Confidence-Aware Active Feedback for Efficient Instance Search

Yue Zhang, Chao Liang, and Longxiang Jiang

Abstract—Relevance feedback is widely used in instance search (INS) tasks to further refine imperfect ranking results, but it often comes with low interaction efficiency. Active learning (AL) technique has achieved great success in improving annotation efficiency in classification tasks. However, considering irrelevant samples’ diversity and class imbalance in INS tasks, existing AL methods cannot always select the most suitable feedback candidates for INS problems. In addition, they are often too computationally complex to be applied in interactive INS scenario. To address the above problems, we propose a confidence-aware active feedback (CAAF) method that can efficiently select the most valuable feedback candidates to improve the re-ranking performance. Specifically, inspired by the explicit sample difficulty modeling in self-paced learning, we utilize a pairwise manifold ranking loss to evaluate the ranking confidence of each unlabeled sample, and formulate the INS process as a confidence-weighted manifold ranking problem. Furthermore, we introduce an approximate optimization scheme to simplify the solution from QP problems with constraints to closed-form expressions, and selects only the top-K samples in the initial ranking list for INS, so that CAAF is able to handle large-scale INS tasks in a short period of time. Extensive experiments on both image and video INS tasks demonstrate the effectiveness of the proposed CAAF method. In particular, CAAF outperforms the first-place record in the public large-scale video INS evaluation of TRECVID 2021.

Index Terms—Active Learning, Interactive Instance Search.

1 INTRODUCTION

With the explosive development of media technology, it is necessary to study how to retrieve information over a large collection of multimedia data. Unlike class retrieval tasks that focus on searching objects of certain categories, instance search (INS) tasks are much more challenging, as they aim at finding the specific object from the whole of its kind [1]. e.g. the former merely need to find buildings from various kinds of objects, while the latter have to find the Big Ben from all kinds of buildings. In the past few years, automatic INS methods have achieved great success in obtaining discriminative feature embeddings [2], [3], [4] or accurate distance metrics [5], [6]. However, due to environmental interference or individual differences [7], fully automatic search methods cannot always generate satisfying results.

Relevance feedback (RF) [8] is often used to refine the existing imperfect ranking result by asking extra supervisory information through an interaction process with users. However, selecting proper feedback candidates is usually a burdensome task. To improve the interaction efficiency, active learning (AL) technique is studied to wisely select valuable feedback candidates, and it has achieved great success in classification problems [9], [10], [11]. But INS is intrinsically a ranking problem, where samples in the gallery are ranked according to their relevance with the probe, rather than a classification problem, where categories are explicitly distinguished and evenly distributed. As illustrated in Figure 1, the irrelevant samples in INS usually consist of multiple categories in the general classification problem, i.e., irrelevant samples’ diversity, and their amount is much more than that of relevant samples, i.e., class imbalance. As a result, the performance of classification-oriented AL methods may not be adequate on INS tasks. Besides, experimental results in Section 4.1 show that they are often too time-consuming to be applied in interactive INS tasks requiring real-time responses. Although there are some ranking-oriented AL methods, they are either designed to select an entire ranking list [12], or based on analogous strategies adopted by classification-oriented methods [13], [14]. Therefore, ranking-oriented methods can be faced with similar problems as classification-oriented ones sometimes.

In order to address the above problem, we propose a confidence-aware active feedback (CAAF) method that

1. All figures in this paper are recommended to be viewed in color mode.
measures ranking confidence of each sample via manifold ranking (MR) [15]. From which, galleries’ ranking scores and confidence scores are mutually coupled in one objective function, and will be simultaneously computed in each round of feedback. Such an idea is inspired by the explicit sample difficulty modeling in self-paced learning (SPL) [16], [17]. But ours has two main differences: (1) CAAF calculates the confidence score with pairwise ranking loss rather than individual classification loss, which better adapts to the characteristics of the ranking problem; (2) CAAF aims at selecting the most valuable candidates for RF, instead of progressively training a classifier, hence difficult galleries with low confidence scores are preferentially selected.

The workflow of the proposed CAAF method is shown in Figure 2. Given a probe and a set of gallery samples, a confidence-weighted data manifold is first constructed. The nodes correspond to the probe and the galleries, and the edges encode their pairwise similarities, whose thickness represents the confidence weight. Based on which, CAAF iteratively optimizes the ranking quality and generates feedback suggestion. In order to solve this problem, we divide the whole process into a ranking step and a suggestion step. The former, ranking step, computes new ranking scores with user relevance feedbacks. It is formulated as a confidence-weighted manifold ranking problem, where samples having higher confidence scores are endowed with larger weights in diffusing their ranking scores to others. e.g. the blue node in Figure 2, which represents the probe, is considered as the only high confident sample in the beginning, hence all edges connected to it are thicker than the other edges. Such confidence-weighted modulation scheme not only effectively relieves model selection stress in the graph-based models [18], but also improves manifold ranking performance by highlighting reliable galleries, and meanwhile, suppressing unreliable ones. The latter, suggestion step, directly estimates galleries’ confidence scores by evaluating their pairwise ranking losses. From the viewpoint of AL [19], galleries with low confidence scores, corresponding to large ranking losses, are ideal feedback candidates for RF. Particularly, due to the merit of MR, the association between labeled and unlabeled samples can be fully considered in the ranking confidence computation process, making the suggestion step select both informative and representative samples [20]. Finally, both ranking results and feedback suggestions are returned to the user. If the user is not content with the ranking result, she/he can make RF of suggested samples and restart CAAF for a new trial.

Moreover, we design two optimization strategies to meet the real-time requirements of interactive INS tasks. On the one hand, we abandon the convergence condition of alternative optimization strategy [21], which is usually adopted in SPL, and approximately solve the quadratic-programming-based objective function by calculating the closed form expressions, i.e., both ranking step and suggestion step are successively taken only once in each round of feedback instead of iteratively solved until convergence, and the constraints in their objective functions are omitted. On the other hand, only top-K samples in the initial ranking result are considered in the INS process, hence the problem scale is mainly confined to K, a number that much smaller than the gallery size. Experimental results in Section 4.4 demonstrate that both strategies hardly affect the performance but significantly reduce the execution time.

