iQPP: A Benchmark for Image Query Performance Prediction

Eduard Poesina, Radu Tudor Ionescu, Josiane Mothe

iQPP: A Benchmark for Image Query Performance Prediction

Eduard Poesina
Department of Computer Science,
University of Bucharest
Bucharest, Romania
eduardgabriel.poe@gmail.com

Radu Tudor Ionescu*
Department of Computer Science,
University of Bucharest
Bucharest, Romania
raducu.ionescu@gmail.com

Josiane Mothe
INSPE, IRIT UMR5505 CNRS,
Université Toulouse Jean-Jaurès
Toulouse, France
josiane.mothe@irit.fr

ABSTRACT

To date, *query performance prediction* (QPP) in the context of content-based image retrieval remains a largely unexplored task, especially in the query-by-example scenario, where the query is an image. To boost the exploration of the QPP task in image retrieval, we propose the first benchmark for *image query performance prediction* (iQPP). First, we establish a set of four data sets (PASCAL VOC 2012, Caltech-101, ROxford5k and RParis6k) and estimate the ground-truth difficulty of each query as the average precision or the precision@k, using two state-of-the-art image retrieval models. Next, we propose and evaluate novel pre-retrieval and post-retrieval query performance predictors, comparing them with existing or adapted (from text to image) predictors. The empirical results show that most predictors do not generalize across evaluation scenarios. Our comprehensive experiments indicate that iQPP is a challenging benchmark, revealing an important research gap that needs to be addressed in future work. We release our code and data as open source at https://github.com/Eduard6421/iQPP, to foster future research.

CCS CONCEPTS

• **Information systems** → **Information retrieval query processing**; **Information retrieval**; **Test collections**; **Retrieval effectiveness**; *Retrieval tasks and goals.*

ACM Reference Format:
Eduard Poesina, Radu Tudor Ionescu, and Josiane Mothe. 2023. iQPP: A Benchmark for Image Query Performance Prediction. In *Proceedings of Arxiv (Preprint).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Query performance prediction (QPP) (also known as query difficulty prediction or query difficulty estimation) is the task of estimating the effectiveness of a set of search results retrieved in response to a query, without relevance judgments [11]. QPP is extremely important in identifying poorly performing queries, which might require the use of a more powerful retrieval system or a query expansion method to improve the search results. By predicting the performance of queries, the system can optimize the search, presenting the most relevant information to the user. This can lead to a better user experience and increased satisfaction. Furthermore, a better understanding of query difficulty can also suggest improvements to the underlying algorithms and systems used in information retrieval, leading to more effective and efficient systems in the future. Hence, QPP is a crucial component for achieving effective retrieval results and improving the overall search experience. The importance of QPP has been widely recognized in text retrieval [3, 10–13, 18, 24–26, 35, 39, 40, 43, 44, 51, 54–56, 58, 63–65, 69, 76, 77], being an actively studied task nowadays [1, 2, 9, 14–16, 22, 32, 60]. However, in the context of content-based image retrieval (CBIR), the QPP task received comparably lower attention from the research community, with only a few studies addressing the topic [33, 34, 42, 45, 46, 68, 70–72, 74]. Furthermore, only a handful of papers [42, 46, 68, 72] study query difficulty prediction in the query-by-example scenario, where the query is an image.

Since we consider that QPP in text and image retrieval is equally important, the goal of this work is to raise the level of exploration of QPP in the image domain to the same level of exploration currently observed in the text domain. To this end, we propose the first benchmark for image query difficulty prediction, which we term iQPP, in the context of query-by-example content-based image retrieval, where images have to be retrieved given an image query. Our benchmark comprises four data sets (PASCAL VOC 2012 [21], Caltech-101 [41], ROxford5k [49] and RParis6k [49]), two image retrieval systems [50, 53], as well as several pre-retrieval and post-retrieval query performance predictors, for which we deliver the predicted and ground-truth performance levels for two effectiveness measures. The data sets are chosen based on their popularity, aiming to accommodate a high variety of images, from landmark photos to pictures of various object classes. The retrieval systems are chosen due to their state-of-the-art performance coupled with the availability of the open source models. Since research on image QPP is scarce, we turn our attention to proposing several novel predictors along with our benchmark. First of all, we propose four novel pre-retrieval predictors, namely (i) the magnitude of the reconstruction error of denoising [73] or masked [28] auto-encoders trained on the database, (ii) the density of the k-means cluster to which the query image embedding is assigned, (iii) the confidence distribution of a classification head attached to the embedding layer of the retrieval model, and (iv) the score predicted by a fine-tuned ViT model [20]. We note that the first three pre-retrieval predictors are unsupervised, while the last one is supervised on labeled training queries. Second of all, we propose four novel post-retrieval predictors, namely (v) a query feedback method redesigned for the image domain, (vi) the intersection over union (IoU) ratio for the results retrieved while iteratively removing the most discriminative features, (vii) the dispersion (variance) of the embeddings of the retrieved results, and ($viii$) the difficulty score predicted by a regression model applied on all the other predictors. Among the

*Corresponding author.

Preprint, 2023, Arxiv

ACM ISBN 978-1-4503-XXXX-X/18/06.
https://doi.org/XXXXXXX.XXXXXXX

proposed post-retrieval predictors, only the last one is supervised. We compare the proposed predictors with existing methods for image difficulty [30, 67] or query performance [68] prediction, as well as the standard deviation of predicted relevance scores, which was originally proposed for QPP in the text domain [13].

We carry out a comprehensive set of experiments on the four aforementioned data sets to evaluate the capacity of the designated predictors in predicting the ground-truth difficulty of the queries. As in previous studies [16, 25], we employ the Pearson correlation and the Kendall τ correlation as evaluation measures. Our results indicate that supervised and post-retrieval predictors tend to achieve better performance. However, the empirical evidence shows that the image QPP task is far from being solved, since none of the existing or proposed predictors can surpass a Kendall τ of 0.65, and there is no predictor to consistently surpass its competitors on all four data sets. This observation indicates that there is an important research gap in image QPP, which needs to be addressed in future work. To ensure the reproducibility of the results and foster future research, we release our code and data as open source at https://github.com/Eduard6421/iQPP.

In summary, our contributions are threefold:

- We develop the first benchmark for image query performance prediction in the query-by-example CBIR setting.
- We propose eight novel pre-retrieval and post-retrieval image query performance predictors.
- We present extensive experiments on the four image data sets included in our benchmark.

2 RELATED WORK

2.1 QPP in ad-hoc text retrieval

Query performance prediction became popular in the mid-2000s in ad-hoc text retrieval [7, 11]. Since then, researchers explored QPP using a broad range of approaches, giving rise to various categories of predictors. For example, query performance predictors can be categorized into pre-retrieval and post-retrieval. Pre-retrieval predictors aim to predict query performance prior to querying the document collection, while post-retrieval ones imply carrying out a search before estimating effectiveness [7]. Some popular pre-retrieval predictors are term rareness, specificity or distribution [17, 27, 61, 79], term ambiguity [17, 44] and query complexity [23, 44]. Hauff et al. [25] analyzed 22 pre-retrieval predictors and concluded that the predictors jointly considering the query and the collection are better than the ones only considering the query. They also observed that the performance of predictors depends on the test collection and the underlying retrieval model. Due to the extra available information, post-retrieval predictors are usually more effective, although less efficient [65]. Many of these predictors mainly consider the homogeneity, magnitude or variance of the retrieved document scores [7, 11, 62, 65, 78]. For example, the clarity score [11] measures the term distribution probability of the retrieved documents and the term distribution probability of the whole collection [11]. Other post-retrieval predictors measure the divergence or stability of the retrieved document list when the query is perturbed using relevance feedback [81], sub-queries [75] or different scoring functions [4]. Individual predictors may have high variance [10, 32, 76] and are not robust across collections. Thus, some studies have considered combining predictors. For non-supervised methods, studies have focused on analyzing classes of queries or the link between system/query features and effectiveness, using for example factorial analysis [6, 19]. Supervised methods represent however the most common means to combine predictors in QPP, comprising approaches based on linear combinations [10, 32, 58, 80], genetic algorithms [5], and neural networks [14, 76]. Supervised predictors are usually evaluated using k-fold or leave-one-out cross-validation [7].

