In Pursuit of Beauty: Aesthetic-Aware and Context-Adaptive Photo Selection in Crowdsensing

Tongqing Zhou, Zhiping Cai, *Member, IEEE*, Fang Liu, and Jinshu Su, *Senior Member, IEEE*

***Abstract*—The pervasive view of the mobile crowd bridges various real-world scenes and people's perceptions with the gathering of distributed crowdsensing photos. To elaborate informative visuals for viewers, existing techniques introduce photo selection as an essential step in crowdsensing. Yet, the aesthetic preference of viewers, at the very heart of their experiences under various crowdsensing contexts (e.g., travel planning), is seldom considered and hardly guaranteed. We propose CrowdPicker, a novel photo selection framework with adaptive aesthetic awareness for crowdsensing. With the observations on aesthetic uncertainty and bias in different crowdsensing contexts, we exploit a joint effort of mobile crowdsourcing and domain adaptation to actively learn contextual knowledge for dynamically tailoring the aesthetic predictor. Concretely, an aesthetic utility measure is invented based on the probabilistic balance formalization to quantify the benefit of photos in improving the adaptation performance. We prove the NP-hardness of sampling the best-utility photos for crowdsourcing annotation and present a (1-1/e) approximate solution. Furthermore, a two-stage distillation-based adaptation architecture is designed based on fusing contextual and common aesthetic preferences. Extensive experiments on three datasets and four raw models demonstrate the performance superiority of CrowdPicker over four photo selection baselines and four typical sampling strategies. Cross-dataset evaluation illustrates the impacts of aesthetic bias on selection.**

***Index Terms*—Computer vision, crowdsourcing, image analysis, mobile computing, transfer learning.**

I. Introduction

Built-in cameras on nowadays pervasive mobile devices can be employed as wide distributed machine eyes for people to visually perceive, experience, and interpret the physical world, which are known as photo (or visual) crowdsensing [1], [2], [3], [4]. Such a paradigm has fueled various real-world applications, including online gallery [5], [6], event sensing [7], [8], and social sharing [9], [10]. For example, the fictional *Dick Whittington*, if were alive today, could view pictures of London via crowdsensing before making his way to the golden streets in rumors [11]. In fact, many platforms (e.g., Beautiful China [12]) have managed to use pictures of resorts and points of interest (PoIs) from mobile users' contributions to facilitate virtual tours online, during the pandemic of COVID-19.

Offloading and scanning all the collected crowdsensing photos, which can have large volumes, is inefficient and inconvenient for the potential users (viewers) [13]. As the ultimate goal of crowdsensing lays in helping one to understand the interesting targets, manually skipping many unexpected photos for the interested will undoubtedly degrade their viewing experience. To this end, photo selection is usually conducted as an essential step in mobile crowdsensing to elaborate highlight and summarized views [14], [15], [16], [17], [18], [19].

Most existing efforts in this field are devoted to filtering noise or finding representative photos by gauging the descriptive characteristics. For example, Hua et al. [16] leverage semantic hash to calculate image similarity and perform deduplication on similar images. Alternatively, visually representative photos are picked out by content clustering and technical quality evaluation [13], [15], [17]. Furthermore, for a selection with spatial representativeness, the methods in [14], [18], [19] formalize photos' conjoint coverage with location and direction information in the metadata. Although a group of informative photos can be obtained with these methods, user preference and affective state in photo viewing [20] are unfortunately ignored in the literature. In fact, as revealed by a user study we conducted (Section III-A), *the aesthetic perception of crowdsensing photos is generally more favored than the above redundancy or representativeness selection criteria under various crowdsensing contexts*. For example, when requesting photos for travel planning [10], users tend to be more interested in those with beautiful scenes, even though some ordinary-looking aspects or angles are not comprehensively captured; when extracting city images for urban design or propagation [21], good-looking pictures are believed to invoke greater empathy, thereby treasured alike their representative counterparts.

With these observations, this work is then motivated to fill the void of aesthetic-aware photo selection in mobile crowdsensing with a novel framework, termed as CrowdPicker. Basically, the pursuit of subjective user preference can facilitate better satisfaction levels, thus making it an effective complement, while not a replacement, for existing objective photo selection strategies. In practice, the aesthetic criterion has already been successfully used in photograph editing [22] and image retrieval [23].

Manuscript received 6 December 2021; revised 21 November 2022; accepted 5 January 2023. Date of publication 24 January 2023; date of current version 8 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62102425, 62072465, 62172155, and U22B2005, in part by the Science and Technology Innovation Program of Hunan Province under Grants 2021RC2071 and 2022RC3061, and in part by the Natural Science Foundation of Hunan Province under Grants 2022JJ40564 and 2022JJ30667. Recommended for acceptance by J. Tang. *(Corresponding author: Fang Liu.)*

Tongqing Zhou, Zhiping Cai, and Jinshu Su are with the College of Computer, National University of Defense Technology, Hunan 410073, China (e-mail: zhoutongqing@nudt.edu.cn; zpcai@nudt.edu.cn; sjs@nudt.edu.cn).

Fang Liu is with the School of Design, Hunan University, Changsha, Hunan 410012, China (e-mail: fangl@hnu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TKDE.2023.3237969, provided by the authors.

Digital Object Identifier 10.1109/TKDE.2023.3237969

Wherein, deep learning models [24], [25] are generally trained for specific and large photo datasets (e.g., [26]) and used to predict photos' aesthetics. Yet, *building a generalized selection framework for different aesthetic contexts is non-trivial*, as the above well-trained models usually have limited performance when applied to crowdsensing tasks (Section III-B). This owes to the uncharted nature of crowdsensing tasks, which is launched to collect photos for seldom-visited targets or even unexplored contexts (e.g., visuals for heritage buildings on a rainy day).

In particular, we are facing two fundamental challenges here: 1) *aesthetic uncertainty* of the crowdsensing context, namely, no prior knowledge (e.g., aesthetic labels, distributions) to validate the preferences and lead the selection; 2) *aesthetic bias* between different crowdsensing tasks, namely, visual domain shift or alternation that renders a static aesthetic assessment model non-adaptive. We point out that these challenges, even each can be solved with myopic annotation [27], [28] and transfer learning (a.k.a., domain adaptation) [29] independently, are non-orthogonal in essence and hard to resolve as a whole. To be specific, *non-informative aesthetic annotations that fails to relieve the context uncertainty would limit the adaptation performance; while improper adaptation would mislead the aesthetic estimation and in turn aggravate the uncertainty.*

In view of these challenges, CrowdPicker is designed on a joint effort of mobile crowdsourcing and domain adaptation that iteratively distill contextual knowledge on aesthetic preference for tuning an adaptive selection dynamically:

1) Given a crowdsourcing budget, we attempt to identify the subset of photos that, if aesthetically annotated, would yield the largest improvement on the adaptation performance. Instead of estimating each photo's utility independently, we argue that the aesthetic predictions of photos, usually represented as a probability distribution, should be ideally covered in equilibrium during sampling. Given the knowledge of such samples, the assessment model can learn to handle photos with different perceptual aesthetics comprehensively, so as to provide an accurate ranking for selection. Based on this insight, *a novel aesthetic utility* is devised by incorporating the orthogonal factors of aesthetic predictions' probability accumulation for overall informativeness and prediction difficulty for the contribution of every single photo. Photo sampling is then performed to find the photo subset with the maximum utility, which is, unfortunately, proved to be NP-hard. To provide a fast solution, we design a greedy-based algorithm with the approximation ratio theoretically bounded by a constant value.

2) We present a two-stage distillation-based adaptation architecture for mitigating aesthetic bias. The contextual preference from annotated samples and common knowledge inherited in a history/raw model are weighted balanced and combined through a loss function. With this, an updated model is progressively tailored to realize aesthetic awareness during photo selection. It is worth mentioning that, due to the adaptive characteristics of CrowdPicker, it is generic for different aesthetic assessment models (e.g., [23], [24]) and extendable to DNN models for different tasks, including visual classification, blind quality estimation, etc.

The main contributions of this paper are as follows:

- We conduct a user study that reveals the importance of aesthetics in crowdsensing photo selection and present observations on the aesthetic uncertainty and bias.
- We propose an adaptive photo selection framework with aesthetic-awareness based on joint exploitation of mobile crowdsourcing and domain adaptation.
- We design an aesthetic utility measure for sampling the most beneficial photos to adaptation. We rigorously prove that finding the maximum utility samples is NP-hard and present an approximate solution. We introduce a distillation-based adaptation and selection architecture using contextual and common aesthetic knowledge fusion.
- We evaluate CrowdPicker with extensive experiments on three datasets, each corresponding to a unique crowdsensing context. Experimental results show that CrowdPicker, built on either of four mainstream raw models, outperforms the photo selection baselines and myopic model adaptation strategies. Cross-dataset evaluation also demonstrates our observations on aesthetic bias among different crowdsensing contexts.

