

3) The Comparative Studies: In comparative studies, we compare our approach DMVPIR with other multi-view re-ranking methods as follows:

1) DMINTIR. We directly reproduce the algorithm in [1] with the analogous parameter setting adopted in our method.
2) DMINTIR-PI. For this approach, we leverage the local multi-view PI features for learning the separating hyperplane \( w \) without taking into account the original multiple...
global feature representations. The online re-ranking is achieved by computing and sorting the distances from the global multi-view projections of the target images to the hyperplane $w$.

3) DQE by Concatenating Averaged Reduced-size Multi-View features for Re-ranking (DQE-CAR-MVR). We first impose PCA on the multi-view features for dimension reduction in both original and privileged space. Thus, we fuse the compressed view-specific features in the two spaces by average pooling and concatenate the pooled features of different views for the holistic representation. Subsequently, analogous to [2], we train a linear SVM model on the resulting representation and compute the signed distance from the separating hyperplane for re-ranking. Note that the reduced feature dimensionality in this method is also set to be 128, which is consistent with the setting in our approach.

4) DQE by Concatenating Averaged Full-size Multi-View features for Re-ranking (DQE-CAF-MVR). This method is essentially the same with DQE-CAR-MVR except that the original dimensionalities of the multi-view features are maintained without dimension reduction.

5) DQE by Averaging Reduced-size Multi-view features for Re-ranking (DQE-AR-MVR). Different from DQE-CAR-MVR and DQE-CAF-MVR, this approach directly utilizes average pooling for fusing all the multi-view features with reduced size in both spaces, which leads to the final image representation to drive the linear SVM model. The reduced feature size is also set to be 128 for the sake of consistency.

6) Late Fusion on DQE with Averaged Reduced-size Multi-view features for Re-ranking (LFDQE-AR-MVR). This method is essentially the same with DQE-CAR-MVR except that the original dimensionalities of the multi-view features are maintained without dimension reduction for respective DQE model training.

To sum up, both DMINTIR and DMINTIR-PI simply take into account the visual information in a single space, whilst our approach along with the other competing methods combine the visual contents from both spaces. In particular, DQE-CAR-MVR, DQE-CAF-MVR as well as DQE-AR-MVR can be viewed as early fusion multi-view re-ranking strategies, whilst LFDQE-AR-MVR and LFDQE-AF-MVR fall into the category of late fusion techniques. These competing algorithms are typical classification-based multi-view re-ranking approaches closely correlated with our proposed method.

Table VII and VIII present the performance of different multi-view re-ranking methods on the two benchmarks. As revealed in the two tables, our scheme demonstrates the unrivalled performance superior to the other competing approaches overall. In particular, the proposed method performs better than both DMINTIR and DMINTIR-PI, which implies the considerable benefit in combining the original visual clues with supplementary PI data for re-ranking. More specifically, DMVPIR reports higher mAP scores surpassing DMINTIR by 1.2% and 0.7% respectively on two datasets. Since there exists the asymmetry between the training and the testing information in DMINTIR-PI, DMVPIR exhibits more dramatic performance advantage against DMINTIR-PI with significant improvements over 15%. In addition, our method also beats the other fusion-based re-ranking methods by achieving substantial performance gains. This sufficiently suggests our subspace-based scheme allows learning the discriminative representation from heterogeneous multi-view features while working better than the methods which perform straightforward fusion strategies.

As shown in Table VIII, the advantage of the proposed method against DMINTIR is not significant on Paris6k. This implies the re-ranking model resulting from our method is more prone to overfitting with higher model complexity than DMINTIR. Besides, since DMINTIR demonstrates promising results on Paris6k with the help of almost twice positive training images (34/16) per query than that on Oxford5k, PI learning demonstrates limited beneficial effect on the performance.
improvement. Note that our scheme does not achieve the best mNDCG results on Oxford5k shown in Table VII. We argue that this results from the evaluation mechanism of NDCG where the junk images with certain ambiguity are also taken into consideration in computing the query-relevance, whereas they are discarded in evaluating mAP score. In this sense, our scheme returns clear groundtruth images at higher ranks than those ambiguous examples.