The contributions of this paper can be summarized as follows:

- We propose a confidence-aware active learning scheme for interactive INS, whose novelty lies in an explicit modeling of the ranking confidence of each gallery sample, which supports more accurate manifold ranking as well as feedback suggestion.
• We design two optimization strategies to efficiently solve our proposed method, so that both refined ranking results and valuable feedback suggestions can be generated in a short period of time, even on large-scale datasets.
• We conduct extensive experiments with state-of-the-art AL methods on both image and video INS. Particularly, we surpass the official first-place record in the public large-scale video INS evaluation of TRECVID 2021.

The remainder of this paper is organized as follows: Section 2 gives a brief introduction of the related work, including AL and SPL. In Section 3, we describe the proposed CAAF method and its solution. Section 4 shows the experimental results, and conclusions are given in Section 5.

2 RELATED WORK
There are abundant literatures in INS, and authors recommend [1] for an extensive overview. In this section, we restrict our discussion to approaches that are directly related to CAAF, including AL methods for feedback selection, and the SPL framework for confidence modeling.

2.1 Active Learning
Existing AL methods can be roughly divided into three categories, including exploitation-driven methods aiming at informative samples, exploration-driven methods focusing on representative samples, and integrated methods seeking both informative and representative samples.

Exploitation-driven methods are the most popular AL approaches, most of which are implemented by either training a single learner and querying the unlabeled instance on which the learner has the least confidence [22], [23], [24], [25], [26], or generating multiple learners and querying the unlabeled instance on which the learners disagree the most [27], [28], [29]. Such sampling schemes enable them to find the most informative samples, but they tend to ignore the information in the large amount of unlabeled instances or mistakenly focus on the outliers [30]. There are some other exploitation-driven methods that aims to minimize the generalization error [31], [32], yet they are generally computationally expensive.

In contrast, exploration-driven methods [9], [29], [33], [34] consider the underlying data distribution of the unlabeled data, so they are able to find more representative samples from the unlabeled set. But since most of these samples cannot provide valuable information, many label requests are required to converge to a good solution [35].

In order to balance informativeness and representativeness, integrated methods [35], [36], [37], [38] are proposed to combine the above two kinds of strategy into one framework, and they have been proved to be more effective than solely exploitation-driven or exploration-driven ones in most cases [35].

Generally speaking, existing AL methods have been successfully applied to reduce the annotation cost in classification problems. But most of them are not compatible with INS tasks, as INS is intrinsically a ranking problem, whose data distribution is much more complex than that in common classification problems. On the one hand, the irrelevant class in INS is actually a combination of many categories in common classification problems; on the other hand, the amount of irrelevant samples are much more than that of relevant ones in INS. In such cases, it’s difficult to configure an accurate decision boundary to select the most informative samples, or find some representative dense regions for the whole irrelevant class in the context of INS. Besides, existing AL methods mainly concern how to train a robust model with as fewer annotations as possible and often ignore the high computational cost, hence they tend to be too time-consuming for interactive INS tasks.

There are some AL methods designed for ranking problems, yet the studies are mostly limited in the literature of document-based information retrieval. Unlike classification-oriented methods that focus on a single sample, ranking-oriented ones either sample on query level [12], document level [13], [39] or both [14], [40], [41]. The query level approaches aim to select the most valuable ranking result that consists of both the query and its whole ranking list, which are definitely not applicable on INS tasks. The document level approaches are very similar to the general ones as they focus on a single document, and they are usually based on analogous strategies mentioned above, e.g. minimizing the generalization error [14]. As a result, they can encounter similar problems as classification-oriented methods sometimes.

CAAF addresses the above issue by introducing a confidence score to explicitly measure galleries’ ranking reliability via pairwise manifold ranking loss. With the merit of MR, CAAF is able to jointly consider both labeled and unlabeled galleries, which supports it to select the most informative and representative feedback candidates. And the confidence score can, on the other way round, serve as the diffusion weight to modulate the manifold structure and improve the search performance.

2.2 Self-paced Learning
SPL [16], [17] techniques are originally studied in the training process of object classification models. It follows curriculum learning (CL) paradigm [42], which is inspired by the learning process of humans and animals that gradually incorporates easy to more complex samples into training. In order to quantitatively estimate the easiness of each gallery sample, SPL defines an objective function where samples with lower classification loss tend to have higher learning priority, which enables them to be preferentially selected for training.

Recently, some researchers begin to apply the idea of SPL into AL. Lin et al. [45] take SPL as a complement to AL, where high-confidence easy samples emphasized by SPL are endowed with pseudo labels, while low-confidence uncertain samples emphasized by SVM-active [22], a classic exploitation-driven method, are selected to acquire human annotations. Tang et al. [44] add a term that estimates the distribution difference between labeled and unlabeled data to the original SPL function, so that valuable samples can be gradually queried from easy to hard.

Unlike existing methods that merely consider easy samples indicated by SPL to train a robust classification model,
CAAF focuses on refining the imperfect ranking result by acquiring labels for low-confidence samples, which are considered to be difficult in SPL. Therefore, CAAF replaces the original classification loss in SPL with a more proper ranking loss, and preferentially selects low-confidence, i.e., difficult samples, with large ranking loss for RF.

3 Proposed Method

In this section, we first present the problem formulation with important notations, and then discuss the solution method.

3.1 Problem Formulation

Given a probe $p$ and a gallery set $G = \{g_i\}_{i=1}^n$, the goal is to refine the ranking score $f = [f_i]_{i=1}^m \in [0, 1]^{m \times 1}$ via human interaction, where $f_i$ is the ranking score of $x_i$. We first construct a new image set $X = G \cup \{p\} = \{x_i\}_{i=1}^m$, where $m = n + 1$. Then the pairwise data affinity can be represented as $W = [\max(w_{ij}, 0)]_{i,j=1}^{m \times m}$, where $w_{ij}$ reflects the similarity between $x_i$ and $x_j$. Apparently, $W$ is a symmetric matrix.