II. Related Work

In this part, we review the relevant literature on crowdsensing photo selection, aesthetic learning, and using crowdsourcing for context knowledge. We also discuss the differences of our work.

Mobile/crowd Photo Selection: There are many efforts on dealing with the large number of photos generated and uploaded by mobile devices, aiming to provide representative content for the viewers. Some work relies on heuristic models to quantify the informativeness of a photo subset in visually depicting the physical world. In [18] and [19], photo utility is calculated using a geographical coverage framework and the selection process attempts to find the photos that maximize the coverage. Considering the redundancy in crowd contributed photos, Yu et al. [16] make use of the DiffServ model to aggregate similar data into the same flow for cost-efficient transmission. On the other hand, data clustering is introduced in [13], [14], [17] to group similar photos based on photos' locations or contents and perform selection on the clusters in a way of scene summarization. However, existing literature mainly focuses on the descriptive capability of photos on potential targets (i.e., PoIs), while human perception of the selection is seldom studied. In contrast, our work focuses on providing aesthetic appeal via photo selection, which provides an effective complement to the existing objective selection criteria.

Learning Photo Aesthetics: Learning and quantifying photos' aesthetic appeal has been widely studied in image processing and supports many applications, such as photo-editing, image retrieval [20], [22]. Hand-crafted features are extracted and used for either full-reference or no-reference quality prediction [30]. By extracting more complex features to represent the perceptual quality, CNNs have greatly promoted the research on aesthetic learning [25]. Among the relevant efforts, some propose to categorize the photos into high and low aesthetics with specific models [22], [23], but is criticized to be unable to give

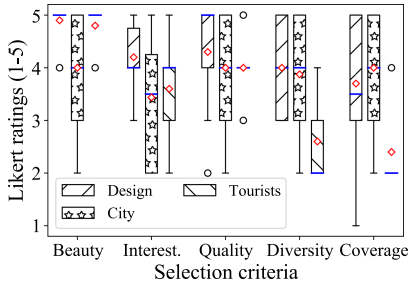


Fig. 1. Importance ratings of selection criteria for crowdsensing photo collection in different fields.

reliable assessment for images from same domain with similar context [24]. Further, in recent studies [24], [25], predictions of photo aesthetic are represented as score distribution and the assessment is performed as ranking tasks. Although accurate predictions can be made, we focus on the adaptivity issues in crowdsensing contexts, where photos are expected to show different visual characteristics with unknown aesthetic preferences. Adapting general models for specific purposes is also investigated in [31]. Given a query image, it proposes to search for a group of similar images from a training database and use these images to train a dedicated model. The assumption of having plenty of reference photos is not true in crowdsensing uncharted targets.

Personalized image aesthetic assessment (PIAA) has recently earn many attentions due to customization demand of users in applications like album management [32]. Relevant techniques generally adopt collaborative filtering [33], user interaction [34], or aesthetic differences estimation [35], [36], [37], [38] for transferring generic assessment into personalized score. For example, residual-based adaptation is used to estimate an aesthetic offset (w.r.t. generic prediction) for each image in [35], while user personality and image aesthetic are jointly learned for personality-aware aesthetic prediction via multi-task learning in [36], [37]. We note that PIAA shares a similar goal as aesthetic-aware photo selection in this work, namely, PIAA attempts to find dedicated photos for individuals and the latter is for dedicated photos on the sensing context. However, technically, PIAA focuses more on profiling individual preference with social factors, interaction, or personality discovery, while crowdsensing photo selection should primarily handle the uncharted nature and uncertain challenge of crowdsensing with limited budget. This leaves sampling and annotation for aesthetic prediction the main problem in this work.

Crowdsourcing for Perception Annotation: With the subjective judgement of human workers, crowdsourcing is often used to perform annotation tasks that are hard for machines [39]. In fact, crowdsensing can be regarded as a special form of crowdsourcing for collecting real-world data. Given usually limited crowdsourcing budgets in practice, an important problem for relevant research is sampling the portion of data that is most valuable for annotation. For example, items that are most influential to the other ones are chosen for crowdsourcing in [40], hoping to better infer the values of the rest if getting them labelled. In [41], the data subset with the maximum inner-differences is sampled. Zhou et al. [42] propose to group the data and send out the centroid of each cluster for crowdsourcing. Unlike those myopic sampling strategies that may miss knowledge on certain data labels, we invent a novel strategy that emphasizes balanced annotation with raw prior distribution. Finally, recruiting competent workers and providing proper incentives are also keys for accurate annotation [43], [44]. Since these are out of the scope of this work, we referred to the recent developments during our implementation.

III. Preliminaries and Overview

This section presents our motivation of studying aesthetic-aware photo selection, followed by the analysis of fundamental challenges in this context and our framework design.

A. Why Aesthetic Matters?: A User Study

To better understand the photo selection preferences of users on different crowdsensing tasks, we conducted a survey with 40 participants from the fields of design, computer science, and ordinary tourists. In practice, they may conduct crowdsensing for designing poll (e.g., building humanity award[1]), smart city management [45], and travel planning [10] purposes. The questionnaire presents participants with toy examples on their familiar crowdsensing contexts (i.e., designer, citizen, tourists), and asks two simple question on their preferences for viewing the provided photos and the possible reasons. The participants also had the chance to comment on the experience at the end.

First, they are asked to rate for the importance of different criteria in selecting photos to view in different contexts. Five criteria were taken from related work, with *Beauty* and *Interestingness* as aesthetic perception indicators [46], and *Quality* [15], *Diversity* [14], [16], and *Coverage* as traditional objective measures [18], [19]. Fig. 1 presents the rating results on a five-point Likert scale between 1 ("not important at all") and 5 ("very important"). We can see that the mean values (indicated with red circles) of the aesthetic factors are generally larger than those of the traditional measures, especially in the design and tourism field, where large margins can be observed. Remarkably, the beauty factor also gives the highest average rating and the smallest deviation.

Furthermore, as indicated by the participants, such aesthetic criteria were favored for their values in providing *inspiration* (40% of the participants) and *recommendation* (20% of the participants). In the comment section of the questionnaire, P05 with a background of computer science and tourist wrote that *"Diversity and coverage are something I cannot evaluate without scanning all the photos, but beauty and quality are easily perceived"*. Designer P11 commented that *"Beautiful and interesting pictures tend to provide me new design thoughts"*. To this end, realizing aesthetic awareness in crowdsensing photo selection is critical on behalf of the requesters/users, yet has been seldom investigated in the literature. This work is thus motivated to fill this void.

Considering the subjective discrepancies of different users on aesthetic, we will hereafter use a score distribution $P_i = [p_i^1, \ldots, p_i^{10}]$ to represent the ground truth of a photo's

[1][Online]. Available: http://cityaward.lifeweek.com.cn/

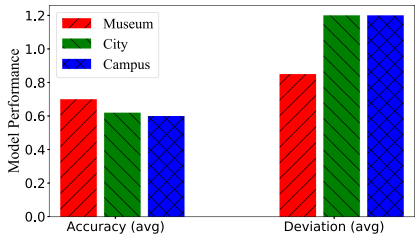


Fig. 2. The aesthetic assessment performance of the SoTA model on photos from three different crowdsensing tasks.

aesthetic level (ranging from 1 to 10), where p_i^j denoting the probability for photo pho_i to be perceived at score level j and $\sum_{j=1}^{10} p_i^j = 1$. The transformation of 5-point scale to 10-point scale is explained in Section IV-C.

B. Aesthetic Bias and Uncertainty

Assessing the aesthetics of crowdsensing photos is non-trivial, because such subjective perceptions will, on one hand, vary among crowdsensing tasks of different visual domains (i.e., *aesthetic bias*) and, on the other hand, are usually unknown a priori (i.e., *aesthetic uncertainty*). Formally, given a model M_0 trained on dataset D_0 and its aesthetic estimations $\widehat{P}_i$ on pho_i of some target dataset D_t, we have

$$Aesthetic\ bias = |f(D_t, M_0) - f(D_0, M_0)|,$$
$$f(\cdot)\ is\ accuracy\ or\ correlation\ measure \quad (1)$$

$$Aesthetic\ uncertainty = DEV(P, \widehat{P}) = \frac{1}{n}\sum |p_i^j - \widehat{p_i^j}|. \quad (2)$$

We owe such limitations to the nature of mobile crowdsensing that is launched to discover uncharted visuals and novel characteristics. Hence, a static aesthetic assessment model cannot work out for photo selection of all the emerging tasks, while designing a model for each task is definitely unaffordable and inscalable for mobile crowdsensing.