To our knowledge, there is no clear benchmark for text QPP. However, there are some common practices. First, several collections are used in a study, allowing researchers to evaluate the robustness of a predictor across collections. QPP for ad-hoc search relies mostly on TREC collections[1] [10, 25, 75]. Evaluation encompasses several aspects, such as the measure used to assess the retrieval system performance, and the one used to evaluate the QPP accuracy. As performance measures, researchers employ the usual ad-hoc retrieval effectiveness measures, e.g. average precision (AP), NDCG, or precision at a certain cut-off point of the retrieved list [10, 25, 32, 57, 75, 79]. The prediction accuracy is measured by considering the actual effectiveness and the predicted performance. To evaluate this relation, most of the studies consider the Pearson correlation. Since the link between the two measures may not be linear, Kendall and Spearman correlations are often employed as additional measures [10, 25, 32, 57, 75, 79]. We follow similar evaluation principles in constructing our iQPP benchmark.

Different for the mainstream area studying QPP in ad-hoc text retrieval, we underline that studies on QPP in image retrieval require new predictors adapted to the image domain, as well as distinct data sets, containing images instead of text documents. Hence, research on image QPP naturally diverges to a different direction, specific to the image domain. We cover this direction in the next section.

2.2 QPP in image search

One of the first contributions on QPP in the image domain is the work of Xing et al. [74]. The authors used query words and context information to compute a set of four text-based pre-retrieval features and train a model for QPP in image retrieval. Subsequent studies turned their attention to post-retrieval predictors. Tian et al. [71] focused on QPP for web image search, where the query is a piece of text and the results are images. The authors proposed the visual clarity score inspired from the clarity score defined for texts [11], which measures the difference in the distribution of the top retrieved images and the whole collection. They also use the coherence score based on the visual similarity among the retrieved images. In a later study, Tian et al. [70] introduced an approach to reconstruct an image query based on the images returned in response to a text query. They estimated query performance via the differences between the ranked lists of the text query and the reconstructed image query. Nie et al. [45] presented a two-stage pipeline for QPP. In the first stage, ranked image lists are classified into person-related and non-person-related. In the second stage, the relevance probability of the query is estimated via graph-based learning, as well as visual content. The authors adapt the visual

[1] Text REtrieval Conference (http://trec.nist.gov)

content representation to the class predicted in the first stage. Jia et al. [33, 34] introduced a post-retrieval predictor that divides the retrieved images into pseudo-positive and pseudo-negative via pseudo-relevance feedback. Next, a voting scheme is applied to label the images as relevant or not. The pseudo-relevance labels are further used to provide an estimate for the AP.

To our knowledge, there are only a few studies [42, 46, 68, 72] that try to predict the performance of image queries. Li et al. [42] proposed a post-retrieval predictor that examines the top ranked images using the clarity score, the spatial consistency of local descriptors, and the appearance consistency of global features. The method is specifically designed for image retrieval models based on the bag-of-visual-words [31, 47]. Pedronette et al. [46] proposed an unsupervised post-retrieval predictor based on the cluster hypothesis [38], considering that the images belonging to a highly effective ranked list should appear in the ranked lists of each other. The authors study different ways to measure the density of reciprocal references among retrieved images, conducting experiments on relatively small data sets (each of about 1,000 images). Sun et al. [68] proposed a supervised post-retrieval predictor that transforms the ranked list of images into a similarity or correlation matrix which is further given as input to a convolutional neural network (CNN). Valem et al. [72] extended the work of Sun et al. [68] by generating synthetic ranked lists as training data for the CNN, requiring a more complex training procedure.

As related studies on QPP in ad-hoc text retrieval, methods studying QPP in image search commonly employ AP and NDCG as effectiveness measures, and the Pearson coefficient as query performance prediction measure [42, 45, 46, 68, 72].

Most of the above studies, e.g. [33, 34, 45, 70, 71, 74], predict the performance of text queries in CBIR. Our study is among the few works [42, 46, 68, 72] studying the performance of image queries. To the best of our knowledge, we are the first to propose a benchmark for image QPP in image retrieval. The benchmark includes four data sets, two retrieval models, and twelve predictors. Among the considered predictors, there are eight novel QPP approaches, adding to the novelty of our study. Furthermore, we are the first to propose pre-retrieval predictors for image queries.

3 PROBLEM FORMULATION

Let $\mathcal{D} = \{x_1, x_2, ..., x_n\}$ be a collection (database) of images, where n is the number of images inside the collection. Given a query image q, an image retrieval system R returns a ranked list of k images denoted as $\rho_{q,R} = \left[x_1^{(q,R)}, x_2^{(q,R)}, ..., x_k^{(q,R)}\right]$, where $k \leq n$ and $x_i^{(q,R)} \in \mathcal{D}$ is the i-th most similar image to q, as evaluated by R. A retrieval model is a tuple $R = (f, \delta)$, where $f : \mathbb{R}^{h \times w} \rightarrow \mathbb{R}^d$ is a model, e.g. neural network, that maps each input image x_i of $h \times w$ pixels to a d-dimensional embedding vector v_i, and $\delta : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a distance or similarity function. To obtain the ranked list, the model f is applied to the query q (online) and the images $x_i \in \mathcal{D}$ (offline). Then, the measure δ is applied on each pair of embedding vectors v_q and v_i, and, based on the returned distance or similarity values, the images are sorted in descending or ascending order, respectively. Let $\rho_q = \left[x_1^{(q)}, x_2^{(q)}, ..., x_k^{(q)}\right]$ denote the ground-truth ranked list of the most similar k images from $\mathcal{D}$ to the query image q. Let $\mathcal{R}_k$ denote the set of all possible ranked lists of k images. Let $P : \mathcal{R}_k \times \mathcal{R}_k \rightarrow \mathbb{R}$ be a performance measure, e.g. average precision or precision@k, that estimates the effectiveness of the retrieval system R on the query q, considering the ground-truth and predicted rankings. For any query image q and retrieval model R, the goal of QPP is to predict the value returned by $P\left(\rho_q, \rho_{q,R}\right)$, without having access to the ground-truth list ρ_q.

4 PRE-RETRIEVAL PREDICTORS

To save space and avoid reiterating through the predictors presented here in a later section, we include hyperparameter choices in the current section. We use the same presentation format for the post-retrieval predictors described in Section 5.

4.1 Baselines

The baselines presented below were previously used for generic image difficulty estimation [30, 67]. We repurpose them as pre-retrieval query performance predictors.

Generic image difficulty. Ionescu et al. [30] proposed an image difficulty predictor trained on a data set collected with the help of human annotators. The ground-truth difficulty score of each image is based on measuring the average time taken by human annotators to search for different objects (present or missing) in the respective image. The image difficulty predictor is based on an ensemble of VGG networks [66] pre-trained on ImageNet [59], where the classification layer is replaced with a Support Vector Regression (SVR) model. We reproduce the original difficulty regressor of Ionescu et al. [30] and apply it on each query image. As discussed in [30], the image difficulty regressor scores images on a continuous scale, such that images with one object in a plain background receive low scores, and images with many objects in a complex scene (background) receive high scores.