**Manifold ranking.** The classic manifold ranking [15] aims to obtain the ranking score $f$ by minimize a pairwise ranking loss function, with a smoothing term to constrain similar samples to have similar ranking scores, and a fitting term to make the ranking score $f$ not deviate too much from the reference ranking score $y = [0, \ldots, 0, 1]^\top \in [0, 1]^{m \times 1}$. The pairwise loss function between $x_i$ and $x_j$ is defined as

$$l_{ij} = w_{ij}(f_i - f_j)^2 + \alpha(f_i - y_i)^2 + \alpha(f_j - y_j)^2$$

(1)

where $\alpha \in (0, 1)$ is a parameter to balance the smoothing term and the fitting term. In order to adapt such a framework to our human-in-the-loop setting, we redefine $y_i = [y_i]_{i=1}^m$ as

$$y_i = \begin{cases} s_i, & \text{if } x_i \in \Psi \\ 0, & \text{otherwise} \end{cases}$$

(2)

where $s_i \in \{0, 1\}$ is the feedback score of $x_i$, and $\Psi$ is the set of all annotated samples. $s_i = 1$ if $x_i$ is relevant to the probe, and $s_i = 0$ otherwise. Particularly, the probe $x_m = p$ can be regarded as a specially annotated sample whose feedback score $s_m = 1$, and $f_i = y_i$ if $x_i \in \Psi$.

**Confidence modeling.** In the context of INS, the confidence of a sample can be defined as the reliability of its ranking score. We denote the confidence score as $v = [v_i]_{i=1}^m \in [0, 1]^{m \times 1}$, where $v_i$ is the confidence score of $x_i$. Apparently, the labeled samples have higher confidence than the unlabeled ones, hence $v_i$ is initialized as

$$v_i = \begin{cases} 1, & \text{if } x_i \in \Psi \\ 0, & \text{otherwise} \end{cases}$$

(3)

For unlabeled samples, however, $v_i$ is negatively correlated with the ranking loss. In other words, the higher ranking loss a sample generates, the lower confidence it has, and it tends to more valuable for RF from the viewpoint of AL. Therefore, $v$ is optimized by minimizing the following loss function

$$\mathcal{L}(f, v) = \frac{1}{m^2} \sum_{i,j}(v_i + v_j)(l_{ij} - \beta)$$

(4)

where $l_{ij}$ is the pairwise manifold ranking loss defined in Eq. 1, and $\beta > 0$ is a loss threshold to measure the confidence of each sample. Specifically, when $l_{ij}$ is smaller than $\beta$, minimizing $\mathcal{L}(f, v)$ will make both $v_i$ and $v_j$ approach 1; and on the contrary, when $l_{ij}$ is larger than $\beta$, minimizing $\mathcal{L}(f, v)$ will make both $v_i$ and $v_j$ approach 0.

By optimizing Eq. 4, $v$ works in two aspects:

- **Feedback generation.** When taking $v$ itself as the optimization objective, it is able to indicate the ranking confidence of each unlabeled sample, from which most valuable feedback suggestions can be generated.

- **Weight modulation.** When taking $f$ as the optimization objective, $v$ becomes an additional weight acting on the manifold ranking loss that increases the impact of high-confidence samples and reduces the influence of the low-confidence ones.

Furthermore, to constrain the element value of $v$, a squared norm regularization term is added to Eq. (4), which is defined as

$$R(v) = \frac{\gamma}{m} \|v\|_2^2$$

(5)

where $\gamma > 0$ is the regularization weight parameter. And finally, we have the following constraint optimization problem

$$\min_{f, v} \mathcal{E}(f, v) = \mathcal{L}(f, v) + R(v)$$

s.t. $0 \preceq f \preceq 1$, $f_i = y_i$ if $x_i \in \Psi$

(6)

$$0 \preceq v \preceq 1, v_i = 1 \text{ if } x_i \in \Psi$$

where $0$ and $1$ represent full 0 and 1 vectors, and $\preceq$ denotes element-wise comparison of $\leq$.

3.2 Solution

In order to solve $f$ and $v$ respectively, we adopt the alternative optimization strategy [21] and decompose Eq. 6 into a ranking step and a suggestion step.

**Ranking step.** In ranking step, we optimize $f$ with fixed $v$, thus Eq. (6) can be simplified by eliminating regularization term $R(v)$ and threshold $\beta$. The objective function is then computed as

$$\mathcal{E}_v(f) = \frac{1}{m} \sum_{i,j} \tilde{v}_{ij} l_{ij}$$

(7)

where $\tilde{v}_{ij} = v_i + v_j$ and the constant coefficient $\frac{1}{m}$ in Eq. (4) is replaced by $\frac{1}{m}$ to simplify the following transformation. By replacing $l_{ij}$ with Eq. (1) and eliminating terms irrelevant with $f$, Eq. (7) can be further rewritten as a confidence-weighted manifold ranking problem

$$\min_{f} \mathcal{E}_v(f) = f^\top (P + Q) f - 2f^\top Q y$$

(8)

where $P$ is the Laplacian matrix of the confidence-weighted affinity graph where the weight of the $(i,j)$-th edge is $\tilde{v}_{ij} w_{ij}$, $Q$ is a diagonal matrix where $q_{ii} = \sum_j \tilde{v}_{ij}$ (see Appendix A for detailed derivation). Since both $P$ and $Q$ are positive semi-definite matrices, Eq. (8) is a standard convex quadratic programming (QP) problem [45].
**Suggestion step.** In suggestion step, $f$ is fixed while $v$ is optimized. The related objective function can be concisely re-written in a matrix form as

$$\begin{align*}
\min_v \ E_f(v) &= \frac{\gamma}{m} v^\top I v + \frac{2}{m^2} l^\top v \\
\text{s.t.} \ 0 \leq v \leq 1, \ v_i = 1 \text{ if } x_i \in \Psi
\end{align*}$$

where $I$ is an $m$-by-$m$ unit matrix, and $l = [l_i]_{i=1}^m \in \mathbb{R}^{m \times 1}$ where $l_i = \sum_j (l_{ij} - \beta)$. Similarly, Eq. (9) is also a standard QP problem.