In our experiments, we conduct more experimental evaluations to further demonstrate the superiority of the proposed method. More specifically, we only use two-view data, namely TE and CNN features, instead of the aforementioned setting in which all the three features are utilized. Table IX shows the comparison of our method and DMINTIR on Paris6k. It can be observed that our proposed method beats DMINTIR across all the three evaluation metrics. More specifically, the advantage varies on different query groups ranging from 0.34% to 2.67% in terms of the AP score, whilst our method consistently beats DMINTIR by returning more top ranked query-related images.

![Fig. 4. Comparison of qualitative retrieval results achieved by baseline (the first row), DMINTIR [1] (the second row) and our scheme (the last row).](image)

Given an image with the annotated query region outlined by the red dashed box shown on the left, the top returned results are displayed accordingly on the right. Note that the junk images and the false alarms are highlighted in green and red boxes, respectively. It is observed that our approach significantly improves the baseline result and exhibits performance advantage against DMINTIR by returning more top ranked query-related images.

![Table VII](image)

**Table VII**

| Methods     | mAP   | mNDCG@50 | mNDCG@100 |
|-------------|-------|----------|-----------|
| DMINTIR    | 80.34 | 82.82    | 81.13     |
| DMINTIR-PI | 61.56 | 66.83    | 66.72     |
| DQE-CAR-MVR | 77.39 | 79.23    | 77.91     |
| DQE-CAP-MVR | 78.72 | 79.72    | 78.01     |
| DQE-AR-MVR  | 40.37 | 49.47    | 49.36     |
| LFQDQE-AR-MVR | 74.42 | 77.43    | 75.80     |
| LFQDQE-AF-MVR | 79.48 | 80.89    | 78.89     |
| Proposed   | **81.51** | 79.30    | 78.18     |

![Table VIII](image)

**Table VIII**

| Methods     | mAP   | mNDCG@50 | mNDCG@100 |
|-------------|-------|----------|-----------|
| DMINTIR    | 77.09 | 94.64    | 87.28     |
| DMINTIR-PI | 61.91 | 85.88    | 76.22     |
| DQE-CAR-MVR | 72.36 | 92.22    | 83.85     |
| DQE-CAP-MVR | 74.90 | 93.85    | 85.13     |
| DQE-AR-MVR  | 46.89 | 63.70    | 58.04     |
| LFQDQE-AR-MVR | 64.54 | 83.00    | 75.69     |
| LFQDQE-AF-MVR | 74.57 | 93.80    | 85.23     |
| Proposed   | **77.83** | 94.76    | **87.74** |

Note that we use the pre-trained CNN features in the aforementioned experiments. In order to explore the impact of fine-tuning on the performance of our method, we deliver the fine-tuned CNN feature to the proposed DMVPIR in our evaluations, while follow the above-mentioned setting to use two-view features in practice. Table X reveals the performance of DMVPIR using pre-trained and fine-tuned CNN codes on Oxford5k. It clearly demonstrates fine-tuning is considerably conducive to improving the performance of our method. Particularly, fine-tuning brings enormous performance boosts exceeding 14% on “all_souls” and “magdalen” in terms of AP, while improves by 7% approximately on “christ_church” and “hertford”, reporting a significantly higher mAP score at **87.65%** against 82.12% resulting from DMVPIR using pre-trained model. Similar performance gains can also be observed on Paris6k shown in Table XI.

Besides, we also compare the proposed DMVPIR method with the state-of-the-arts in recent years in Table XII. The algorithms shown in Table XII have been selected in our comparative studies, since they are similar to our proposed method in a way that makes use of local visual cues for instance search instead of only performing global feature encoding. For instance, [38] and [39] take advantage of feature encoding on the local regions to build an image search and re-ranking pipeline. As illustrated in Table XII, DMVPIR achieves performance on par with both traditional BoW-based and recent CNN-based re-ranking approaches, suggesting the promise of the proposed framework. In particular, compared with CNN-based methods, our approach significantly surpasses [40] on Oxford5k by over 14% and reports comparable result on Paris6k with the same feature size. Additionally, DMVPIR consistently beats Faster R-CNN+CA-SR+QE which also makes use of the deep model pre-trained with VGG16 architecture [30] while enjoys a more compact representation. Although fine-tuning the VGG16 network brings further performance gains, DMVPIR still achieves higher re-ranking accuracy than Faster R-CNN+CS-SR+QE on Oxford5k and rivals the performance on Paris6k.