One can observe these fundamental challenges by applying a well-trained SoTA model to predict photo aesthetics in different contexts. As an example, Fig. 2 shows the basic prediction performance of NIMA-aes(MN) [25] (i.e., D_0) on three crowdsensing photo collections (i.e., D_t). As shown, compared with the achieved accuracy $f(D_0, M_0) = 0.81$, the accuracy of classifying photos into a high and low quality is rather low for photos from different contexts (even close to the performance of random guesses for the Campus dataset), indicating the existence of aesthetic bias. Further, the prediction deviation is around 1 on average, showing obvious uncertainty in inferring the aesthetic scores. More results for the cross-task cases are provided in Section VI-C4. Hence, the crux here is how to properly generalize a model and adapt it to different crowdsensing contexts.

C. Framework Overview

1) Basic Idea: Intuitively, for the uncertainty issues, mobile crowdsourcing can be used to collect aesthetic ratings for photos, while domain adaptation techniques are widely used to handle domain bias. Yet, we note that *the aesthetic bias and uncertainty challenges are non-orthogonal here, so an independent "annotation"+"adaptation" solution is inadequate*. That is, the non-informative annotation will limit the performance improvement of adaptation (shown in Section VI-C2), while improper adaptation will aggravate the uncertainty (shown in Section VI-C1). Hence, to mitigate the aesthetic estimation bias between tasks, explicit aesthetic knowledge of each context is needed to specify the preferences, while annotating all photos would easily exhaust the budget. This leads to the first research question:

RQ#1: "Which subsets of photos can provide proper context-specific aesthetic knowledge?"

On the other hand, to relieve the uncertainty in selecting photos with top-ranked aesthetics, generalized and adaptive aesthetic assessment is expected. For this, we need to answer:

RQ#2: "How to adapt the model for better aesthetic assessment performance in different contexts?"

From a high-level view, we attempt to build a measure for photos' aesthetic utility, leverage crowdsourcing to acquire context knowledge from the best-utility photo samples, and tune the model by carefully fusing this contextual knowledge and general aesthetic preference. Conventional aesthetic assessment models (a.k.a., raw models in this work), trained on large datasets, encode generic aesthetic perception, thus making a much better training basis compared with training from the scratch in crowdsensing contexts. The proposed selection framework leverages such encoded knowledge in raw models [25], [35] with agnostic backbones (e.g., MobileNet, VGG16) and data sources (e.g., AVA, FLICKR-AES) as aesthetic references, which formally represented as probability $\widehat{p_i^j}$, $j \in \{1, .., 10\}$ for each photo pho_i. We will study the impact of raw models in Section VI.

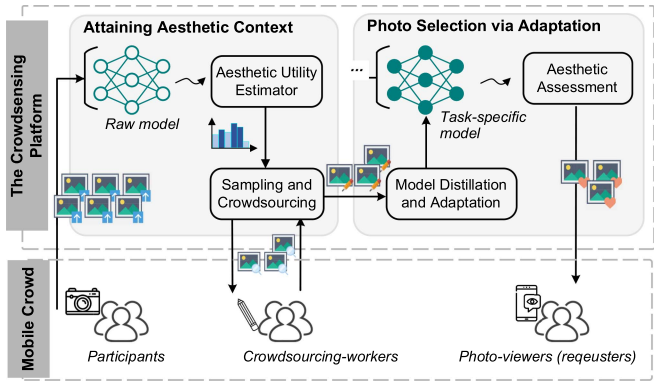
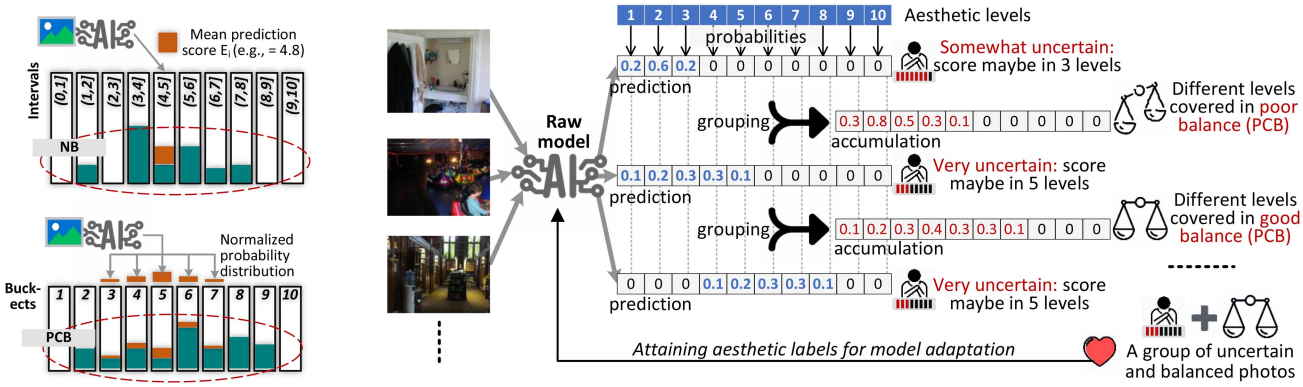


Fig. 3. Framework of CrowdPicker. Crowdsensing photos are gathered from the participants to the platform, who will perform aesthetic-aware selection to elaborate potentially preferred photos for the requesters.

2) Design: Fig. 3 illustrates the framework of CrowdPicker. As shown, it consists of four roles/entities: the photo-viewer (a.k.a., requester) that sends out a request by launching a photo collection task; the participants that join in the task and upload photos for the target domain; the crowdsensing platform that collects photos and accommodates the crowdsourcing and adaptation; and the crowdsourcing-workers that respond to aesthetic



(a) Different balance measures.

(b) A toy example of photo aesthetic utility estimation.

Fig. 4. The predictions of a raw model are used to measure photos' aesthetic utility in improving model performance once labeled. (a) An illustrative example for the NB measure (Section IV-A1) and the PCB measure (Section IV-A2). (b) The joint estimation workflow of a photo group's PCB degree and uncertainty degree (Section IV-A3). (a) Different balance measures. (b) A toy example of photo aesthetic utility estimation.

annotation tasks using their contextual knowledge. CrowdPicker works in two phases:

Phase I: Aesthetic context attaining to overcome uncertainty (investigated in Section IV). After gathering a set of crowdsensing photos $\mathbb{A} = \{pho_i\}$ $(i = 1, \ldots, n)$, CrowdPicker first gauges their aesthetic utility with the estimator. By carefully referring to the inaccurate predictions (i.e., $\widehat{P}$), it attempts to measure the benefit of a subset of photos, once labeled, in improving the aesthetic perception of the current crowdsensing context. Second, CrowdPicker finds and samples the photos with the best aesthetic utility and crowdsources them to workers for manual ratings. A budget will be considered here to limit the number of sampled and crowdsourced photos on behalf of the platform (Section IV-B).

Here we assume that the collected photos were taken from different locations and contain no duplication.[2] From the aspect of active learning strategy, such a label learning process belongs to the pool-based sampling.

Phase II: Photo selection via adaptation to mitigate bias (investigated in Section V). Given the aesthetic-labeled photos, CrowdPicker uses them as context preferences and performs adaptation on the raw model. Instead of simply fine-tuning, it first distills valuable general knowledge from the history training, which is integrated with the contextual knowledge to lead to a robust task-specific model. Final assessments for photo aesthetics are made using the updated model, with the high-ranked photos selected for the requesters.

IV. Attaining the Aesthetic Context Actively

To overcome aesthetic uncertainty, CrowdPicker uses mobile crowdsourcing for photo aesthetic annotation and gain contextual knowledge. Given the crowdsourcing budget limitation, it needs to first identify the most useful photos through sampling. For this, we propose a novel aesthetic utility in this part by jointly pursuing knowledge on different score ranges and the difficulties in assessing their aesthetics. Then an algorithm for sampling the best-utility photos is designed.

A. A Novel Aesthetic Utility Measure

Fig. 4 gives an illustrative depiction of our utility measure, where the deep model is the raw model. Intuitively, we claim that raw model's prediction $\widehat{P}_i$, although inaccurate for different crowdsensing contexts, provides valuable relative ranking differences between different photos. For example, assuming we have photos A and B with real scores 2 and 6, their aesthetics are predicted as 4 and 5 (biasedly to a medium level) by the raw model. Obviously, the predictions $\widehat{P}_i(A)$ and $\widehat{P}_i(B)$ are inaccurate, but the predictions' difference is considered a good indicator that the real aesthetic levels of these two photos are different. Such an essential differential property is tested and observed on the three datasets by performing t-test to statistically distinguish real-different&predicted-different photo pairs and real-similar&predicted-different photo pairs. With p-values significantly smaller than 0.05 (0.0019, 0.0074, and 0.0056 for Museum, City, and Campus, respectively), we consider prediction difference effective in picking out aesthetically different photos.