In each round of feedback, the ranking step and suggestion step are alternatively solved till convergence. After that, $q$ unlabeled samples with the smallest $v$s are taken as the feedback candidates as the model is not confident about their ranking scores.

### 3.3 Implementation Details

Real-time response is highly demanded for interactive INS tasks, yet it may be too time-consuming to iteratively solve the above QP problems, especially on large-scale datasets where the problem scale $m$ is extremely large. Hence, we design two optimization strategies to efficiently solve the proposed model, including an approximate solution for iteratively solved QP problems and a top-$K$ search scheme to reduce the problem scale.

**Approximate solution.** In order to reduce the execution time, we abandon the convergence condition in alternative optimization strategy and successively solve the ranking step and suggestion step only once in each round of feedback. In this case, we don’t have to strictly follow the constraints in Eq. (8) and Eq. (9), and both $f$ and $v$ can be solved by closed-form expressions.

In ranking step, $f$ can be solved by

$$f_i = \begin{cases} y_i, & \text{if } x_i \in \Psi \\ f_i - \min(f) \left(\frac{\max(f) - \min(f)}{\max(f)}\right), & \text{otherwise} \end{cases}$$

(10)

where $\min(\cdot)$ and $\max(\cdot)$ denote the minimum and maximum value of a vector, respectively, and $\hat{f}$ is solved by

$$\hat{f} = (P + Q)^{-1} Q y$$

(11)

And in suggestion step, $v$ can be solved by

$$v = -\frac{l}{\gamma m}$$

(12)

Since we only concern about the relative magnitude of $v$ instead of its real value, the constant terms can be further omitted. Hence, $v$ is solved by

$$v = -\hat{l}$$

(13)

where $\hat{l} = [\hat{l}_i]_{i=1}^n$, and $\hat{l}_i = \sum_j l_{ij}$.

Since Eq. (10) and (13) only involve basic matrix operations, we can use GPU to accelerate the computational process. And the comparative results between the original QP solution and the GPU-accelerated approximate solution will be discussed in Section 4.4.

After successively solving $f$ and $v$, the user is asked to provide feedback scores for $q$ least confident unlabeled samples. Then in the beginning of the next round of feedback, we re-use Eq. (3) to update $v$, i.e., unlabeled samples are all considered to be low-confidence while labeled samples are all regarded as high-confidence ones. Once the predetermined maximum round of feedback $T$ is reached, the ranking score of the probe $f_m$ is removed, and the ranking score is denoted as $f^* = [f_i]_{i=1}^n$. The overall procedure of CAAF is summarized in Algorithm 1.

**Top-K search scheme.** Since relevant samples tend to be concentrated in the top of the initial ranking list [46], we follow a common practice that selects only the top-$K$ samples with the highest initial ranking scores to form the gallery set $\mathcal{G}$ [47], [48], i.e., the number of galleries $n = K$, the total number of samples $m = K + 1$. Therefore, we can set a relatively small $K$ to balance the performance and computational cost. Eventually, the ranking list concatenates two parts: (1) the re-ranked top-$K$ galleries, and (2) the remaining galleries in the initial ranking orders. The rationality of such a practice will also be discussed in Section 4.4.

### 4 Experiments

This section is divided into four parts. In the first part, we compare the proposed CAAF with existing AL methods on three image-based INS datasets. Then in the second part, we report the evaluation results in large-scale video-based INS task in TRECVID 2021. The third part analyzes the proposed confidence modeling scheme from both the distribution of feedback samples and the effect of weight modulation. And finally, the rationality of implementation details, including approximate solution and top-$K$ search scheme, is discussed in the last part.

#### 4.1 Comparison with Existing AL Methods

**4.1.1 Datasets**

In this section, we evaluate our method on 3 image-based retrieval datasets:

- **Holidays** [49] is one of the most widely used image retrieval benchmark. There are 1,491 images of 500 landscapes collected from personal holiday albums, and each landscape has one query.
- **Oxford5k** [30] is also a popular dataset for image retrieval. It contains 5,062 images of 11 buildings in
Unless otherwise specified, we set $\alpha = 0.01$ and $K = 300$ in all our experiments. Besides, we set $q = 5$ and $T = 4$, i.e., $q \times T = 20$ feedback samples are generated for each probe in total.

### 4.1.5 Baseline Methods

We choose 9 baseline methods to compare:

- **Rand** is a commonly used baseline for AL approaches, where feedback samples are randomly chosen from unlabeled set.

- **GappedTopK** [39] partitions all galleries into $K$ clusters based solely on the ranking scores. Then from each cluster the sample with the highest ranking score is selected to form the feedback set.

- **LossMin** [13] designs an expected hinge rank loss to select samples that correspond to the lower bound on the AUC. By selecting samples that will minimize the expected hinge rank loss, the rank-learning measured by the AUC is guaranteed to be maximized.

- **VM** [14] uses a DCG-like gain function to measure the ranked list sampled from the rank distribution, and unlabeled samples that cause the largest variance in the gain will be selected for user feedback.

- **CoreSet** [9] is one of the state-of-the-art geometric techniques that tries to choose a representative subset of the entire dataset.

- **VAAL** [10] tries to select the most representative samples by training a variational auto encoder (VAE) and an adversarial network to discriminate between unlabeled and labeled data. Then probability associated with the discriminator’s predictions serves as a score to select low-confidence samples for feedback.

- **SPAL** [44] simultaneously considers the potential value and easiness of an instance by integrating a self-paced regularizer and a term estimates the distribution difference between labeled and unlabeled data into the same objective function.

- **UncertainGCN** [11] is a state-of-the-art AL method that constructs a sequential graph convolutional network (GCN) to distinguish unlabeled samples from labeled ones. The outputs of the GCN are served as confidence scores, and an uncertainty sampling approach is applied to select samples whose confidence scores are closest to a pre-defined small margin.