Number of objects divided by their area. In object detection, Soviany et al. [67] observed that the number of detected objects and the total area covered by the objects are positively and negatively correlated with image difficulty [30], respectively. In other words, an image depicting one object covering the entire image is easy for visual search, while an image depicting multiple objects covering a small area, e.g. because the objects are photographed from far away, is difficult. Thus, following the intuition of Soviany et al. [67] and transposing it to image QPP, the performance of an image query can be estimated as the number of objects divided by their average bounding box area. Let $\mathcal{B} = \{b_1, b_2, ..., b_m\}$ be the set of m bounding boxes detected in query image q by a pre-trained object detector, namely a Faster R-CNN [52] with a ResNet-50 [29] backbone. Let h_i and w_i denote the height and width of bounding box b_i. The difficulty score of an image q is defined as follows:

$$s(q, \mathcal{B}) = \frac{m}{\frac{1}{m} \sum_{i=1}^{m} w_i \cdot h_i}. \quad (1)$$

4.2 Proposed

Auto-encoder reconstruction. We train two types of auto-encoders (AEs) to reconstruct the images in the collection $\mathcal{D}$, namely denoising AEs [73] and masked AEs [28]. A denoising AE corrupts input

images with Gaussian noise and learns to reconstruct the original (uncorrupted) inputs. By adding noise, the AE avoids learning the identity mapping. Masked auto-encoders represent a modern attempt to learn discriminative representations by dividing input images into a grid of patches and eliminating a significant amount of patches (typically 75%). A masked AE learns to reconstruct the missing patches. Both AE models embed images into a latent manifold that captures the most important patterns in the training data distribution. Auto-encoders are known for their capacity to represent images from the training distribution very well. However, as soon as an image from a different distribution is given as input, AEs exhibit poor reconstruction capabilities. We leverage this property and propose to train a denoising or masked AE on the collection of images $\mathcal{D}$ and apply it on the query image q. Intuitively, if the query image can be accurately reconstructed by the model, then the query is likely to be easy, i.e. a low reconstruction error means that the query image belongs to the training data distribution. In contrast, a query image that is poorly reconstructed by the model indicates that it is not well represented by the training distribution. Moreover, we consider that the higher the reconstruction error, the farther away the query image is from the training distribution.

The denoising AE is based on a convolutional architecture. The encoder is composed of four convolutional layers with ReLU activations and 2D batch normalization, while the decoder is formed of four convolutional layers with ReLU activations and nearest neighbor upsampling applied between layers. The masked AE is based on transformer blocks. The encoder uses an embedding dimension of 768 and comprises 18 transformer blocks, each with 16 attention heads. The decoder is lighter, having only 8 transformer blocks based on 512-dimensional embeddings. We employ Adam [36] and the mean squared error (MSE) to train both AEs. We set the learning rate to 10^{-3} and the mini-batch size to 12 for both models. During inference, the MSE is used as effectiveness score for the query images.

K-means cluster density. We propose to cluster the embedding vectors given by the model f for all images in the collection $\mathcal{D}$, using k-means clustering. The embedding of the query image $v_q = f(q)$ is assigned to one of the clusters denoted as C_j, represented by the centroid c_j, where $1 \leq j \leq K$ and K is the number of clusters. We consider that the query is *easy* if the cluster C_j has many points densely packed together, and *hard* if the cluster C_j has only a few points spread over a large area. Aside from the cluster density, which is the same for all queries assigned to cluster C_j, the relation between a query and the cluster centroid provides another clue about the difficulty of the query. More precisely, the farther the embedding v_q is from the cluster centroid, the more difficult the query. We combine the aforementioned conjectures into a closed form equation and compute the difficulty score $s(q)$ of query q as follows:

$$s(q) = \frac{\delta(c_j, v_q) + var(C_j)}{|C_j|}, \tag{2}$$

where $|C_j|$ is the cardinal of cluster C_j and $var(C_j)$ is the variance of cluster C_j. We tune the hyperparameter K taking values between 50 and 300, with a step of 50. We obtain optimal results with $K = 150$.

Confidence distribution of self-supervised classification head. We propose to equip the embedding model f with a softmax classification head and train it on the image database $\mathcal{D}$. After embedding each query with the model f, we additionally pass it through the classification head to obtain a class distribution. We conjecture that easy queries are likely to be assigned to one class with high confidence. At the same time, hard queries will be assigned to multiple classes with various confidence levels, indicating that the classifier is not certain what class label should be assigned to the query image. We alternatively employ two measures to estimate the class confidence distribution: dispersion and kurtosis.

Since not all image collections have class labels attached, we train the classification head in a self-supervised manner. Following Caron et al. [8], we first cluster the embedding vectors v_i into K clusters via k-means. Then, we use the cluster assignments as target class labels for our classification head. The head comprises two hidden layers of 50 neurons each with ReLU activations, and a softmax layer. We employ the Adam optimizer [36] with a learning rate of 10^{-4} to minimize the cross-entropy loss. During training, the embedding model f is frozen. We tune the number of clusters in the range 50-300, at a step of 50. The optimal value for K is 150.

Fine-tuned ViT. Visual transformers [20] are a family of deep learning architectures that apply the self-attention mechanism to visual data. This allows the model to handle long-range dependencies and spatially-varying information in images, recently leading to improved performance in tasks such as image classification, segmentation, and generation. The power of these models relies on using a two-stage training process: (i) large scale pre-training and (ii) fine-tuning on downstream tasks. Following this procedure, we propose to fine-tune a visual transformer (ViT) model [20] to predict the performance levels of training query images. We select the ViT-B16 backbone pre-trained on ImageNet [59], and fine-tune it on the QPP task for 100 epochs with the Adam optimizer. We tune the learning rate (considering 10^{-2}, 10^{-3} and 10^{-4} as possible values) and the mini-batch size (considering 8, 16 and 32 as possible values) using grid-search.

5 POST-RETRIEVAL PREDICTORS

All post-retrieval predictors take into account the returned list of results. Since we use the precision@100 to determine the ground-truth effectiveness of queries, we set $k = 100$ and use the top k retrieved results for the post-retrieval predictors presented below.

5.1 Baselines

Score variance. We use the score variance introduced by Cummins et al. [13] as our first baseline post-retrieval predictor. The score variance behaves as an estimator for the influence of characteristics that are not related to the query. Because of its simplicity, this predictor can be applied to images without any adaptation. Formally, the estimated query performance is computed as follows:

$$s(q) = var\left(\delta\left(v_q, v_i^{(q,R)}\right)\right). \tag{3}$$

Correlation-based CNN. Sun et al. [68] proposed to train a CNN on similarity or correlation matrices. A ranked list of images is

turned into a matrix by computing the pairwise similarities between all pairs of returned images. The authors found that the best approach to determine the pairwise similarity between two images is to pass them through another CNN and compute the similarity between the resulting embedding vectors. To obtain these embedding vectors, we employ the embedding model f which comes with the retrieval system. In our case, the model f is a ResNet-101 [29].

The CNN that learns to predict query performance based on similarity matrices is formed of three convolutional-pooling blocks, followed by two dense layers. We use the exact same configuration for the CNN as Sun et al. [68]. The model is optimized to minimize the mean squared error between the ground-truth and the predicted query performance. Following Sun et al. [68], we train the CNN for 100 epochs with the Adam optimizer. We tune the learning rate (considering 10^{-2}, 10^{-3} and 10^{-4} as possible values) and the mini-batch size (considering 8, 16 and 32 as possible values) using grid-search.

5.2 Proposed

Adapted query feedback. Our first post-retrieval predictor is a redesigned version of the query feedback proposed by Zhou et al. [81]. In text retrieval, the query feedback is computed as the overlap between the ranked lists retrieved by a system for the original query and the expanded query, respectively. The query expansion is performed considering the list of documents retrieved for the original query. In the image domain, we perform query expansion by finding the median image in the list $\rho_{q,R}$ returned by the system R for the query q. We define the median image as the image $q' \in \rho_{q,R}$ having the closest embedding to the average embedding of the returned list:

$$q' = \underset{x_i^{(q,R)} \in \rho_{q,R}}{\operatorname{argmin}} \ \delta\left(f\left(x_i^{(q,R)}\right), \frac{1}{k}\sum_{j=1}^{k} f\left(x_j^{(q,R)}\right)\right), \tag{4}$$

where f is the embedding model. Finally, our adapted query feedback is given by the IoU ratio between $\rho_{q,R}$ and $\rho_{q',R}$.