1) Naïve Balance Measure: Based on the above idea, a simple balance measure (NB) investigates the distribution of the mean scores of a photo subset (as shown in Fig. 4(a)). For photo subset $\mathbb{S}$, we can calculate the mean score of each $pho_i \in \mathbb{S}$ as $\widehat{E_i} = \sum_{j=1}^{10} j \cdot \widehat{p_i^j}$. Since E_i may not be an integer, we set 10 score intervals $\{\mathbb{I}_j | \mathbb{I}_j = (j-1, j]\}$ $(j = 1, \ldots, 10)$ together with a counter m_j for each interval and study the distribution of $\widehat{E_i}$ in these intervals. We increase counter m_j by 1 if having $\widehat{E_i} \in \mathbb{I}_j$ for each photo in $\mathbb{S}$. By considering the overall count for each interval, we can then calculate the naïve balance measure as:

$$\mathrm{NB}(\mathbb{S}) = -\sum_{j=1}^{10} \frac{m_j}{|\mathbb{S}|} \cdot \log_2 \frac{m_j}{|\mathbb{S}|} \tag{3}$$

[2]One can simply compare the content similarities between photos to exclude duplications before performing an aesthetic assessment.

where $|\mathbb{S}| = \sum_{j=1}^{10} m_j$ denotes the size of sampled photos (i.e., cardinality) and the $m_j = 0$ terms are ignored.

2) Probabilistic Coverage Balance Measure: We note that, in fact, pho_i's score falls in the range of $(\widehat{E_i} - 1.96\delta_i, \widehat{E_i} + 1.96\delta_i)$ with a probability of 0.95 (i.e., 95% confidence interval), where δ_i is the standard deviation of the raw prediction and is normally larger than 1 in our cases. Such a range will roughly span about 4 intervals we defined in the NB measure (i.e., ≥ 4). As a result, simply using 1 interval to represent the photo score in NB will lead to an inaccurate estimation of the overall sampling distribution.

As a remedy, we jointly exploit the probabilities for every $pho_i \in \mathbb{S}$ to be rated as different scores and propose a probabilistic coverage balance measure (i.e., PCB in Fig. 4(a)). First, in order to differentiate ratings with similar values and identify their contributions to the overall balance performance, an offset embedding step is first performed to calibrate the predictions for each photo. Specifically, we normalize the mean value and calculate the offset amplitude Δ_i for each photo with:

$$\mathrm{N}(\widehat{E_i}) = 1 + \frac{9 \times (\widehat{E_i} - \min(\widehat{E_i}))}{(\max(\widehat{E_i}) - \min(\widehat{E_i}))} \tag{4}$$

$$\text{and} \quad \Delta_i = \widehat{E_i} - \mathrm{N}(\widehat{E_i}). \tag{5}$$

Then, we construct 10 buckets corresponding to the 10 score levels and denote $B_j^{\mathbb{S}}$ as the accumulated depth of the jth bucket (shown in the upper dotted red circle in Fig. 4(a)). Initially, $B_j^{\mathbb{S}}$ is set to 0 for $j \in \{1, \ldots, 10\}$. By going through the normalized prediction distribution $\widehat{P}_i$ for each $pho_i \in \mathbb{S}$, the depth variable for each bucket is updated with:

$$B_j^{\mathbb{S}} = \begin{cases} B_j^{\mathbb{S}} + \widehat{p_i^t}, & t = j + \Delta_i,\ 1 \leq t \leq 10 \\ B_j^{\mathbb{S}}, & \text{otherwise.} \end{cases} \tag{6}$$

That is, each probability component of one raw prediction is added to the depth corresponding to its normalized score level, which can be compared to the process of filling in the buckets with certain amounts of water. A larger B_j indicates that there are probably more photos rated at score j in the sampled subset. Finally, we denote the PCB degree of $\mathbb{S}$ as its expected coverage on all the aesthetic levels, which can be estimated based on the overall distribution of $B_j^{\mathbb{S}}$:

$$\mathrm{PCB}(\mathbb{S}) = -\sum_{j=1}^{10} \frac{B_j^{\mathbb{S}}}{M} \cdot \log_2 \frac{B_j^{\mathbb{S}}}{M} \tag{7}$$

where $M = \sum_{j=1}^{10} B_j^{\mathbb{S}}$ denotes the summation of depths for all the buckets corresponding to $\mathbb{S}$.

3) Hybrid Aesthetic Utility: Among all the raw predictions on a photo collection, some may be precise enough. It is not wise to crowdsource these photos as their aesthetic knowledge may have already been learned by the raw model, that is, they can facilitate limited performance improvement for adaptation. In other words, the raw model provides the buckets with some uneven initial depths, so we should maintain the balance during subsequent photo sampling given such a non-zero start. This requires measuring the difficulty of predicting photos' aesthetics in addition to their PCB.

Such characteristics can be quantified with the commonly used uncertainty measure in active learning, which builds on the observation that a photo with almost identical probabilities at different scores will confuse the estimator. We then denote the uncertainty degree of a sampled photo subset as the raw model's unfamiliar degree on its photos and calculate it as:

$$\mathrm{D}(\mathbb{S}) = -\frac{1}{|\mathbb{S}|} \cdot \sum_{pho_i \in \mathbb{S}} \sum_{j=1}^{10} \widehat{p_i^j} \cdot \log_2(\widehat{p_i^j}). \tag{8}$$

Finally, as shown in Fig. 4(b), given a candidate sampling $\mathbb{S}$ (a subset of the whole photo collection), its *aesthetic utility is denoted as an indicator of the expected performance gain of the raw model if knowing the aesthetic annotations for its photos*. Formally, we estimate aesthetic utility by integrating a subset's probabilistic coverage balance degree (informativeness of balanced aesthetic levels) and uncertainty degree (informativeness of aesthetically unseen samples):

$$\mathrm{U}(\mathbb{S}) = \alpha \cdot \mathrm{PCB}(\mathbb{S}) + (1 - \alpha) \cdot \mathrm{D}(\mathbb{S}) \tag{9}$$

where both factors are in the range of $[0, \log_2 10]$. W.l.o.g, we set weight α to be 0.5 to treat the two factors with equal importance. In practice, different weights can be assigned according to the availability of additional contextual knowledge. For example, we can increase α if the visual domain of the raw model and the target domain is similar.

B. Utility-Based Photo Sampling

1) Problem Formulation: Considering the utility model in (9) and the constraint, the sampling process to find the most context-aware photos for crowdsourcing can be formalized as follows:

Utility-based Sampling Problem: Given a photo collection $\mathbb{A}$, a raw aesthetic assessment model, and a threshold b, the problem asks for a photo subset $\mathbb{S}^* \subseteq \mathbb{A}$ that can maximize its utility $\mathrm{U}(\mathbb{S}^*)$ in terms of the raw model, while satisfying $|\mathbb{S}^*| \leq b$.

Theorem 1: The Utility-based Sampling Problem is NP-hard.

Proof: We prove the NP-hardness by reducing the Maximum Set Coverage problem to a special case of this problem. Detailed proof can be found in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2023.3237969.□

2) An Approximate Algorithm: In order to efficiently find the proper photos for aesthetic annotation, we propose an approximate algorithm to solve the intractable NP-hard problem based on greedy strategy. The pseudo-code is listed in Algorithm 1.

Initially, the algorithm inputs all photos to the raw model to infer their aesthetic predictions and computes the prediction offset and difficulty for each photo. Then it iteratively selects the next best photo by calculating the utility gain each photo candidate can bring. Specifically, in each round, the algorithm examines every photo $pho_i \in \mathbb{A} - \mathbb{S}^*$, estimates the bucket depths $B_j^{\mathbb{S}^* \cup pho_i}$ one can attain if adding pho_i to the current $\mathbb{S}^*$, and computes the corresponding balance performance with $\mathrm{PCB}(\cdot)$ after involving pho_i. The utility gain of a photo, denoted

Algorithm 1: Aesthetic Utility-Based Sampling.