- **CoreGCN** [11] integrates geometric information between the labeled and unlabeled graph representation by the CoreSet strategy on the $l_2$ distances of features extracted by the GCN in UncertainGCN.

We categorize the above methods into 3 groups according to their original application scenario, and summarize them in Table 1.

All baseline methods except for VAAL are implemented with their default settings in our experiment. As for VAAL, it requires a relatively large number of labeled samples to train the VAE and the adversarial network, yet it’s impractical for our interactive retrieval settings where the probe is the only labeled sample at the very beginning. Therefore, we randomly select $q$ samples in the first round of feedback, and VAAL is applied from the second round of feedback. For fair comparison, all baseline methods, like CAAF, are applied to the top-$K$ samples in the initial ranking list. And they are only used to generate feedback suggestions for RF, then the feedback scores are diffused by the same ranking process described in Sec. 3.2.
VM are two exceptions, as the former does not require an extra model, and the latter requires multiple sampling of unlabeled samples to estimate their feedback value. And thanks to our approximate optimization scheme, CAAF is able to generate feedback samples in less than 5 ms, which is equivalent to most ranking-oriented methods as well as Rand.

Figure 3 shows the dynamic performance of all AL methods as the feedback round $T$ increases on three datasets. We can observe that classification-oriented methods generally perform worse than ranking-oriented ones—it makes sense since INS is intrinsically a ranking problem instead of a classification problem. But it’s worth noting that different from the experimental results reported in other literatures, some classification-oriented methods, e.g., CoreSet and VAAL, can be inferior to Rand, which can be owing to the lack of labeled samples—unlike classification tasks where a subset taking up $5 \sim 10\%$ of the whole unlabeled set is pre-labeled for a cold start, INS tasks take the probe as the only labeled sample in the first round of feedback. And the total feedback samples in classification tasks is usually far more than that in INS tasks, as it requires a large amount of labeled samples to train a robust classification model. However, it is unfriendly to ask the user to check too many samples during an interactive search.

We can also observe from Figure 3 that CAAF surpasses almost every other baseline methods. Among them, GappedTopK is the closest competitor of our method. In the last round of feedback, GappedTopK achieves 87.1% on Holidays, 60.98% on Oxford5k, and 89.96% on CUHK03. However, CAAF achieves 87.46%, 61.51% and 91.33% on Holidays, Oxford5k and CUHK03, respectively, which is the highest performance among that of all baseline methods. Such results demonstrate the efficiency of CAAF in selecting valuable feedback candidates.

4.2 Evaluation on TRECVID 2021

In this section, we evaluate our method on the INS task of TRECVID 2021 to demonstrate the capability of handling extremely large-scale datasets.

4.2.1 Introduction of TRECVID INS Evaluation

The INS task of TRECVID 2021 contains 20 query topics about specific persons doing specific actions, and participants are asked to retrieve all relevant video segments from EastEnders series. The entire dataset contains more than 470,000 video shots coming from 244 episodes, with a total length of more than 464 hours. Each participant can submit 4 fully automatic runs that take official topics as input and generate results without any human intervention, and 4 interactive runs that take human interactions to filter or re-rank automatically generated search results for up to a period of 5 elapsed minutes per search and 1 user per system run.

4.2.2 Baseline Methods

In this paper, we report the official evaluation results of the top-3 interactive runs as well as their corresponding automatic runs in INS task of TRECVID 2021:

- F_M_A_B_WHU_NERCMS.21_2 (Auto1) is the first-ranked automatic run in INS task of TRECVID 2021.
- F_M_A_B_WHU_NERCMS.21_4 (Auto2) is the second-ranked automatic run in INS task of TRECVID 2021.

1. https://www-nlpir.nist.gov/projects/tv2021/ins.html
In order to adapt to the INS task of TRECVID 2021, we make some necessary modifications on the proposed CAAF: (1) We concatenate the facial visual feature of the character appeared in each video segment and the category encoding of action taken by the character as the semantic feature, as there is no specific query sample in the INS task. (2) We take the average feature of each video segment. (3) We add a temporal expansion term to reflect the temporal similarity between each video segments.\(W_{ij} = e^{-\lambda|t_i - t_j|}W_{ij}\), where \(|t_i - t_j|\) calculates the temporal distance between shot \(i\) and shot \(j\), and we set \(\lambda = 0.005\). The intuition is that temporally consecutive video segments tend to share the same topic. (4) We replace the binary reference score \(y\) with the relevance score obtained by Auto2 and set \(v = 1\) in the first round of feedback, as the semantic features are somewhat weak to provide adequate information. (5) We set \(q = 18\) in the contest, but \(T\) is no longer fixed since the total time is fixed, and we try to take as many feedback rounds as possible.

### 4.2.3 Supplementary Evaluation

In addition to the official evaluation results, we supplement an interactive run (Auto2+CAAF*) where the total feedback number of each topic is the same as that of Auto2+TopK, i.e., 176 samples are fed back to the user in Topic 9319 altogether, 169 samples are fed back to the user in Topic 9320, and so on.

![Fig. 5. The 11-point interpolated precision-recall curve of the supplementary Auto2+CAAF* and all baseline methods in TRECVID 2021, where the values in the legend represent the mAP of each method.](image)

As presented Table 2, with the same initial search result, Auto2+TopK performs 0.1% better than Auto2+CAAF when the total interaction time is constrained to 5 minutes. However, the user has to check 105 video segments on average with Auto2+TopK, which is 28 more than that of Auto2+CAAF. And with the same total feedback number, Auto2+CAAF* outperforms Auto2+TopK by 5.9%.

Moreover, Figure 5 shows that Auto2+CAAF* even surpasses the first-ranked Auto1+TopK, although the performance of Auto2 itself is inferior to that of Auto1.

### 4.2.4 Result Analysis

As presented Table 2, with the same initial search result, Auto2+TopK performs 0.1% better than Auto2+CAAF when the total interaction time is constrained to 5 minutes. However, the user has to check 105 video segments on average with Auto2+TopK, which is 28 more than that of Auto2+CAAF. And with the same total feedback number, Auto2+CAAF* outperforms Auto2+TopK by 5.9%.