Iterative removal of discriminative features. We propose an approach for QPP inspired by the unmasking technique of Koppel et al. [37]. The technique was initially proposed for the task of text author identification, serving as an estimator for how distinguishable texts are from each other. The method relies on gradually removing distinguishable features learned by a linear classifier, leveraging the idea that if two texts are written by the same author, then they should differ by a relatively small amount of features.

We adapt the aforementioned principle with the purpose of identifying the level of similarity between the query and the retrieved images. We take the top k retrieved images and compute the Hadamard product between their embeddings and the query embedding v_q to identify the features with higher correlation. We sort the features in descending order of the correlation and remove the top m features from the embeddings of the query and the database images. This process is repeated l times. To measure query performance, we employ the IoU score computed over the sets of images retrieved at all iterations. Intuitively, if a query is easy to handle, systematic removals of features should not strongly deter the original answer, as the query exhibits a larger set of highly correlated features. We employ grid-search to perform hyperparameter tuning, obtaining optimal results with $m = 50$ and $l = 15$.

Table 1: Important statistics about the data sets included in the iQPP benchmark.

Data set	#images	#train queries	#test queries
PASCAL VOC 2012	17,125	700	700
Caltech-101	9,146	700	700
ROxford5k	5,063	-	70
RParis6k	6,392	-	70

Embedding variance. Inspired by the intuition behind the predictor based on score variance [65], we propose a predictor based on estimating the variance of the embeddings of the retrieved images. Formally, the effectiveness of query q is estimated by:

$$s(q) = var(f(\rho_{q,R})), \tag{5}$$

where f is the embedding model and $\rho_{q,R}$ is the list of k images returned by the system R, as defined in Section 3. From another perspective, we can regard the predictor defined in Eq. (5) as a post-retrieval version of the pre-retrieval predictor based on k-means cluster density defined in Eq. (2). Indeed, the returned images $\rho_{q,R}$ are likely to have a high overlap with the images in cluster C_j. The overlap is higher, as the query image is closer to the cluster centroid.

Meta-regressor. We propose a meta-regression model to leverage the results of all of the previously described methods. We first normalize the values of all the other predictors between 0 and 1. Next, we train a Support Vector Regression (SVR) model and employ grid-search to identify the optimal values of the penalty term $C \in \{0.1, 1, 10, 100\}$, the fraction of support vectors $\nu \in \{0.1, 0.25, 0.5, 0.75\}$, and the kernel type (linear or RBF). We identify $C = 100$, $\nu = 0.25$ and the Radial Basis Function (RBF) kernel as the optimal choices.

6 BENCHMARK RESOURCES

6.1 Data sets

To evaluate the performance of the predictors, we assemble a lineup of four image data sets to form our novel benchmark, namely PASCAL VOC 2012 [21], Caltech-101 [41], ROxford5k [49] and RParis6k [49]. While ROxford5k and RParis6k are well established CBIR data sets, we repurpose PASCAL VOC 2012 and Caltech-101 to accommodate the QPP task. The well-established ROxford5k and RParis6k data sets consist of 70 queries each. To achieve a more robust evaluation on PASCAL VOC 2012 and Caltech-101, we establish a set of 700 test queries for these two data sets, increasing the number of queries by an order of magnitude compared to the well-established ROxford5k and RParis6k, as shown in Table 1.

In Figure 1, we illustrate two queries (one easy and one hard) per data set, and the top five results for each query. Note that ROxford5k and RParis6k contain landmark images, while PASCAL VOC 2012 and Caltech-101 contain images of various objects.

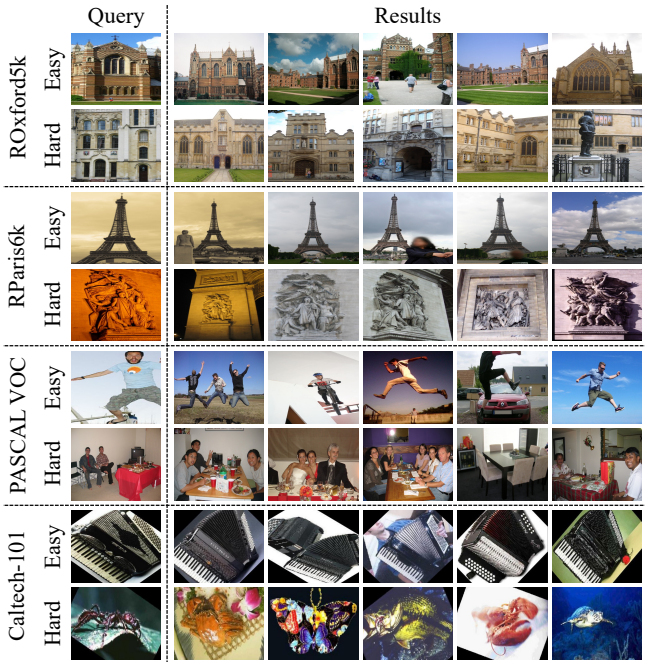


Figure 1: Examples of easy (high performance) and hard (low performance) queries from the four data sets included in the iQPP benchmark. For each query, we show the top five results returned by the system of Radenović et al. [50] to better illustrate the performance levels of the chosen queries. Best viewed in color.

PASCAL VOC 2012. PASCAL VOC 2012 [21] is a data set of 17,125 images covering a wide range of computer vision tasks such as image segmentation, object detection and object recognition. It contains a broad array of real-life scenes, depicting more than 20 object categories. The data set provides annotations for bounding boxes, object classes and contours.

Our work increases the value of the data set by supplementing the available tasks with an additional entry, that of content-based image retrieval. We generate a selection of query images which are held out from the image list. We consider two variations of the retrieval process: easy and hard. For the easy track, we crop the bounding boxes of random objects and use the cropped objects as query images, marking as positive any image that contains the searched object class. The difficult track focuses on multi-object queries (illustrating certain activities, e.g. person riding a bike, cats playing), marking as positive any image that contains all of the object classes seen in the query image. We generate 350 single-object queries and 350 multi-object queries for evaluation purposes. For supervised predictors, we generate an equally large and identically balanced set of queries (700) for training.

Caltech-101. Caltech-101 [41] is an object recognition data set, which we repurpose to address the CBIR task. The data set consists of images depicting objects from one of 100 object classes. Additional images are provided to represent a visual clutter (background) class. The data set contains 9,146 images. We keep 700 images as training queries and another 700 images as test queries.

ROxford5k. ROxford5k [49] is a version of the Oxford5k [47] data set curated by Radenović et al. [49]. Oxford5k is a popular object retrieval data set that contains images depicting landmarks from the city of Oxford. The curation consists of annotation corrections, new queries and multiple difficulty tracks. ROxford5k is composed of 5,063 images, out of which 70 are held out as queries. The data samples are split into four different categories based on how clearly they depict the subject: easy, hard, negative and unclear. Images labeled as easy have minor viewpoint changes and illumination conditions similar to the query, while images labeled as hard have more difficult viewing conditions. Unclear images depict landmarks that cannot be accurately identified without contextual information. Since ROxford5k [49] comes with only 70 queries, we propose to evaluate supervised predictors using a 5-fold cross-validation procedure. We publicly release the folds with the benchmark.

RParis6k. RParis6k [49] is an enhancement of Paris6k [48] consisting of 6,392 images and 70 queries. The data set follows an identical structure to ROxford5k, since it is also curated by Radenović et al. [49]. RParis6k contains various landmarks from the city of Paris from multiple viewing points and illumination conditions. To evaluate supervised query performance predictors, we employ 5-fold cross-validation. As for ROxford5k, we make the folds public to facilitate future comparisons.