In addition to the above quantitative evaluations, we also present the qualitative results of different methods as shown in Fig. 4. It is observed that our scheme not only significantly
TABLE IX
COMPARISON OF DMINTIR AND OUR DMVPIR METHOD ON PARIS6K WITH ONLY TWO-VIEW DATA INVOLVED (%)

| Query   | DMINTIR | Proposed |
|---------|----------|----------|
|         | AP   | NDCG@50 | NDCG@100 | AP   | NDCG@50 | NDCG@100 |
| defense | 75.98 | 95.03   | 80.72   | 77.79 | 95.64   | 80.48    |
| eiffel  | 64.71 | 93.38   | 87.69   | 65.32 | 92.99   | 86.60    |
| invalids| 69.66 | 99.25   | 95.20   | 70.00 | 98.77   | 94.69    |
| louvre  | 68.68 | 96.20   | 75.68   | 72.02 | 87.70   | 77.49    |
| moulinrouge | 80.79 | 98.62 | 96.43   | 82.09 | 98.80   | 96.61    |
| museedorsey | 63.56 | 82.29 | 67.40   | 66.23 | 85.09   | 69.61    |
| notredame | 77.77 | 98.88 | 87.23   | 79.00 | 98.42   | 89.08    |
| pantheon | 90.75 | 99.38 | 95.35   | 92.98 | 99.81   | 95.72    |
| pompidou | 93.90 | 89.36   | 87.63   | 92.81 | 89.50   | 87.88    |
| sacrecoeur | 81.66 | 97.95 | 92.15   | 83.69 | 97.43   | 92.25    |
| triomphe | 84.61 | 99.25   | 97.43   | 87.06 | 99.26   | 97.48    |
| mean    | 77.46 | 94.56   | 87.54   | 79.00 | 94.86   | 87.99    |

TABLE X
PERFORMANCE OF THE PROPOSED DMVPIR USING PRE-TRAINED AND FINE-TUNED CNN FEATURES ON OXFORD5K

| Query      | Pre-trained | Fine-tuned |
|------------|-------------|------------|
|            | AP   | NDCG@50 | AP   | NDCG@50 |
| all_souls  | 0.7270 | 0.7469 | 0.8694 | 0.7605 |
| ashmolean  | 0.8161 | 0.8056 | 0.8569 | 0.8282 |
| balliol    | 0.8445 | 0.7354 | 0.9391 | 0.7655 |
| bodian     | 0.9332 | 0.9072 | 0.9710 | 0.9233 |
| christ_church | 0.7079 | 0.8084 | 0.7774 | 0.8540 |
| cirmarket  | 0.8507 | 0.8664 | 0.8676 | 0.9272 |
| hertford   | 0.8500 | 0.8136 | 0.9293 | 0.8570 |
| kelsie     | 1.0000 | 0.9712 | 1.0000 | 0.9962 |
| magdalen   | 0.3184 | 0.4760 | 0.4736 | 0.5838 |
| pit_rivers | 0.9882 | 0.9236 | 1.0000 | 0.9546 |
| radcliffe_camera | 0.9350 | 0.7866 | 0.9575 | 0.8024 |
| mean       | 0.8212 | 0.7917 | 0.8765 | 0.8330 |

TABLE XI
PERFORMANCE OF THE PROPOSED DMVPIR USING PRE-TRAINED AND FINE-TUNED CNN FEATURES ON PARIS6K

| Query      | Pre-trained | Fine-tuned |
|------------|-------------|------------|
|            | AP   | NDCG@50 | AP   | NDCG@50 |
| defense    | 0.7779 | 0.8048 | 0.8317 | 0.8522 |
| eiffel     | 0.6532 | 0.8660 | 0.6484 | 0.8735 |
| invalids   | 0.7000 | 0.9469 | 0.7749 | 0.9769 |
| louvre     | 0.7202 | 0.7749 | 0.7839 | 0.8120 |
| moulinrouge| 0.8209 | 0.9661 | 0.8698 | 0.9875 |
| museedorsey| 0.6623 | 0.6961 | 0.7981 | 0.8105 |
| notredame  | 0.7900 | 0.8908 | 0.8371 | 0.9342 |
| pantheon   | 0.9298 | 0.9572 | 0.9556 | 0.9749 |
| pompidou   | 0.9281 | 0.8788 | 0.9411 | 0.8892 |
| sacrecoeur | 0.8369 | 0.9225 | 0.9358 | 0.9417 |
| triomphe   | 0.8706 | 0.9748 | 0.8545 | 0.9729 |
| mean       | 0.7900 | 0.8799 | 0.8392 | 0.9114 |