Input: Photo dataset $\mathbb{A}$, crowdsourcing budget b;
Output: A photo subset $\mathbb{S}^*$ for annotation;

$\mathbb{S}^* \leftarrow \emptyset$;
for *each photo* pho_i *in* $\mathbb{A}$ **do**
  Infer the raw prediction $\widehat{\mathrm{P}}$ *and the mean value* $\widehat{E_i}$;
  Compute the offset amplitude Δ_i ;
  Compute the prediction difficulty $\mathrm{D}(pho_i)$;
end
repeat
  for *each photo* pho_i *in* $\mathbb{A} - \mathbb{S}^*$ **do**
    for j *in* $[1, 10]$ **do**
      Compute $B_j^{\mathbb{S}^* \cup pho_i}$ *using* $\widehat{P}_i$ *and* Δ_i ;
    end
    $M \leftarrow \sum_{j=1}^{10} B_j^{\mathbb{S}^* \cup pho_i}$;
    $\mathrm{PCB}(\mathbb{S}^* \cup pho_i) \leftarrow -\sum_{j=1}^{10} \frac{B_j^{\mathbb{S}^* \cup pho_i}}{M} \cdot \log_2 \frac{B_j^{\mathbb{S}^* \cup pho_i}}{M}$;
    $\Delta\mathrm{U}(pho_i) \leftarrow \mathrm{PCB}(\mathbb{S}^* \cup pho_i) + \mathrm{D}(pho_i)$;
  end
  $\mathbb{S}^* \leftarrow \mathbb{S}^* \cup \underset{pho_i}{argmax}\{\Delta\mathrm{U}(pho_i)\}$;
until $|\mathbb{S}^*| > b$ *or* $\mathbb{A} - \mathbb{S}^* = \emptyset$;
return $\mathbb{S}^*$

as $\Delta\mathrm{U}(pho_i)$, is calculated by adding the balance level it yields and its prediction difficulty. Then it selects the photo with the maximum $\Delta\mathrm{U}(pho_i)$ and merges it into $\mathbb{S}^*$, wherein ties are broken by selecting the one with the lowest index. Once selected, a photo is removed from future consideration. The iteration runs until the number constraint is active or all photos have been selected. Its time complexity is $O(nb)$.

Theorem 2: Our algorithm provides a $(1 - 1/e)$ approximation of the optimal solution, where e is the base of the natural logarithm.

Proof: From the proof of Theorem 1, a selection of b subsets (in the MSC problem) implies a valid sampling of b photos, while the hybrid aesthetic utility of photos is maximized when the corresponding subsets cover the maximum number of elements. Moreover, the subset selected by greedily picking the set that covers the most number of uncovered elements can yield a cardinality that is at least $(1 - 1/e)$ times the optimal value according to [47]. Therefore, the utility of the greedily sampled photos is also $(1 - 1/e)$-optimality in the utility lower bound.□

C. Crowdsourcing Aesthetic Annotation

Given the sampled photos, the sampling and crowdsourcing module proceeds to package them into crowdsourcing tasks and assign each task to five different workers. A task starts with a brief description of the corresponding crowdsensing target and contains 10 different photos. Meanwhile, sentinels with obvious aesthetic quality are randomly embedded to filter out unreliable responses. Inspired by the context proximity and sharing aesthetic tendency of mobile users [27], we propose to disseminate the task through social platforms (e.g., WeChat). In this way, CrowdPicker carefully gathers the crowd wisdom on specific crowdsensing contexts.

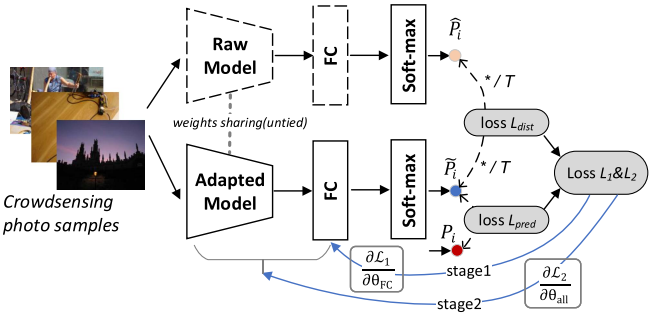


Fig. 5. Two-stage distillation-based model adaptation architecture. Upper: a well-trained raw model; Lower: our adapted model for the current crowdsensing context.

Workers are asked to rate the aesthetics of the assigned photos on a 5-point scale, which will be transformed to a 10-point rating to align with the Raw model using $\frac{(rating-1)}{4} * 10$ (Preston2000). We then use the mean and deviation values from the ratings of a photo to fit a normal distribution based on maximum entropy optimization. Such a fitting operation will implicitly relieve unavoidable discrepancies of the workers from common people, as the possibilities of those scores that are not assigned are also statistically taken into consideration. Photos and their distributions are then used as samples for later adaptation.

V. Photo Selection Via Adaptation

Basic Architecture: The adaptation architecture in the 2nd phase of CrowdPicker is shown in Fig. 5. All the input samples are re-scaled and cropped into the size of 224×224. Basically, the backbone network (MobileNet, VGG, and Inception-BN in this work) is followed by a full connection layer (FC) with 10 neurons. Soft-max activations are used in the end as the classifier to estimate the probabilities of falling in different score buckets. MobileNet is recommended in practice for it shows better training and inference efficiency in mobile crowdsensing scenarios.

Initially, we share the weights of the 27 shallow layers from the Raw model to our adapted model and apply the Lecun uniform initializer for its FC layer. During adaptation, model distillation technique is utilized to learn from the samples' contextual knowledge and to retain common aesthetic knowledge inherited from the Raw model simultaneously. Such a design is expected to make up for the small sample volume in our budget-limited dynamical adaptation cases. We point out that the existing unsupervised adaptation methods are not good choices in the crowdsensing scenarios, as the domain shift (i.e., aesthetic bias) between different crowdsensing tasks can be very large, which will be evaluated in Tables II and V.

Two-Stage Adaptation. We conduct the adaptation in two stages: (i) the FC layer of the adapted model is fine-tuned with the rest layers frozen to make use of the common knowledge for

aesthetic feature extraction; (ii) the whole model are put together for fine-tuning with smaller learning rate. Detailed settings are presented in Section VI-A.

As shown in Fig. 5, CrowdPicker calculates the loss value based on a hard/predication loss $\mathcal{L}_{pred}$ and a soft/distillation loss $\mathcal{L}^j_{dist}$ with j the stage indicator. We adopt the Earth Mover's Distance [25] as the hard loss:

$$\mathcal{L}_{pred} = EMD(\widehat{P}_i, P_i) = \left(\frac{1}{N}\sum_{r=1}^{N}|CDF_{P_i}(r) - CDF_{\widehat{P_i}}(r)|\right)^{\frac{1}{2}}, \quad (10)$$

where $CDF_{\mathrm{P}}(r)$ denotes the cumulative distribution function following the distribution of P. For the soft loss, a temperature factor T_j is introduced for the jth stage on the logit outputs of both the Raw and the adapted models to generate softer probability:

$$q_i^t = \frac{exp(p_i^t/T_j)}{\sum_k exp(p_i^k/T_j)},$$

which denotes the probability of pho_i to be with score t. It is helpful to obtain much of the information about the learned function that resides in the ratios of very small probabilities. Given the predication of the adapted model $\widetilde{P}$, we have:

$$\mathcal{L}^j_{dist} = EMD(softmax(\widetilde{P_i}/T_j), softmax(P_i/T_j)). \quad (11)$$

For each stage, we use a weight to combine these two loss together and present the overall loss function as:

$$\mathcal{L}_j = \beta_j \cdot \mathcal{L}_{pred} + (1-\beta_j)\mathcal{L}^j_{dist},\ j = 1\ or\ 2. \quad (12)$$

Finally, we infer the aesthetics of all the crowdsensing photos with our updated adapted model. The platform can then flexibly choose to retrieve the photos according to the aesthetic ranking in a progressive way or send only the top-K aesthetic-pleasing photos to the requester.

VI. Experiments

The evaluation attempt to answer the following questions: 1) how does CrowdPicker compare to state-of-the-art photo selection methods on resolving aesthetic uncertainty and bias, and improving selection performance? (Section VI-C1); 2) what is the impact of the sampling parameter α (9) in photo selection performance? (Section VI-C2); 3) how does the sampling strategy of CrowdPicker compares to state-of-the-art sampling strategies? (Section VI-C2). After answering these questions, our evaluation studies the time costs (Section VI-C3) and dives into the cross-context performance to further emphasize the design motivation in closing (Section VI-C4).