Moreover, Figure 5 shows that Auto2+CAAF* even surpasses the first-ranked Auto1+TopK, although the performance of Auto2 itself is inferior to that of Auto1.

Such experimental results demonstrate that CAAF is capable of dealing with massive data, and that CAAF is more efficient than directly checking the top-ranked samples in the initial ranking result.

### 4.3 Analysis on Confidence Modeling

This section analyzes the proposed confidence modeling scheme from two aspects: (1) the distribution of feedback samples, and (2) the effects of weight modulation.

### Table 2: Comparison Between TopK and CAAF in TRECVID 2021

| Topic     | Official #FB | Official AP | Supplementary #FB | Supplementary AP |
|-----------|--------------|-------------|-------------------|------------------|
| 9319      | 176          | 59.8        | 176               | 70.7             |
| 9320      | 169          | 79.0        | 169               | 80.2             |
| 9321      | 182          | 65.3        | 182               | 72.1             |
| 9322      | 97           | 56.4        | 57                | 59.1             |
| 9323      | 90           | 51.5        | 90                | 57.8             |
| 9324      | 91           | 49.4        | 91                | 56.0             |
| 9325      | 120          | 71.9        | 120               | 75.3             |
| 9326      | 138          | 74.5        | 138               | 78.4             |
| 9327      | 130          | 74.0        | 130               | 75.7             |
| 9328      | 98           | 20.9        | 98                | 34.3             |
| 9329      | 114          | 21.5        | 114               | 24.5             |
| 9330      | 97           | 51.9        | 97                | 55.2             |
| 9331      | 104          | 39.8        | 104               | 45.9             |
| 9332      | 116          | 51.6        | 116               | 60.2             |
| 9333      | 77           | 40.4        | 77                | 45.9             |
| 9334      | 79           | 35.2        | 79                | 45.1             |
| 9335      | 60           | 6.5         | 60                | 14.0             |
| 9336      | 67           | 16.5        | 67                | 24.0             |
| 9337      | 50           | 12.8        | 50                | 15.9             |
| 9338      | 53           | 41.8        | 53                | 46.7             |

| Topic     | Mean #FB | Mean AP | Mean #FB | Mean AP |
|-----------|----------|---------|----------|---------|
| 9319      | 105.4    | 46.0    | 78.4     | 45.9    |

*Authorized access level
1. **Initial Ranking**

| T  | Holidays | Oxford5k | CUHK03 |
|----|----------|----------|--------|
| 1  | 2430 66  | 264 11  | 6955 5 |
| 2  | 1104 720 | 129 83  | 4013 168 |
| 3  | 751 970  | 46 100 | 889 2365 |
| 4  | 295 1047 | 28 69  | 532 1908 |

2. **Fig. 6.** The heat map about how feedback samples distribute in the initial ranking list, where the deeper color reflects the denser distribution.

3. **Fig. 7.** The ratio of relevant and irrelevant samples in feedback samples generated by CAAF.

4. **Fig. 8.** A visualized example of the feedback samples selected by CAAF on Holidays dataset, where the first line of each row shows the feedback samples in the T-th round, and the second line shows their initial rankings. Images with blue and orange bounding boxes are relevant and irrelevant samples, respectively.
### 4.3.1 Analysis on Feedback Samples

We count the distribution of the feedback samples in the initial ranking list, and the heat map is illustrated in Figure 6. We can observe that in the first round of feedback, the feedback samples tend to be concentrated in the top of the initial ranking list. Then as T increases, CAAF selects samples from the middle and back segments of the initial ranking list. This indicates that CAAF is able to select diverse feedback samples that are not so similar to the probe, which can be owing to the MR framework that enables CAAF to consider the relationship between both labeled and unlabeled samples [20].

We further count the labels of the feedback samples, and the ratio of relevant and irrelevant samples is illustrated in Figure 7. We can observe that the proportion of relevant feedback samples gradually decreases as the round of feedback increases. Combined with Figure 6, it indicates that relevant samples are more likely to appear in the top of the initial ranking list, demonstrating the rationality of selecting only the top-K samples for interactive INS.

A visualized example of the feedback samples selected by CAAF is shown in Figure 8. As T increases, the number of relevant samples gets fewer and fewer, while the initial rankings of the selected samples gets lower and lower.

### 4.3.2 Analysis on Weight Modulation

As we claimed in Section 3.1, the proposed confidence modeling scheme can not only indicate valuable feedback samples for AL, but also modulate the propagation weight in MR and improve the ranking accuracy. Experiments in Section 4.1 have demonstrated the first half of the claim. Here we set $\nu = 1$ in the beginning of each round of feedback to testify the second half of the claim. In this case, the ranking step of CAAF degrades into the classical MR.

As shown in Table 3, although CAAF performs slightly worse than MR when $T = 0$, i.e., no sample is labeled except for the probe, it consistently outperforms MR on all three datasets when $T \geq 1$. And the difference between CAAF and MR can be up to 1.72%, 3.12% and 1.72% on Holidays, Oxford5k and CUHK03, respectively. Such results demonstrate that $\nu$ does help improve the ranking accuracy by modulating the affinity matrix in the graph construction of MR.

### 4.4 Rationality of Implementation Details

This section studies the rationality of the two optimization strategies introduced in Section 3.3, including approximate optimization and top-K search scheme. The experiments are conducted on the three image-based datasets with default settings introduced in Section 4.1.

#### 4.4.1 Rationality of Approximate Solution

We compare the search accuracy and the execution time taken in each round of feedback of both GPU-accelerated approximate solution (Appr.) and the original QP solution (QP), and the results are shown in Table 4. We can observe that Appr. achieves equivalent performance as QP, yet the former’s execution time is only around one tenth of the latter’s. Therefore, it is rational to simplify the optimization process through the proposed approximation scheme.