6.2 Evaluation protocol

To assess the difficulty level of a query, we consider two alternative measures of retrieval effectiveness, namely the average precision (AP) and the precision for the top k retrieved results (P@k). The precision@k is given by the number of true positive images divided by k. The recall is given by the ratio between the number of true positive images and the number of images labeled as positive for the query. The AP is given by the area under the precision-recall curve, which takes into account all possible thresholds k. Although P@10 is sometimes used in text QPP [75], we found that a high percentage of the test queries (between 29% and 82%, depending on the data set) have a P@10 score of 1. For a better estimation of query difficulty, we decided to use P@100.

To estimate the performance level of a predictor, we employ the Pearson and Kendall τ correlation coefficients between the predicted and the actual effectiveness levels of all test queries, following the conventional evaluation procedure in text QPP [10, 22, 75, 79]. Moreover, we apply a Student's t-test at a confidence score of 0.01 to test significance [54]. Although some data sets separate queries into different difficulty tracks, we aim to evaluate the capacity of predictors to estimate the actual AP or P@100 scores rather than classifying the queries as easy or hard, since we consider that the regression task is more suitable for revealing the true abilities of the predictors.

6.3 Image retrieval models

The first image retrieval system used in our benchmark was proposed by Radenović et al. [50][2]. The model is based on fine-tuning a convolutional neural network on a large set of annotation-free images. The authors leverage the use of geometry and camera

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

Table 2: Pearson and Kendall τ correlations of the query performance predictors on PASCAL VOC 2012 and Caltech-101, which contain images of various object classes. Underlined results are significantly better than the random chance baseline, according to a Student's t-test with a p-value lower than 0.01. The top pre-retrieval and post-retrieval scores are highlighted in bold.

Type	Supervised	Method	PASCAL VOC 2012								Caltech-101							
			Radenović et al. [50]				Revaud et al. [53]				Radenović et al. [50]				Revaud et al. [53]			
			AP		P@100		AP		P@100		AP		P@100		AP		P@100	
			Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
		Random	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pre-retrieval		#objects / area [67]	0.02	**0.22**	0.03	**0.25**	0.02	**0.27**	0.03	**0.25**	0.01	0.08	0.01	0.04	0.04	0.06	0.03	0.04
		Image difficulty [30]	**0.25**	0.19	**0.33**	0.23	**0.32**	0.24	**0.31**	0.22	−0.01	−0.02	−0.07	−0.07	0.00	−0.02	−0.07	−0.06
		Denoising AE	0.15	0.16	0.06	0.08	0.11	0.12	0.08	0.09	0.03	0.02	0.06	0.03	0.12	0.07	0.13	0.07
		Masked AE	0.11	0.11	0.01	0.05	0.01	0.06	−0.01	0.03	−0.04	−0.04	0.01	0.00	0.03	0.02	0.09	0.05
		Class head kurtosis	0.05	0.08	0.09	0.07	0.12	0.09	0.12	0.08	0.16	0.17	0.26	0.30	0.23	0.17	0.13	0.10
		Class head dispersion	0.08	0.09	0.13	0.08	0.17	0.11	0.17	0.10	0.25	0.20	**0.48**	**0.38**	0.32	0.23	0.21	0.15
		Cluster density	0.13	0.12	0.00	0.01	−0.02	−0.04	−0.01	−0.01	0.15	0.09	0.41	0.24	−0.13	0.09	−0.03	−0.4
	✓	Fine-tuned ViT	0.04	0.02	0.20	0.10	0.17	0.06	0.14	0.05	**0.54**	**0.38**	0.27	0.15	**0.65**	**0.47**	**0.41**	**0.20**
Post-retrieval		Score Variance [13]	0.02	0.05	−0.02	0.02	0.23	0.19	0.26	0.20	0.11	0.01	0.21	0.01	0.51	0.51	0.30	0.39
	✓	Correlation CNN [68]	0.27	0.07	0.32	0.16	0.32	0.15	0.26	0.11	**0.83**	**0.65**	**0.76**	**0.51**	**0.78**	**0.60**	**0.71**	**0.50**
		Adapted query feedback	0.23	0.16	0.37	0.21	0.41	0.26	0.41	0.24	0.60	0.43	0.60	0.46	0.56	0.40	0.60	0.44
		Iterative removal	0.16	0.13	0.35	0.20	0.41	0.26	0.40	0.23	0.57	0.41	0.57	0.42	0.31	0.20	0.40	0.23
		Embedding Variance	0.29	0.20	0.33	0.21	0.43	0.22	0.37	0.20	0.28	0.20	0.49	0.28	0.26	0.18	0.49	0.26
	✓	Meta-regressor	**0.36**	**0.28**	**0.45**	**0.29**	**0.51**	**0.34**	**0.48**	**0.30**	0.71	0.53	0.72	**0.51**	0.76	0.57	0.70	0.49

positioning of 3D models returned by a structure-from-motion framework to guide the selection of matching and non-matching image pairs, eliminating the need for manually annotated data. The proposed architecture employs a novel pooling layer with trainable parameters that induce a particular case of the generalized mean (GeM).

The second image retrieval system from iQPP was introduced by Revaud et al. [53][3]. The authors apply a histogram binning approximation to make the AP differentiable, enabling its use as a loss function for training deep networks. The system uses a ResNet-101 [29] backbone pre-trained on ImageNet [59]. Following Radenović et al. [50], a GeM pooling layer is integrated into the architecture.

7 BENCHMARK RESULTS

We group our experiments based on the type of images from the chosen data sets. Hence, in Section 7.1, we present the results on data sets composed of images of various natural or man-made objects, namely PASCAL VOC 2012 and Caltech-101. In Section 7.2, we discuss the results on ROxford5k and RParis6k, as both data sets contain landmark images. Finally, we make a few observations about the overall results in Section 7.3.

7.1 Results on PASCAL VOC and Caltech-101

We report the results on PASCAL VOC 2012 and Caltech-101 in Table 2. We discuss the reported results below, making several interesting observations.

[3]https://github.com/naver/deep-image-retrieval

Results of pre-retrieval predictors. On PASCAL VOC, the best pre-retrieval predictors are the baselines based on image difficulty or the number of objects divided by their area (see the first two rows after the random baseline in Table 2). These pre-retrieval predictors mainly rely on the presence of multiple objects from various object categories inside the query image, e.g. people, cars or dogs, being suitable for PASCAL VOC queries. However, the two baseline predictors exhibit poor performance on Caltech-101 images, which typically contain only one object per image. Interestingly, on PASCAL VOC, image difficulty gives a better Pearson correlation, while the number of objects divided by their area gives a better Kendall τ. This observation points towards the importance of using multiple correlation measures to better assess predictors' behavior.

On Caltech-101, the best pre-retrieval predictor for the AP measure is the fine-tuned ViT. However, when we consider P@100 as the ground-truth query performance, the fine-tuned ViT is surpassed by the classification head dispersion in one scenario. Some predictors seem to be better suited for certain effectiveness measures. To find robust predictors across effectiveness measures, it is thus important to include more than a single measure to estimate ground-truth query performance. Finally, we observe that the two data sets have distinct top scoring pre-retrieval predictors, demonstrating that it is not sufficient to use a single data set to find generic predictors.

Results of post-retrieval predictors. The baseline predictor based on score variance seems to be a suitable estimator for the system of Revaud et al. [53]. However, the predictions given by score variance for the system of Radenović et al. [50] are close to random chance. These results demonstrate that predictors can depend on

Table 3: Pearson and Kendall τ correlations of the query performance predictors on ROxford5k and RParis6k, which contain landmark images. Underlined results are significantly better than the random chance baseline, according to a Student's t-test with a p-value lower than 0.01. The top pre-retrieval and post-retrieval scores are highlighted in bold.