A. Implementation Details

The proposed CNN models are implemented using TensorFlow. Different from traditional model training that splits a dataset into recommended proportions for training and testing, the sampling for annotation process in CrowdPicker splits datasets into training and test subsets with the proportion of training data bounded by the crowdsourcing budget b during the experiments. To maintain a reasonable crowdsourcing overhead, we set the budget to be smaller than 35 photos, which yields the train-test proportion to be smaller than 1:7. Accordingly, the batch size is set to the number of the sampled/training data in all the experiments, as the amount of training samples is quite small.

TABLE I
Basic Statistics of the Adopted Datasets. Near-Duplicates Have Been Filtered as a Pre-Selection Step Based on Hand-Crafted Visual Similarities

Datasets	Visual Domain	Geographic Scale	# Photos
Museum	Humanity: scene, objects	Inside a museum	251
City	Trip: buildings, resort	Around the CN tower	270
Campus	Campus life: buildings, activities	The Oxford Campus	700

For determining hyper-parameter, we perform grid search with 10-fold cross-validation. Although separate validation is more often used, it is not suitable for the crowdsensing contexts where all photos are with no labels at the outset. That is, given the small number of labeled images attained with crowdsourcing, splitting a held-out subset for validation is counter-productive for with at most 7 photos [48]. Hence, we alternatively use 10-fold validation with the dynamically aesthetic-annotated photos, while using grid search to optimize the learning rate (dense), learning rate (all), learning epochs (dense), learning epochs (all), and dropout in parameter spaces {1e-3, 1e-4}, {3e-5, 3e-6, 3e-7}, {12, 15, 18}, {8, 12, 15}, and {0.1, 0.4, 0.7} for each test case, respectively. Adam optimizer is used to dynamically adjust the learning rates.

For the distillation process, we set the weights β_1 and β_2 for the hard and soft loss to 0.8 and 0.5 for the two stages, and use temperatures 2 and 4 for them, as recommended in [49] for relatively fewer samples cases. The underlying consideration is to *lean more on the target context to build the FC layer, while just slightly calibrating the deep features to retain the common aesthetic knowledge*. All experiments were conducted on a workstation with 2.7 GHz Intel(R) i7 CPU and 64 GB RAM, and model training and inference tasks were executed on a plug-and-play NVIDIA 3080Ti GPU.

B. Setup

1) Datasets: Our experiments need crowdsensing photo collections, which, different from large online photo collections (e.g., ImageNet, COCO), consist of number-constrained photos, sharing strong correlations on the crowdsensing contexts. For example, in a task that gathers photos of some Campus Opening Day, requesters expect elaborated photos explicitly related to this spatial and semantic context.

Given the unavailability of such a dataset, we choose to aggregate three photo datasets using Museum [50], City [51], and Campus [52], to simulate three different crowdsensing tasks. The basic information of the datasets is summarized in Table I. Note that hundreds of photos are sufficient in the

crowdsensing context [14], [18], [19], due to spatial, temporal, and semantic constraints of the tasks. In fact, as indicated in [6], the composition property of crowdsensing tasks determines that one can always decompose a city-scale task and its collections into smaller regions and photo sets for an independent analysis.

To obtain ground truth aesthetic ratings for the photos, we collected annotations from 23 students (12 males, 11 females) in a lab setting, whose controllable environment factors are believed to provide reliability for such subjective tasks. Given a brief description of the crowdsensing context (e.g., task targets, viewer background, scenes), the workers were asked to rate on a five-point scale, like the crowdsourcing task we mentioned above in Section IV-C. Note that the scale of these datasets, although much smaller than AVA due to its prohibitive annotation form of long-term online competition [26], is comparable with typical image aesthetic datasets, e.g., CUHKPQ with 10 raters [53], AADB with 5 raters [24], FLIKR-AES with 5 raters [35], and REAL-CUR with 14 raters [35]. Given the 23 ratings of each photo, we further calculated its rating distribution by fitting it to a normal distribution using maximum entropy optimization, relieving individual differences by extracting statistical common opinions, as suggested in [54].

To test whether the ratings show consistent judgment on aesthetics, we use the Intra-class Correlation Coefficient (ICC) [55] and Spearman's ranking correlation as reliability indicators to analyze annotation reliability. For Museum, City, and Campus, the ICC results are 0.9, 0.928, and 0.937, indicating that most variance can be explained by image differences instead of ratings, while 89%, 93.7%, and 99.6% photos have significant agreement among raters (p-values smaller than FDR level 0.05), respectively. The analysis results show that annotations for the three datasets are reliable for scientific research and, more importantly, their consistency implicitly reflects general judgments on aesthetics.

2) Metrics: Four metrics are tested by comparing to the human rating baseline (i.e., GT scores) in the experiments. In particular, *we use binary classification accuracy ACC and rating deviation DEV to test the aesthetic assessment performance (1) and (2), while using the Spearman coefficient ρ and top-k recall kRec to quantify the aesthetic-aware selection performance.*

- *ACC* measures the coarse (i.e., high or low) aesthetic assessment capacity. A cut-off threshold 0.5 is used, thereby yielding $Acc = \frac{1}{n} \cdot |((\mathbf{R}_P - 5) \odot (\mathbf{R}_G - 5)) > 0|$, wherein $\mathbf{R}_P$ and $\mathbf{R}_G$ correspond to the prediction and GT rating vectors for all the photos, and $\odot$ stands for an element-wise product.
- *DEV* is calculated as the mean averaged error of the assessed scores from the GT mean values, which also reflects the accuracy of score estimation.
- ρ estimates the overall ranking correlation of the assessed and the GT scores as $\rho = 1 - 6\frac{\sum d_i^2}{n^3 - n}$ (d_i denotes the distance in the two ranking positions of the same photo pho_i). ρ lays in the range of $[-1, 1]$, where a larger, positive (negative) value indicates a higher, positive (negative) correlation in two rankings.
- *kRec* denotes the hitting ratio of the first K photos in our ranking in terms of the GT ranking. It is computed as $kRec = \frac{|\mathbb{A}_G^{\{1,..,k\}} \bigcap \mathbb{A}_P^{\{1,..,k\}}|}{k}$, where $\mathbb{A}_G^{\{1,..,k\}}$ and $\mathbb{A}_P^{\{1,..,k\}}$ are the real and predicted top-k photo subset.

Note that ρ and *kRec* are particularly useful since the goal of our framework is to pick out aesthetically pleasing photos from the whole collection and send them to the requesters.

3) Baselines: First, to evaluate the aesthetics selection performance, we introduce both a traditional heuristic photo selection method and some variants of the adaptation-based selection methods for comparison.

- *Cov.* [19] selects crowdsensing photos for certain coverage of the target by finding the photos with largest SIFT similarities to the others.
- *Raw* uses the predictions of different raw models for aesthetic assessment, indicating an no-adaptation method.
- *Scratch* uses the crowdsourcing annotated photos to train a deep model from the scratch (learning rate=0.01, epoch=20).
- *Pseudo* [56] realizes a typical unsupervised adaptation method, which assigns each photo a pseudo-label using $\arg\max_j \widehat{P_i^j}$ based on the raw prediction and then performs training with the Scratch method.

Second, since different sampling strategies will pick out and attain labeled samples with different merits for aesthetic selection, we also investigate the impact of different sampling strategies during crowdsourcing in addition to our aesthetic utility-based design. For this, we adopt the following baselines:

- *Certainty* samples photos with the biggest overall prediction difficulty (i.e., $\mathrm{D}(\mathbb{S})$ in (8)).
- *Balance* chooses the most balanced photo subset with the probabilistic balance measure in (7).
- *Diff* [28], [35] finds the photo subset with the maximum differences, measured by the cross entropy, among its contained photos during sampling. This process is NP-hard (can be reduced from the classical *weighted K-clique problem*). A multi-round sampling is thus performed by choosing the most different photo iteratively, because it can yield a $(1 - 1/e)$ approximation ratio, given its monotone and submodular properties [57].
- *Random* samples photo for annotation randomly.

We use four conventional aesthetic assessment models as raw models, which are *NIMA-aes(MN)* [25] as a MobileNet model trained on AVA [26], *NIMA-tec(MN)* [25] as a MobileNet model trained on TID2013 [58], *NIMA-aes(VGG)* [25] as a VGG16 model trained on AVA [26], and *PAM* [35] as an Inception-BN model trained on FLICKR-AES [35]. CrowdPicker and the baselines (except for Cov.) are tested by directly using its predictions (Raw) or tuning them.