#### 4.4.2 Rationality of Top-K Search Scheme

We also analyze the search accuracy and execution time with varying K’s. As illustrated in Figure 9, the performance tends to be stable when $K \geq 300$. Meanwhile, the execution time per round of query exponentially increases as K goes up. Such results demonstrate that selecting only the top-K galleries is a reasonable practice for balancing search accuracy as well as time cost, and we set $K = 300$ to balance the performance and computational cost.

### 5 Conclusions

This paper investigates a confidence-aware active feedback (CAAF) method specifically designed for interactive INS.
The core idea and main novelty lies in an explicit assessment of the ranking confidence of each gallery sample, which supports more accurate feedback suggestion as well as manifold ranking. Furthermore, with an approximate solution and a top-K search scheme, CAAF can be efficiently applied to interactive INS on large-scale datasets. Extensive experiments on both image and video INS tasks demonstrate the effectiveness of our proposed method. In particular, CAAF surpasses the first-ranked record in the public large-scale video INS evaluation of TRECVID 2021.

In the future, we will pay more attention to noisy human annotations, as we find that the quality of user feedback can severely affect the search performance.

**APPENDIX A**

**Detailed Derivation of the Ranking Step**

The detailed derivation form Eq. 7 to Eq. 8 is demonstrated as follows:

\[
\mathcal{E}_v(f) = \frac{1}{2} \sum_{i,j} \tilde{v}_{ij}^2
\]

\[
= \frac{1}{2} \sum_{i,j} \tilde{v}_{ij} w_{ij} (f_i - f_j)^2 + \sum_{i,j} \alpha \tilde{v}_{ij} (f_i - y_i)^2
\]

\[
= \frac{1}{2} \sum_{i,j} \tilde{v}_{ij} w_{ij} (f_i^2 + f_j^2 - 2f_i f_j) + \sum_{i,j} \alpha \tilde{v}_{ij} (f_i^2 + y_i^2 - 2f_i y_i)
\]

\[
= \sum_{i,j} \tilde{w}_{ij} f_i^2 - \sum_{i,j} \tilde{w}_{ij} f_i f_j + \sum_{i,j} \alpha \tilde{v}_{ij} f_i^2 - 2 \sum_{i,j} \alpha \tilde{v}_{ij} f_i b_i + \sum_{i,j} \alpha \tilde{v}_{ij} y_i^2
\]

\[
= f^\top Df - f^\top \tilde{W}f + f^\top Qf - 2 f^\top Qy + C
\]

\[
= f^\top (P + Q)f - 2 f^\top Qy + C
\]

where \( \tilde{w}_{ij} = \tilde{v}_{ij} w_{ij} \) is the confidence-weighted affinity between \( x_i \) and \( x_j \), \( D \) is a diagonal matrix where \( d_{ii} = \sum_{j} \tilde{w}_{ij} \), \( Q \) is a diagonal matrix where \( q_{ii} = \sum_{j} \alpha \tilde{v}_{ij} \), \( P = D - \tilde{W} \) is the Laplacian matrix of the confidence-weighted graph, and \( C = \sum_{i,j} \alpha \tilde{v}_{ij} y_i^2 \) is a term irrelevant with \( f \). Hence, the ranking step of CAAF can be considered as a confidence-weighted manifold ranking problem.

**REFERENCES**

[1] L. Zheng, Y. Yang, and Q. Tian, “SIFT meets CNN: A decade survey of instance retrieval,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1224–1244, 2018.

[2] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline),” in ECCV (4), ser. Lecture Notes in Computer Science, vol. 11208. Springer, 2018, pp. 501–518.

[3] Y. Gu, C. Li, and Y. Jiang, “Towards optimal CNN descriptors for large-scale image retrieval,” in ACM Multimedia. ACM, 2019, pp. 1768–1776.

[4] F. Radenovic, G. Toliyas, and O. Chum, “Fine-tuning CNN image retrieval with no human annotation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1655–1668, 2019.

[5] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” in CVPR. IEEE Computer Society, 2014, pp. 1386–1393.

[6] M. Wray, H. Doughty, and D. Damen, “On semantic similarity in video retrieval,” in CVPR. Computer Vision Foundation / IEEE, 2013, pp. 3650–3660.

[7] C. Liang, B. Huang, R. Hu, C. Zhang, X. Jing, and J. Xiao, “A unsupervised person re-identification method using model based representation and ranking,” in ACM Multimedia. ACM, 2015, pp. 771–774.

[8] D. Harman, “Relevance feedback revisited,” in SIGIR. ACM, 1992, pp. 1–10.

[9] O. Sener and S. Savarese, “Active learning for convolutional neural networks: A core-set approach,” in ICML (Poster). OpenReview.net, 2018.

[10] S. Sinha, S. Ebrahimi, and T. Darrell, “Variational adversarial active learning,” in ICCV. IEEE, 2019, pp. 5971–5980.

[11] R. Caramalau, B. Bhattachar, and T. Kim, “Sequential graph convolutional network for active learning,” in CVPR. Computer Vision Foundation / IEEE, 2021, pp. 9583–9592.

[12] P. Cai, W. Gao, A. Zhou, and K. Wong, “Relevant knowledge helps in choosing right teacher: active query selection for ranking adaptation,” in SIGIR. ACM, 2011, pp. 115–124.

[13] P. Donmez and J. G. Carbonell, “Active sampling for ranking learning via optimizing the area under the ROC curve,” in ECIR, ser. Lecture Notes in Computer Science, vol. 5478. Springer, 2009, pp. 78–89.

[14] W. Cai and Y. Zhang, “Variance maximization via noise injection for active sampling in learning to rank,” in CIKM. ACM, 2012, pp. 1809–1813.

[15] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, “Ranking on data manifolds,” in NIPS. MIT Press, 2003, pp. 169–176.

[16] M. P. Kumar, B. Packer, and D. Koller, “Self-paced learning for latent variable models,” in NIPS. Curran Associates, Inc., 2010, pp. 1189–1197.

[17] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, and A. G. Hauptmann, “Self-paced learning with diversity,” in NIPS, 2014, pp. 2078–2086.