Type	Supervised	Method	ROxford5k								RParis6k							
			Radenović et al. [50]				Revaud et al. [53]				Radenović et al. [50]				Revaud et al. [53]			
			AP		P@100		AP		P@100		AP		P@100		AP		P@100	
			Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
		Random	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.02	0.01
Pre-retrieval		#objects / area [67]	0.17	0.05	0.14	0.03	0.16	0.02	−0.05	0.04	0.01	−0.01	−0.17	−0.16	0.00	−0.07	−0.12	−0.19
		Image difficulty [30]	−0.04	−0.04	−0.18	−0.11	−0.11	−0.10	0.12	0.09	−0.06	−0.03	−0.04	0.05	−0.19	−0.06	−0.09	−0.01
		Denoising AE	−0.07	−0.02	−0.04	0.06	−0.07	−0.03	0.06	0.09	−0.18	−0.20	0.31	0.19	0.02	0.00	**0.45**	0.28
		Masked AE	0.10	0.07	0.22	0.20	−0.09	0.07	0.11	0.11	−0.21	−0.20	0.18	0.14	0.10	0.02	0.40	0.26
		Class head kurtosis	0.27	0.20	0.20	0.18	0.28	0.18	0.00	0.04	0.19	0.18	−0.01	−0.09	0.18	0.00	0.04	−0.07
		Class head dispersion	0.34	0.24	0.27	**0.23**	0.35	0.21	0.12	0.10	**0.34**	**0.24**	0.27	**0.23**	**0.35**	0.21	0.12	0.10
		Cluster density	0.22	0.02	0.38	0.14	0.01	−0.05	0.23	0.01	0.32	0.22	**0.36**	0.21	0.20	0.17	0.23	0.17
	✓	Fine-tuned ViT	**0.41**	**0.31**	**0.40**	0.21	**0.43**	**0.28**	**0.36**	**0.22**	0.18	0.21	0.23	0.17	0.25	**0.24**	0.38	**0.38**
Post-retrieval		Score variance [13]	0.01	−0.02	0.21	0.22	0.27	0.28	0.41	0.35	0.41	0.17	0.45	0.35	−0.02	0.02	−0.38	−0.12
	✓	Correlation CNN [68]	0.43	0.30	**0.69**	0.49	**0.60**	0.44	**0.90**	0.62	0.65	0.33	**0.77**	0.55	**0.56**	**0.46**	**0.66**	0.37
		Adapted query feedback	0.22	0.11	0.36	0.24	0.42	0.29	0.77	0.52	0.33	0.13	0.56	0.32	0.32	0.16	0.32	0.26
		Iterative removal	0.28	0.22	0.31	0.33	0.36	0.23	0.75	0.53	0.30	0.17	0.40	0.19	0.35	0.14	0.50	0.32
		Embedding variance	0.30	0.15	0.54	0.32	0.34	0.23	0.84	0.57	0.51	0.27	0.69	0.45	0.36	0.18	0.47	0.40
	✓	Meta-regressor	**0.49**	**0.37**	0.62	**0.51**	0.58	**0.45**	0.88	**0.65**	**0.69**	**0.51**	0.76	**0.60**	0.37	0.37	0.65	**0.56**

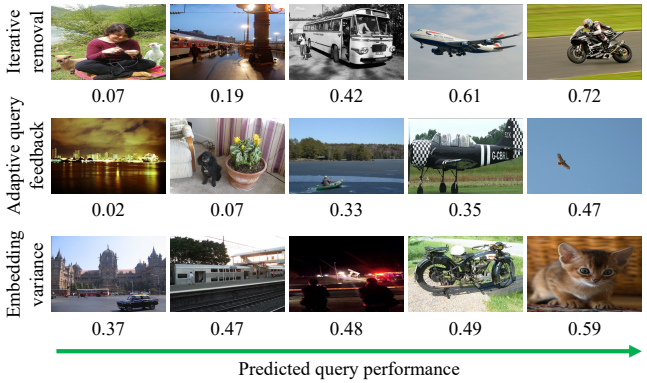


Figure 2: Image queries from PASCAL VOC 2012 displayed in increasing order of predicted performance, from left to right. Examples are illustrated for three of the best unsupervised predictors: iterative feature removal (top row), adaptive query feedback (middle row), and embedding variance (bottom row). Best viewed in color.

the reference system. By including multiple retrieval systems in our benchmark, we are able to identify predictors that are inconsistent across different retrieval models.

On PASCAL VOC 2012, our largest data set, the meta-regressor outperforms all competing predictors, leveraging the use of information from the other predictors to surpass them. However, this does not happen on Caltech-101, where the best predictor is the correlation-based CNN. Regardless of the data set, it is clear that supervised post-retrieval predictors are generally better, surpassing the unsupervised post-retrieval predictors. Another expected outcome is that post-retrieval predictors obtain superior results compared with pre-retrieval predictors.

The proposed unsupervised post-retrieval predictors (adapted query feedback, iterative feature removal and embedding variance) reach reasonably good correlation levels, always surpassing the random predictor baseline by statistically significant margins. To further analyze the behavior of these predictors, we illustrate some randomly chosen queries from PASCAL VOC 2012 in Figure 2, organizing them in increasing order of predicted performance. The figure shows that all three predictors find query images with fewer objects and plain background as more likely to exhibit high performance. In contrast, images with multiple objects, photographed in poor illumination conditions are associated with low performance levels. In summary, we find strong connections between the visual content of queries and the performance scores predicted by the unsupervised post-retrieval predictors.

7.2 Results on ROxford5k and RParis6k

We report the results on ROxford5k and RParis6k in Table 3.

Results of pre-retrieval predictors. Since the ROxford5k and RParis6k data sets contain landmark images, the predictors based on image difficulty and the number of objects divided by their area obtain generally poor performance. This happens because the images contain landmarks, such as the Eiffel Tower, rather than objects, such as dogs and horses.

The denoising and masked auto-encoders are better correlated with P@100 than with the AP measure, likely because P@100 associates higher penalties to queries with very few positive results. Since these queries likely reside in a sparse area of the data distribution (due to the low number of similar images), AE models are unable to reconstruct the corresponding queries very well. This observation shows the importance of using more than one ground-truth query performance metric to find predictors that are robust across multiple target performance measures, supporting our decision to use both the AP and P@100 measures for our benchmark.

Among the pre-retrieval predictors, the fine-tuned ViT obtains the best results in most cases. This is rather unsurprising, since ViT is a supervised predictor, while all the other pre-retrieval predictors are unsupervised. Interestingly, the ViT model is often challenged by the classification head dispersion. We thus consider the latter model as the best unsupervised pre-retrieval predictor on the ROxford5k and RParis6k data sets.

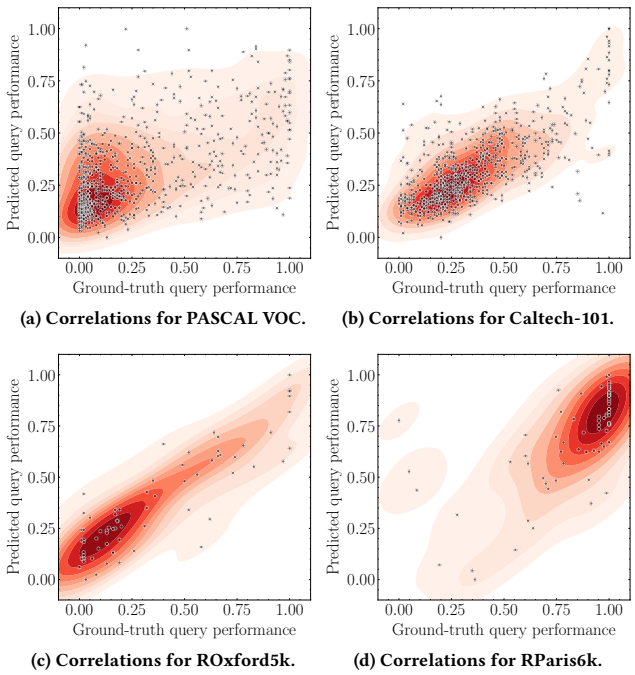
Results of post-retrieval predictors. The post-retrieval predictor based on score variance obtains inconsistent results, being the weakest post-retrieval predictor. Except for the score variance, the post-retrieval predictors surpass the pre-retrieval ones in the majority of cases. Comparing the unsupervised post-retrieval predictors among each other, we observe that embedding variance provides the best scores, generally surpassing the iterative feature removal and the adapted query feedback, respectively.