C. Experimental Results

1) Comparisons With Other Selection Methods: Table II presents the comparison results of CrowdPicker and the listed photo selection baselines in terms of aesthetic awareness. Since Cov. provides a selected subset without overall score assessment

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT PHOTO SELECTION METHODS WITH DIFFERENT FOUR RAW MODELS ON THREE DATASETS. FOR SCRATCH AND OUR CROWDPICKER, THE RATIO (NUMBER) OF SAMPLED PHOTOS FOR CROWDSOURCING IS 0.04 (10), 0.04 (11), AND 0.01 (7) FOR MUSEUM, CITY, AND CAMPUS, RESPECTIVELY

Methods		Raw models	Museum				City				Campus			
			ACC	*DEV*	ρ	*kRec*	*ACC*	*DEV*	ρ	*kRec*	*ACC*	*DEV*	ρ	*kRec*
heuristic selection	Cov.[19]	agnostic	-	-	-	0.1	-	-	-	0	-	-	-	0
adaptation -based selection	Raw	NIMA-aes(MN)	0.76	0.71	0.61	0.2	0.58	1.04	0.42	0.1	0.58	1.13	0.39	**0.1**
		NIMA-tec(MN)	0.5	1.02	0.07	0	0.43	1.21	-0.16	0	0.44	1.34	-0.17	0.05
		NIMA-aes(VGG)	0.58	0.90	0.49	0.3	0.45	1.12	0.36	0.3	0.51	1.22	0.35	0.05
		PAM	0.73	0.74	0.64	0.3	0.59	1.17	0.6	0.5	0.62	1.1	0.68	0.3
	Scratch	NIMA-aes(MN)	0.46	1.47	0.08	0.1	0.5	1.44	0.15	0.08	0.46	1.62	0.12	0.03
		NIMA-tec(MN)	0.44	1.69	0.02	0.07	0.48	1.53	-0.01	0.03	0.45	1.73	-0.04	0.03
		NIMA-aes(VGG)	0.55	1.29	0.15	0.3	0.52	1.38	0.12	0.17	0.57	1.54	0.15	0.06
		PAM	0.57	0.98	0.01	0.1	0.46	1.33	0.04	0.05	0.46	1.41	0.01	0.04
	Pseudo[56]	NIMA-aes(MN)	0.52	0.82	0.5	0.2	0.51	1.02	0.25	**0.4**	0.47	1.17	0.23	0.05
		NIMA-tec(MN)	0.5	2.21	-0.05	0	0.52	2.23	-0.09	0	0.52	2.27	-0.17	0.05
		NIMA-aes(VGG)	0.57	0.89	0.34	0.3	0.43	1.03	0.3	0.4	0.53	1.18	0.29	0.05
		PAM	0.73	3.45	0.52	0.2	**0.69**	3.74	0.49	**0.6**	**0.71**	3.53	0.6	**0.5**
	Ours	NIMA-aes(MN)	**0.78**	**0.67**	**0.69**	**0.4**	**0.69**	**1**	**0.52**	0.24	**0.62**	**1.1**	**0.4**	0.09
		NIMA-tec(MN)	**0.64**	**0.91**	**0.27**	**0.2**	**0.62**	**1.12**	**0.24**	**0.06**	**0.59**	**1.29**	**0.22**	**0.18**
		NIMA-aes(VGG)	**0.73**	**0.67**	**0.66**	**0.4**	**0.67**	**0.89**	**0.47**	**0.48**	**0.66**	**1.06**	**0.46**	**0.17**
		PAM	**0.74**	**0.67**	**0.68**	**0.34**	0.61	**1**	**0.64**	0.53	**0.71**	**0.96**	**0.72**	0.31

and ranking, only the top-k performance is tested. Here we use small sampling ratios of 0.04, 0.04, and 0.01 for the three datasets to explicitly set a strict budget (around 10 photos for annotation).

Generally, by adapting to the context with two-stage distillation, CrowdPicker improves (reduces) the assessment accuracy (deviation) significantly and takes care of photos in different score ranges, which benefits its ranking performance. Overall, for the 12 test cases on different raw models and different datasets, CrowdPicker outperforms both the heuristic and adaptation-based baselines, with 3 exceptions caused by Pseudo. For the special case where Pseudo presents larger *kRec*, it is caused by a useless ranking result with too many ties (the predictions of Pseudo concentrates on a quite small range of 0.5-3). Meanwhile, as expected, all the adaptation-based selection methods satisfy the aesthetic preference better (larger *kRec*) than the heuristic selection, because the latter ignores the subjective experiences in design.

For different raw models, NIMA-aes(VGG) and PAM tend to facilitate the best performance for each method on every dataset, while NIMA-tec(MN) presents an obviously worse performance in each case. We owe this to the joint impact of the backbone and the training dataset, namely, advanced backbones (e.g., VGG and Inception-BN) could better encode aesthetic perception and samples, especially from daily lives (e.g., FLICKR-AES), can provide more generic knowledge during adaptation. It is worth noting that CrowdPicker successfully improves ranks from negative correlation to positive correlation in City and Campus under raw model NIMA-tec(mn).

Last but not least, the DEV performance advantage of CrowdPicker over Raw demonstrates that it effectively addresses the aesthetic uncertainty concerns; while by providing similar *ACC* and higher *kRec* compared with the reported performance for the Raw model on AVA (around 0.8 and 0.6), CrowdPicker is also believed to have significantly relieved the aesthetic bias. We present some photo aesthetic prediction examples in Appendix B, available in the online supplemental material to better illustrate the performance superiority.

2) Comparisons on Different Sampling Strategies: To evaluate the effectiveness of utility-based sampling, we conduct *ablation study* with the Certainty (i.e., α=0) and Balance (i.e., α=1) methods, and also compare its performance with a SoTA Diff method and a simple Random strategy. The tests are conducted with four raw models on three datasets.

Ranking Performance: The ranking correlations of different methods under varying crowdsourcing ratio settings are presented in Fig. 6. All methods' performance experience slight improvement with an increasing number of sampled photos. Except for NIMA-tec(MN) with the worst learning basis, our sampling strategy is shown to provide generally better or similar good aesthetic ranking correlation than others. Under NIMA-tec(MN), CrowdPicker requires a higher sampling ratio to attain performance superiority, so we note that a weaker raw model is still useful when giving sufficient adaptation knowledge. We notice that Diff and CrowdPicker have similar performance on PAM, because Raw of PAM has already provided sufficiently high correlations (> 0.6). In this case, certainty-oriented adaptation may result in degraded performance, even worse than Random, as the sampled uncertain photos for PAM could have ambiguous annotations.

Photo Selection Performance: We emphasize that the ultimate goal of CrowdPicker is to select some aesthetically representative photos from the collected contents for the requester, so $kRec$ is a more straightforward performance indicator. W.l.o.g, we calculate the averaged top-K recall of different sampling methods using $K = 10, 20, 30$ separately and present the results in Fig. 7. One can observe more obvious performance gaps

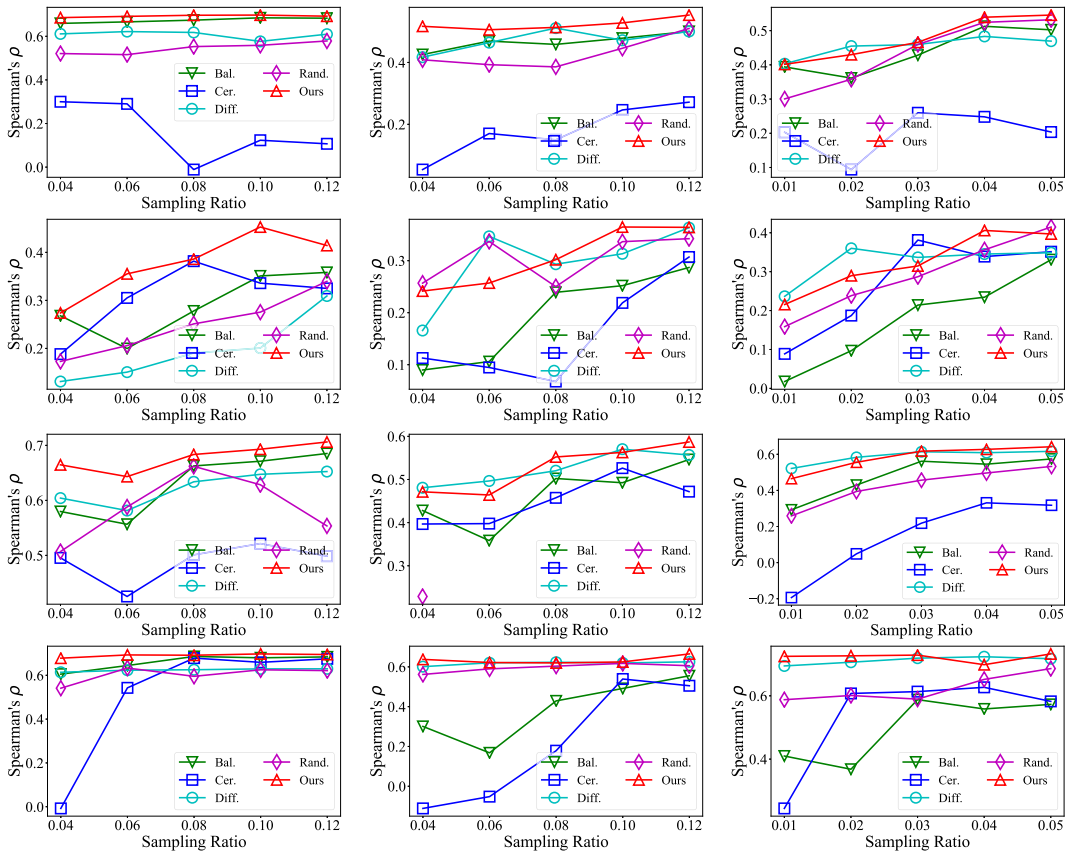


Fig. 6. Performance on the ranking correlation for the assessed aesthetic scores of different sampling strategies (Row 1-4: with raw model NIMA-aes(MN), NIMA-tec(MN), NIMA-aes(VGG), and PAM; Column 1-3: Dataset Museum, City, and Campus.).