TABLE XII
COMPARISON OF OUR APPROACH AND THE STATE-OF-THE-ART RE-RANKING METHODS ON TWO DATASETS (MAp), d AND K REFERS TO THE FEATURE DIMENSIONALITY AND THE VOCABULARY SIZE RESPECTIVELY

| BoW-based Methods | d | K | Oxford5k | Paris6k |
|-------------------|---|---|----------|---------|
| Reciprocal NN [41]| 500k | 0.814 | 0.803 |
| Database Saliency [42]| 1024 | 0.835 | 0.814 |
| HE+MA+PGM [43]| 100k | 0.737 | - |
| LS+R+LQE [44]| 25k | 0.788 | 0.848 |

| CNN-based Methods | d | K | Oxford5k | Paris6k |
|-------------------|---|---|----------|---------|
| CroW + QE [40]| 256 | 0.718 | 0.815 |
| Faster R-CNN+CA-SR+QE [39]| 512 | 0.749 | 0.848 |
| Faster R-CNN+CS-SR+QE* [39]| 512 | 0.678 | 0.784 |
| siaMAC+MAC+R+QE [45]| 512 | 0.850 | 0.865 |
| siaMAC+R+MAC+R+QE [45]| 512 | 0.829 | 0.836 |
| ROMIR [46]| 1024 | 0.8668 | 0.8331 |
| NetVLAD+ReSW+QE [47]| 512 | 0.766 | 0.870 |
| siaMAC+ReSW+QE [47]| 512 | 0.883 | 0.896 |

| Proposed using pre-trained CNN | d | K | Oxford5k | Paris6k |
|-------------------------------|---|---|----------|---------|
| Proposed using fine-tuned CNN| 128 | 0.8212 | 0.790 |

*achieved with two different fine-tuning strategies

improves the retrieval accuracy of the baseline but also demonstrates better performance than the state-of-the-art DMINTIR method. Specifically, with the help of PI learning, our approach is able to return more top ranked ground-truth images even when the query-related instances account for small regions with the surrounding complex visual background or are partially occluded by other objects (e.g., tree, person, lamp post) in the image.

4) Large-Scale Evaluations: In order to evaluate the performance of our DMVIPR approach at a larger scale, we merge the Flickr100k with Oxford5k and Paris6k, producing Oxford105k and Paris106k datasets respectively. Different from the aforementioned setup, VLAD+ and fine-tuned
CNN codes are used as two-view features in DMVPIR for large-scale evaluations. Fig. 5 illustrates the performance of DMVPIR when the distractor images are incrementally added to Oxford5k and Paris6k. It is shown that DMVPIR exhibits limited performance drop with an increasing dataset size. Particularly, when the complete 100k images are incorporated, DMVPIR reports respective mAP scores at 85.15% and 83.68% on Oxford105k and Paris106k respectively, only decreasing by 0.8% and 1.97% compared with the accuracies of 85.95% and 85.65% achieved on Oxford5k and Oxford6k. This sufficiently suggests the desirable scalability of our method.

Table XIII gives the comparison of DMVPIR with the state-of-the-art re-ranking methods on both large-scale datasets (MAP%).

| Method          | Dim | Oxford105k | Paris106k |
|-----------------|-----|------------|------------|
| NetVLAD+ResW+QE | 512 | 72.5       | 82.5       |
| siaMAC+ResW+QE  | 512 | 86.4       | 84.8       |
| siaMAC+MAC+R+QE | 512 | 81.8       | 78.8       |
| Proposed        | 128 | 85.15      | 83.68      |

Table XIV gives the time cost of our scheme on two datasets(s).

| Datasets | model training (s) | re-ranking (s) |
|----------|--------------------|---------------|
| Oxford5k | 79.35              | 0.14          |
| Paris6k  | 78.66              | 0.18          |

G. Parameter Analysis

We now thoroughly discuss the impact of various parameters in the proposed DMVPIR framework on the re-ranking performance, including the vocabulary size $k$ for generating the TE and VLAD+ features, the six hyperparameters to tune in Eq. (8), the length of the shortlist $K$ for user interaction and the subspace dimensionality $d$.