[18] B. Zhao, F. Wang, C. Zhang, and Y. Song, “Active model selection for graph-based semi-supervised learning,” in ICASSP. IEEE, 2008, pp. 1881–1884.

[19] B. Settles, “Active learning literature survey,” University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[20] S. Huang, R. Jin, and Z. Zhou, “Active learning by querying informative and representative examples,” in NIPS. Curran Associates, Inc., 2010, pp. 892–900.

[21] J. C. Bezdek and R. J. Hathaway, “Convergence of alternating optimization,” Neural Parallel Sci. Comput., vol. 11, no. 4, pp. 351–368, 2003.

[22] S. Tong and E. Y. Chang, “Support vector machine active learning for image retrieval,” in ACM Multimedia. ACM, 2001, pp. 107–118.

[23] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, “Multi-class active learning for image classification,” in CVPR. IEEE Computer Society, 2009, pp. 2372–2379.

[24] M. Rothmann, K. Kahl, and H. Gottschalk, “Deep bayesian active semi-supervised learning,” in ICMLA. IEEE, 2018, pp. 158–164.

[25] W. H. Beluch, T. Genewein, A. Nürnberg, and J. M. Köhler, “The power of ensembles for active learning in image classification,” in CVPR. IEEE Computer Society, 2018, pp. 9366–9377.

[26] Z. Liu, J. Wang, S. Gong, H. Lu, and D. Tao, “Deep Reinforcement Active Learning for Human-in-the-Loop Person Re-Identification,” in ICCV, 2019.

[27] H. S. Seung, M. Oppler, and H. Sompolinsky, “Query by committee,” in COLT. ACM, 1992, pp. 287–294.

[28] N. Abe and H. Mamitsuka, “Query learning strategies using latent variable models,” in COLT. ACM, 2011, pp. 115–124.

[29] A. McCallum and K. Nigam, “Employing EM and pool-based active learning for text classification,” in ICML. Morgan Kaufmann, 1999, pp. 99–106.

[30] A. McCallum and K. Nigam, “Self-paced learning with diversity,” in NIPS. Curran Associates, Inc., 2010, pp. 1189–1197.

[31] W. H. Beluch, T. Genewein, A. Nürnberg, and J. M. Köhler, “The power of ensembles for active learning in image classification,” in CVPR. IEEE Computer Society, 2018, pp. 9366–9377.

[32] Z. Liu, J. Wang, S. Gong, H. Lu, and D. Tao, “Deep Reinforcement Active Learning for Human-in-the-Loop Person Re-Identification,” in ICCV, 2019.

[33] H. S. Seung, M. Oppler, and H. Sompolinsky, “Query by committee,” in COLT. ACM, 1992, pp. 287–294.

[34] N. Abe and H. Mamitsuka, “Query learning strategies using latent variable models,” in COLT. ACM, 2011, pp. 115–124.
[33] Y. Baram, R. El-Yaniv, and K. Luz, “Online choice of active learning algorithms,” in ICML. AAAI Press, 2003, pp. 19–26.
[34] N. Cebon and M. R. Berthold, “Active learning for object classification: from exploration to exploitation,” Data Min. Knowl. Discov., vol. 18, no. 2, pp. 283–299, 2009.
[35] S. Ebert, M. Fritz, and B. Schiele, “RALF: A reinforced active learning formulation for object class recognition,” in CVPR. IEEE Computer Society, 2012, pp. 3626–3633.
[36] A. Bondu, V. Lemaire, and M. Boulle, “Exploration vs. exploitation in active learning: A bayesian approach,” in The 2010 International Joint Conference on Neural Networks (IJCNN). IEEE, 2010, pp. 1–7.
[37] T. T. Osugi, K. Deng, and S. D. Scott, “Balancing exploration and exploitation: A new algorithm for active machine learning,” in ICDM. IEEE Computer Society, 2005, pp. 330–337.
[38] B. Zhang, L. Li, S. Yang, S. Wang, Z. Zha, and Q. Huang, “State-relabeling adversarial active learning,” in CVPR. Computer Vision Foundation / IEEE, 2020, pp. 8753–8762.
[39] X. Shen and C. Zhai, “Active feedback in ad hoc information retrieval,” in SIGIR. ACM, 2005, pp. 59–66.
[40] B. Long, J. Bian, O. Chapelle, Y. Zhang, Y. Inagaki, and Y. Chang, “Active learning for ranking through expected loss optimization,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 5, pp. 1180–1191, 2015.
[41] Q. Wang, W. Wu, Y. Qi, and Z. Xin, “Combination of active learning and self-paced learning for deep answer selection with bayesian neural network,” in ECAL, ser. Frontiers in Artificial Intelligence and Applications, vol. 325. IOS Press, 2020, pp. 1587–1594.
[42] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in ICML, ser. ACM International Conference Proceeding Series, vol. 382. ACM, 2009, pp. 41–48.
[43] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, “Active self-paced learning for cost-effective and progressive face identification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 1, pp. 7–19, 2018.
[44] Y. Tang and S. Huang, “Self-paced active learning: Query the right thing at the right time,” in AAAI. AAAI Press, 2019, pp. 5117–5124.
[45] S. P. Boyd, L. Vandenberghe, S. P. Boyd, and L. Vandenberghe, Convex optimization. Cambridge university press, 2004.
[46] J. Zobel, “How reliable are the results of large-scale information retrieval experiments?” in SIGIR. ACM, 1998, pp. 307–314.
[47] H. Jegou, M. Douze, and C. Schmid, “Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search,” in ECCV (1), ser. Lecture Notes in Computer Science, vol. 5302. Springer, 2008, pp. 304–317.
[48] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in CVPR. IEEE Computer Society, 2007.
[49] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in CVPR. IEEE Computer Society, 2014, pp. 152–159.
[50] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in CVPR. IEEE Computer Society, 2017, pp. 3652–3661.
[51] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in ICLR, 2015.
[52] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, “Imagenet: A large-scale hierarchical image database,” in CVPR. IEEE Computer Society, 2009, pp. 248–255.