The best post-retrieval predictors are the supervised ones, namely the correlation-based CNN and the meta-regressor. In 9 out 16 cases, the meta-regressor obtains the best correlations. In the other 7 cases, the correlation-based CNN outperforms all competing predictors. Interestingly, we observe that there are 5 situations where the correlation-based CNN gives a better Pearson correlation than the meta-regressor, while the meta-regressor surpasses the correlation-based CNN in terms of Kendall τ. This shows the importance of using multiple measures to evaluate QPP methods, indicating that our decision to consider both Pearson and Kendall τ for iQPP is useful in finding predictors that are consistent across QPP evaluation measures.

7.3 Generic discussion

Our empirical results reveal that many predictors are only suitable for certain data sets (for example, image difficulty for PASCAL VOC), ground-truth measures (for example, auto-encoders for P@100), retrieval systems (for example, score variance for the system of Revaud et al. [53]) or correlation coefficients (for example, the correlation-based CNN for the Pearson correlation). Hence, the results demonstrate the importance of building a comprehensive benchmark based on multiple data sets, retrieval systems and metrics, to establish the generalization capacity of predictors and their robustness to variations of the above components.

The iQPP benchmark includes a high variety of scenarios, being difficult to find a single predictor that is consistently better over all scenarios. However, there are some generic observable trends. First, we notice that post-retrieval predictors generally obtain higher correlations than pre-retrieval predictors. Second, we observe that supervised predictors generally outperform the unsupervised ones. The meta-regressor appears to be the best predictor, being closely



(a) Correlations for PASCAL VOC. (b) Correlations for Caltech-101.

(c) Correlations for ROxford5k. (d) Correlations for RParis6k.

Figure 3: Correlation plots between the ground-truth performance given by the P@100 measure for the system of Revaud et al. [53], and the query performance predicted by the meta-regressor. The intensity of red is proportional to the density of points in the corresponding region.

followed by the correlation-based CNN, confirming the observed trends. Since high Pearson or Kendall τ correlation scores can be sometimes misleading, we illustrate the correlation plots between the ground-truth query performance and the performance predicted by the meta-regressor in Figure 3. We observe that the plots correspond to the reported numbers, i.e. the higher correlations on Caltech-101, ROxford5k and RParis6k, and the lower correlation on PASCAL VOC 2012 are visually confirmed by the plots in Figure 3.

8 CONCLUSION

In this paper, we introduced the first benchmark for image QPP, comprising four data sets, two retrieval systems and twelve query performance predictors. We studied a wide variety of query performance predictors for CBIR, including state-of-the-art methods [30, 67, 68], adaptations of text QPP methods [13], as well as novel proposals. Our benchmark shows that the problem of QPP in image search is still open, as none of the predictors obtained high performance across all data sets and retrieval methods. The empirical results show that our new benchmark exhibits a high variety of evaluation scenarios, representing a real challenge for current and future work on QPP. We thus envision our benchmark as a stepping stone for future research on QPP in the image domain.

In future work, we aim to increase the value of our benchmark by expanding the pool of data sets and retrieval methods. Furthermore, we aim to study novel predictors by leveraging the insights gained from our current experiments. We also plan to delve deeper into the analysis of queries, and separately study easy and hard queries.

REFERENCES

[1] Negar Arabzadeh, Amin Bigdeli, Morteza Zihayat, and Ebrahim Bagheri. 2021. Query Performance Prediction Through Retrieval Coherency. In *Proceedings of ECIR 2021*. 193–200.

[2] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. BERT-QPP: Contextualized Pre-trained transformers for Query Performance Prediction. In *Proceedings of CIKM*. 2857–2861.

[3] Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, Feras Al-Obeidat, and Ebrahim Bagheri. 2020. Neural embedding-based specificity metrics for pre-retrieval query performance prediction. *Information Processing & Management* 57, 4 (2020), 102248.

[4] Javed A. Aslam and Virgil Pavlu. 2007. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Proceedings of ECIR*. Springer, 198–209.

[5] Shariq Bashir. 2014. Combining pre-retrieval query quality predictors using genetic programming. *Applied intelligence* 40 (2014), 525–535.

[6] Anthony Bigot, Claude Chrisment, Taoufiq Dkaki, Gilles Hubert, and Josiane Mothe. 2011. Fusing different information retrieval systems according to query-topics: a study based on correlation in information retrieval systems and TREC topics. *Information Retrieval* 14, 6 (2011), 617.

[7] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. 2006. What makes a query difficult?. In *Proceedings of SIGIR*. 390–397.

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep Clustering for Unsupervised Learning of Visual Features. In *Proceedings of ECCV*, Vol. 11218. 139–156.

[9] Xiaoyang Chen, Ben He, and Le Sun. 2022. Groupwise Query Performance Prediction with BERT. In *Proceedings of ECIR*. 64–74.

[10] Adrian-Gabriel Chifu, Léa Laporte, Josiane Mothe, and Md Zia Ullah. 2018. Query Performance Prediction Focused on Summarized Letor Features. In *Proceedings of SIGIR*. 1177–1180.

[11] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proceedings of SIGIR*. 299–306.

[12] Ronan Cummins. 2014. Document score distribution models for query performance inference and prediction. *ACM Transactions on Information Systems* 32, 1 (2014), 2.

[13] Ronan Cummins, Joemon Jose, and Colm O'Riordan. 2011. Improved query performance prediction using standard deviation. In *Proceedings of SIGIR*. 1089–1090.

[14] Suchana Datta, Debasis Ganguly, Derek Greene, and Mandar Mitra. 2022. Deep-QPP: A Pairwise Interaction-based Deep Learning Model for Supervised Query Performance Prediction. In *Proceedings of WSDM*. 201–209.

[15] Suchana Datta, Debasis Ganguly, Mandar Mitra, and Derek Greene. 2022. A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants. *ACM Transactions on Information Systems* 41, 2 (2022), 1–31.

[16] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. A'Pointwise-Query, Listwise-Document'based Query Performance Prediction Approach. In *Proceedings of SIGIR*. 2148–2153.

[17] Claude De Loupy and Patrice Bellot. 2000. Evaluation of document retrieval systems and query difficulty. In *Proceedings of LREC*. 32–39.

[18] Sébastien Déjean, Radu Tudor Ionescu, Josiane Mothe, and Md Zia Ullah. 2020. Forward and backward feature selection for query performance prediction. In *Proceedings of SAC*. 690–697.

[19] Bekir Taner Dinçer. 2007. Statistical Principal Components Analysis for Retrieval Experiments. *Journal of the American Society for Information Science and Technology* 58, 4 (2007), 560–574.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of ICLR*.

[21] Mark Everingham, S.M. Ali Eslami, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. 2015. The PASCAL Visual Object Classes Challenge: A Retrospective. *International journal of computer vision* 111 (2015), 98–136.

[22] Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. 2022. sMARE: a new paradigm to evaluate and understand query performance prediction methods. *Information Retrieval Journal* 25, 2 (2022), 94–122.

[23] Lorraine Goeuriot, Liadh Kelly, and Johannes Leveling. 2014. An analysis of query difficulty for information retrieval in the medical domain. In *Proceedings of SIGIR*. 1007–1010.

[24] Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. 2009. The combination and evaluation of query performance prediction methods. In *Proceedings of ECIR*. 301–312.

[25] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proceedings of CIKM*. 1419–1420.

[26] Ben He and Iadh Ounis. 2004. Inferring query performance using pre-retrieval predictors. In *Proceedings of SPIRE*. 43–54.

[27] Ben He and Iadh Ounis. 2006. Query performance prediction. *Information Systems* 31, 7 (2006), 585–594.