1) The Impact of the Vocabulary Size: In our method, the re-ranking performance largely depends on the multiple features including TE and VLAD+ both of which need a well-trained vocabulary. As for the TE signature, we follow the standard practice [34], [48], [49] to set the vocabulary size $k$ as 64 for generating a 8,064 dimensional feature with low frequency dimensions removed, since further increasing $k$ yields limited boost in performance while severely compromises the computational efficiency [34]. In terms of VLAD+, we use the vocabulary of the same size as in [1]. In order to explore the impact of $k$ on the performance, we further increase $k$ to 256 and 1024 respectively. Consistent with [1], the resulting performance gains consist in less than 1% and 1.5% at the cost of considerable growth in memory footprint and computational overhead. Therefore, we use $k = 256$ for VLAD+ in all tests for the tradeoff between accuracy and efficiency.

2) The Impact of Tradeoff Hyperparameters: For model selection, we evaluate different combinations of hyperparameters $(\lambda, \lambda^*, \beta, \beta^*, \gamma, C)$ on a single query group “all_souls” to obtain the optimal ones. For the sake of simplicity, we set $\lambda = \lambda^*$ and $\beta = \beta^*$ in our case. As illustrated in Fig. 6, the highest AP score is achieved at 90.44% when the hyperparameters take the values of $\{0.6, 0.6, 0.1, 0.1, 1.5, 1.3\}$. Thus, we use this parameter setting for evaluations on both datasets. Overall, limited fluctuation in the performance with different hyperparameter combinations is observed, which, to some extent, implies the desirable property of DMVPIR in hyperparameter insensitivity. This can be explained by the fact that introducing PI into our framework brings the performance boost varying within a certain range dependent on the tradeoff between respective regularization terms.

3) The Influence of the User Interaction: Analogous to [1], user interaction is involved in training DMVPIR for obtaining the query-relevant positive images with annotated PI regions from the top returned shortlist. Thus, it is essential to explore the effect of the shortlist size $K$ on the re-ranking performance. Fig. 7 gives the DMVPIR performance with varying $K$ on the query “all_souls.” It is shown that the
re-ranking accuracy improves with an increase in $K$, yet the growth stabilizes; i.e., further increasing $K$ leads to limited performance improvements in spite of the additional user interaction and human workload. In practise, we do not take into account the case when $K$ is greater than 50 to avoid a user click indicating the query-relevance and annotating the PI region is required in our case. Therefore, larger $K$ incurs unaffordable burden and thus adversely affect the efficiency of the whole system. In implementation, we assume that $K = 40$ is a reasonable choice with desirable compromise between accuracy and efficiency. Since the images with low ranks are recognized as the negative training data without using any user interaction, the size of the negative set is empirically set to be 100. Thus, we use this parameter setting (40/100) for all query groups.

4) The Overhead of the Human PI Annotations: As mentioned above, in addition to the binary labels indicating the query-relevance, the human PI annotations are also required for obtaining the privileged data in our approach. Since image labeling in our scenario includes both user clicks and PI annotations, it can be observed from Table II that approximately 16 and 34 images are manually labeled for each query instance. Thus, a total of 880 and 1870 samples accounting for a comparatively small proportion of the database size are
annotated on both datasets. Since one annotation is to simply plot a ROI bounding box, it costs 1s approximately, and thus the overall time of human annotations on both datasets amounts to roughly 15 and 32 min with an affordable overhead. More importantly, these annotations can be pre-computed without affecting the on-the-fly efficiency.

5) The Effect of the Subspace Dimension: Fig. 8 illustrates the performance of our approach with different low-dimensional subspaces on query “all souls.” Overall, the retrieval performance grows with an increase in $d$ when the highest mAP score is reported at 72.09% with $d = 128$. Besides, a slight performance drop is observed when the subspace dimension exceeds 128. Interestingly, increasing the subspace dimension may not bring performance boost, perhaps due to the feature redundancy present in the original multi-view spaces. As a result, we use the 128-dimensional subspace in our scenario.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we propose a discriminative multi-view PI-aware image re-ranking method termed as DMVPIR. Different from the conventional multi-view re-ranking approaches, we take into consideration the supplementary PI cues, since they are capable of characterising the dominant information in the image that captures the query intention. In model training, the auxiliary PI data and the original training data are simultaneously delivered to the unified multi-view embedding framework for producing a PI-aware subspace with sufficient discriminating power. For accurate re-ranking, the PI-aware latent representations can be obtained by projecting the multi-view features of the target images onto the underlying space for efficient similarity measure. Extensive evaluations on the public datasets for landmark retrieval task demonstrate that our scheme outperforms the classical multi-view re-ranking strategies and achieves the comparable results on par with the state-of-the-arts.

Despite being effective, DMVPIR relies on the user interaction for PI annotation. In the future, we will further study the generalization capability of the re-ranking model when the PI cues are limited.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and all anonymous reviewers for their valuable and constructive comments to improve the quality of the manuscript.

REFERENCES

[1] J. Li, C. Xu, W. Yang, C. Sun, and D. Tao, “Discriminative multi-view interactive image re-ranking,” IEEE Trans. Image Process., vol. 26, no. 7, pp. 3113–3127, Jul. 2017.
[2] R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2911–2918.
[3] V. Vapnik and A. Vashist, “A new learning paradigm: Learning using privileged information,” Neural Netw., vol. 22, nos. 5-6, pp. 544–557, Jul. 2009.
[4] V. Sharma and S. Quatramento, and C. H. Lampert, “Learning to rank using privileged information,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 825–832.
[5] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, “Salient object detection by composition,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 1028–1035.
[6] Y. Yu, Y. Rui, and D. Tao, “Click prediction for Web image reranking using multimodal sparse coding,” IEEE Trans. Image Process., vol. 23, no. 5, pp. 2019–2032, May 2014.
[7] J. Yu, D. Tao, M. Wang, and Y. Rui, “Learning to rank using user clicks and visual features for image retrieval,” IEEE Trans. Cybern., vol. 45, no. 4, pp. 767–779, Apr. 2015.
[8] J. Yu, Y. Rui, and B. Chen, “Exploiting click constraints and multi-view features for image re-ranking,” IEEE Trans. Multimedia, vol. 16, no. 1, pp. 159–168, Jan. 2014.
[9] Y. Li, B. Geng, D. Tao, Z.-J. Zha, L. Yang, and C. Xu, “Difficulty guided image retrieval using linear multiple feature embedding,” IEEE Trans. Multimedia, vol. 14, no. 6, pp. 1618–1630, Dec. 2012.
[10] C. Deng, R. Ji, D. Tao, X. Gao, and X. Li, “Weakly supervised multi-graph learning for robust image reranking,” IEEE Trans. Multimedia, vol. 16, no. 3, pp. 785–795, Apr. 2014.
[11] F. Yang, B. Matei, and L. S. Davis, “Re-ranking by multi-feature fusion with diffusion for image retrieval,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., Jan. 2015, pp. 572–579.
[12] C. Xu, D. Tao, and C. Xu, “Multi-view intact space learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 12, pp. 2531–2544, Dec. 2015.
[13] Z. Ji, Y. Yu, Y. Pang, J. Guo, and Z. Zhang, “Manifold regularized cross-modal embedding for zero-shot learning,” Inf. Sci., vol. 378, pp. 48–58, Feb. 2017.
[14] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui, “Click-through-based cross-view learning for image search,” in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), 2014, pp. 717–726.
[15] S. Pang, J. Ma, J. Xue, J. Zha, and V. Ordonez, “Deep feature aggregation and image re-ranking with heat diffusion for image retrieval,” IEEE Trans. Multimedia, vol. 21, no. 6, pp. 1513–1523, Jun. 2019.
[16] X. Xu, W. Li, and D. Xu, “Distance metric learning using privileged information for face verification and person re-identification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 12, pp. 3150–3162, Dec. 2015.
[17] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, “Image classification by cross-media active learning with privileged information,” IEEE Trans. Multimedia, vol. 18, no. 12, pp. 2494–2502, Dec. 2016.
[18] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, “Information bottleneck learning using privileged information for visual recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1496–1505.
[19] H. Yang, J. T. Zhou, J. Cai, and Y. S. Ong, “MIML-FCN+ Multi-instance multi-label learning via fully convolutional networks with privileged information,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1577–1585.
[20] Z. Shi and T.-K. Kim, “Learning and refining of privileged information-based RNNs for action recognition from depth sequences,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3461–3470.
[21] A. Gordo and D. Larlus, “Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6589–6598.