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of CVPR*. 16000–16009.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*. 770–778.

[30] Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P. Papadopoulos, and Vittorio Ferrari. 2016. How Hard Can It Be? Estimating the Difficulty of Visual Search in an Image. In *Proceedings of CVPR*. 2157–2166.

[31] Radu Tudor Ionescu and Marius Popescu. 2016. Object recognition with the bag of visual words model. *Knowledge Transfer between Computer Vision and Text Mining* (2016), 99–132.

[32] Parastoo Jafarzadeh and Faezeh Ensan. 2022. A semantic approach to post-retrieval query performance prediction. *Information Processing & Management* 59, 1 (2022), 102746.

[33] Qianghuai Jia and Xinmei Tian. 2015. Query difficulty estimation via relevance prediction for image retrieval. *Signal Processing* 110 (2015), 232–243.

[34] Qianghuai Jia, Xinmei Tian, and Tao Mei. 2014. Query difficulty estimation via pseudo relevance feedback for image search. In *Proceedings of ICME*. 1–6.

[35] Gilad Katz, Anna Shtock, Oren Kurland, Bracha Shapira, and Lior Rokach. 2014. Wikipedia-based query performance prediction. In *Proceedings of SIGIR*. 1235–1238.

[36] Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic gradient descent. In *Proceedings of ICLR*.

[37] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research* 8, 6 (2007), 1261–1276.

[38] Oren Kurland. 2014. The cluster hypothesis in information retrieval. In *Proceedings of ECIR*. 823–826.

[39] Oren Kurland, Fiana Raiber, and Anna Shtok. 2012. Query-performance prediction and cluster ranking: Two sides of the same coin. In *Proceedings of CIKM*. 2459–2462.

[40] Oren Kurland, Anna Shtok, Shay Hummel, Fiana Raiber, David Carmel, and Ofri Rom. 2012. Back to the roots: A probabilistic framework for query-performance prediction. In *Proceedings of CIKM*. 823–832.

[41] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. 2022. Caltech 101. https://doi.org/10.22002/D1.20086

[42] Yangxi Li, Bo Geng, Linjun Yang, Chao Xu, and Wei Bian. 2012. Query difficulty estimation for image retrieval. *Neurocomputing* 95 (2012), 48–53.

[43] Stefano Mizzaro, Josiane Mothe, Kevin Roitero, and Md Zia Ullah. 2018. Query performance prediction and effectiveness evaluation without relevance judgments: Two sides of the same coin. In *Proceedings of SIGIR*. 1233–1236.

[44] Josiane Mothe and Ludovic Tanguy. 2005. Linguistic features to predict query difficulty. In *Proceedings of SIGIR*. 7–10.

[45] Liqiang Nie, Meng Wang, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Oracle in image search: a content-based approach to performance prediction. *ACM Transactions on Information Systems* 30, 2 (2012), 1–23.

[46] Daniel Carlos Guimarães Pedronette and Ricardo da S. Torres. 2015. Unsupervised effectiveness estimation for image retrieval using reciprocal rank information. In *Proceedings of SIBGRAPI*. 321–328.

[47] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of CVPR*. 1–8.

[48] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *Proceedings of CVPR*.

[49] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *Proceedings of CVPR*. 5706–5715.

[50] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 7 (2019), 1655–1668.

[51] Fiana Raiber and Oren Kurland. 2014. Query-performance prediction: setting the expectations straight. In *Proceedings of SIGIR*. 13–22.

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of NIPS*. 91–99.

[53] Jerome Revaud, Jon Almazán, Rafael S. Rezende, and Cesar Roberto de Souza. 2019. Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In *Proceedings of ICCV*. 5107–5116.

[54] Haggai Roitman. 2018. An extended query performance prediction framework utilizing passage-level information. In *Proceedings of SIGIR*. 35–42.

[55] Haggai Roitman, Shai Erera, Oren Sar-Shalom, and Bar Weiner. 2017. Enhanced mean retrieval score estimation for query performance prediction. In *Proceedings of SIGIR*. 35–42.

[56] Haggai Roitman, Shai Erera, and Bar Weiner. 2017. Robust standard deviation estimation for query performance prediction. In *Proceedings of SIGIR*. 245–248.

[57] Haggai Roitman and Oren Kurland. 2019. Query performance prediction for pseudo-feedback-based retrieval. In *Proceedings of SIGIR*. 1261–1264.

[58] Dwaipayan Roy, Debasis Ganguly, Mandar Mitra, and Gareth J.F. Jones. 2019. Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. *Information Processing & Management* 56, 3 (2019), 1026–1045.

[59] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115 (2015), 211–252.

[60] Ghulam Sarwar and Colm O'Riordan. 2021. Passage Based Answer-Set Graph Approach for Query Performance Prediction. In *Proceedings of ADCS*. 1–6.

[61] Falk Scholer, Hugh E. Williams, and Andrew Turpin. 2004. Query association surrogates for web search. *Journal of the American Society for Information Science and Technology* 55, 7 (2004), 637–650.

[62] Aditya Kumar Sehgal and Padmini Srinivasan. 2005. Predicting performance for gene queries. In *Proceedings of Workshop on Predicting Query Difficulty–Methods and Applications*. 1–3.

[63] Anna Shtok, Oren Kurland, and David Carmel. 2010. Using statistical decision theory and relevance models for query-performance prediction. In *Proceedings of SIGIR*. 259–266.

[64] Anna Shtok, Oren Kurland, and David Carmel. 2016. Query performance prediction using reference lists. *ACM Transactions on Information Systems* 34, 4 (2016), 1–34.

[65] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems* 30, 2 (2012), 1–35.

[66] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of ICLR*.

[67] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2021. Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding* 204 (2021), 103–166.

[68] Shaoyan Sun, Wengang Zhou, Qi Tian, Ming Yang, and Houqiang Li. 2018. Assessing image retrieval quality at the first glance. *IEEE Transactions on Image Processing* 27, 12 (2018), 6124–6134.

[69] Yongquan Tao and Shengli Wu. 2014. Query performance prediction by considering score magnitude and variance together. In *Proceedings of CIKM*. 1891–1894.

[70] Xinmei Tian, Qianghuai Jia, and Tao Mei. 2015. Query difficulty estimation for image search with query reconstruction error. *IEEE Transactions on Multimedia* 17, 1 (2015), 79–91.

[71] Xinmei Tian, Yijuan Lu, and Linjun Yang. 2012. Query difficulty prediction for Web image search. *IEEE Transactions on Multimedia* 14, 4 (2012), 951–962.

[72] Lucas Pascotti Valem and Daniel Carlos Guimarães Pedronette. 2021. A denoising convolutional neural network for self-supervised rank effectiveness estimation on image retrieval. In *Proceedings of ICMR*. 294–302.

[73] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research* 11 (2010), 3371–3408.

[74] Xing Xing, Yi Zhang, and Mei Han. 2010. Query difficulty prediction for contextual image retrieval. In *Proceedings of ECIR*. 581–585.

[75] Elad Yom-Tov, Shai Fine, David Carmel, and Adam Darlow. 2005. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proceedings of SIGIR*. 512–519.

[76] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural query performance prediction using weak supervision from multiple signals. In *Proceedings of SIGIR*. 105–114.

[77] Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpepper. 2019. Information needs, queries, and query performance prediction. In *Proceedings of SIGIR*. 395–404.

[78] Zhongmin Zhang, Jiawei Chen, and Shengli Wu. 2018. Query performance prediction and classification for information search systems. In *Proceedings of APWeb-WAIM*. Springer, 277–285.

[79] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proceedings of ECIR*. Springer, 52–64.

[80] Yun Zhou and W. Bruce Croft. 2006. Ranking robustness: a novel framework to predict query performance. In *Proceedings of CIKM*. 567–574.

[81] Yun Zhou and W. Bruce Croft. 2007. Query Performance Prediction in Web Search Environments. In *Proceedings of SIGIR*. 543–550.