TABLE III
PHOTO SELECTION EFFICIENCY ANALYSIS (S) ON CAMPUS

Raw models	Cov.	Raw	Pseudo (adap., inf.)	Scratch (samp., adap., inf.)	Ours (samp., adap., inf.)
NIMA-aes(MN)	1493.73	7.2	(187.4, 7.2)	(1.3, 18.45, 7.2)	(1.3, 19.15, 7.2)
NIMA-tec(MN)		7.3	(188.1, 7.3)	(1.3, 15.64, 7.3)	(1.3, 20.66, 7.3)
NIMA-aes(VGG)		6.8	(179.4, 6.8)	(1.3, 88.31, 6.8)	(1.3, 99.91, 6.8)
PAM		12.6	(390.4, 12.6)	(1.3, 65.99, 12.6)	(1.3, 84.49, 12.6)

TABLE IV
SAMPLING TIME COST STATISTICS (S)

Datasets	Bal.	Cer.	Diff.	Rand.	Ours
Museum	0.28	0.002	0.92	<0.001	0.41
City	0.35	0.002	1.05	<0.001	0.42
Campus	0.75	0.01	2	<0.001	1.3

among methods than the correlation tests, where CrowdPicker showcases the best top-K performance in almost every case. Specifically, the Certainty and Random methods have fluctuated performance as they adopt myopic sampling strategies, while the performance of other methods is either steady or increasing slowly with the sampling ratio. Remarkably, we find that the selection performance is impressive even with small sampling ratios (i.e., around 10 annotated samples for 10% hit rate), which demonstrates the practical value of this framework.

3) Time Cost Analysis: We analyze the efficiency of different photo selection methods with detailed time overhead composition in Table III (only the result on Campus is presented here for space limitation). Coverage is agnostic to the raw models and requires the most time for SIFT feature extraction and complex matching operations. The time cost of Raw is the inference time of different raw models. As shown, VGG is the most efficient one in making real-time predictions. The adaptation (a.k.a., fine-tuning) time of Pseudo is much longer than Scratch and CrowdPicker, which two has similar time costs, for it trains on every sample during this stage. For CrowdPicker, the time cost on those better-performance raw models (e.g., PAM) is 4-5 times higher than their counterparts (e.g., NIMA-tec(MN)), which reminds the overhead of tuning selection performance. We also test the required time for sampling photos to get annotation. As shown in Table IV, each strategy takes an acceptable time for choosing samples, wherein CrowdPicker's time consumption is relatively longer than that of Balance and Certainty. We also note that sampling time on larger datasets (e.g., Campus) is longer. Finally, we defer the analysis of the trade-off between crowdsourcing overhead (e.g., response latency) and selection performance to future work.

4) Cross-Context Evaluation: As discussed in Section III-B, aesthetic bias hinders the generalizability of photo selection

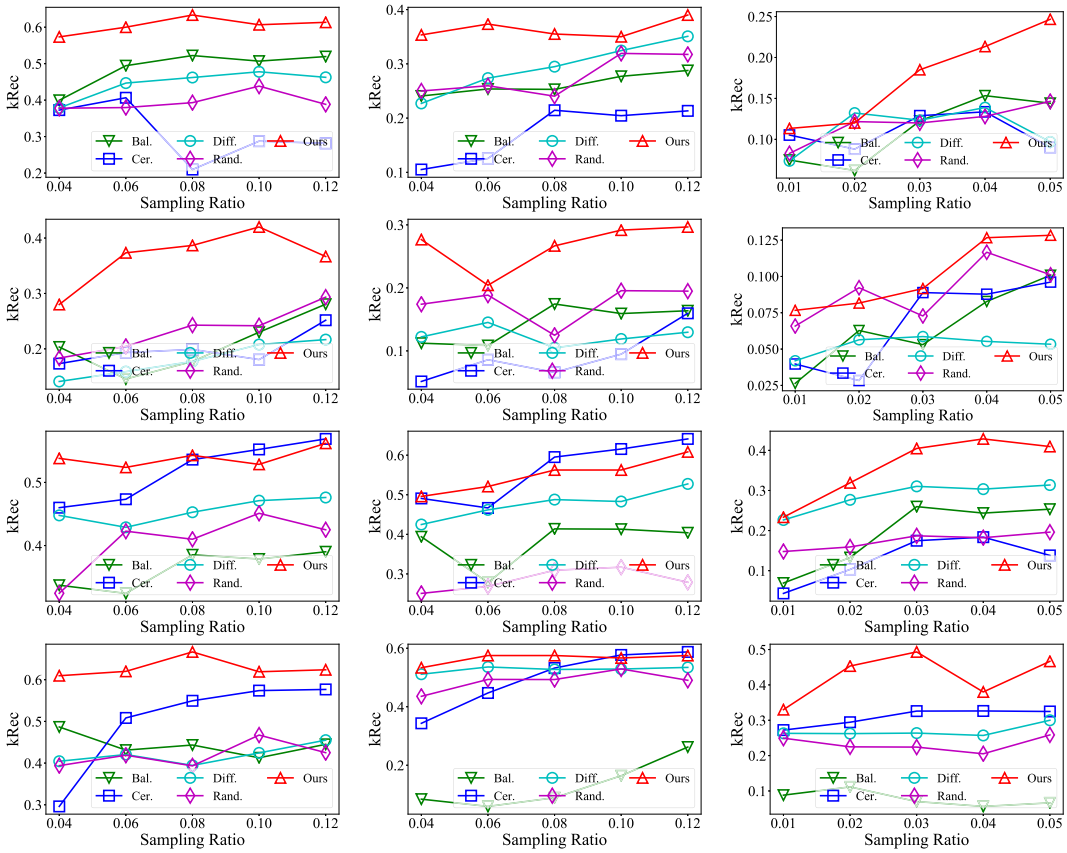


Fig. 7. Performance on the averaged top-K recall for the assessed aesthetic scores of different sampling strategies (Row 1-4: with raw model NIMA-aes(MN), NIMA-tec(MN), NIMA-aes(VGG), and PAM; Column 1-3: Dataset Museum, City, and Campus.).

TABLE V
EVALUATION ON THE CROSS-CONTEXT/DATASET PERFORMANCE. THE RESULTS ALSO SHOWCASE AESTHETIC BIAS

(ρ, *kRec*)		Target contexts		
		Museum	City	Campus
Source contexts	Museum	**(0.69, 0.55)**	(0.43, 0.16)	(0.44, 0.09)
	City	(0.13, 0.05)	**(0.55, 0.35)**	(0.23, 0.1)
	Campus	(0.66, 0.42)	(0.39, 0.3)	**(0.55, 0.19)**

* Ratios of the sampled photos are 0.1, 0.1, and 0.04.

under different crowdsensing contexts. To gain more insights into this motivation, we evaluate whether models tailored to one dataset perform well on the others. Table V provides an illustrative comparison of the cross-context performance under raw model NIMA-aes(MN). Remarkably, we observe that each cross-context pair presents very limited "transferability," which owes to the different visual domains and feature distributions. These results further demonstrate our initial observation that aesthetic bias widely exists in crowdsensing tasks and significantly impacts the selection performance.

VII. DISCUSSIONS

Not a Replacement: We propose CrowdPicker to fill the void of subjective perception-driven photo selection, which works as a complement to existing representativeness-driven solutions, but not a replacement. The importance of aesthetics and representativeness criteria varies from scene to scene. For example, one would prefer the former for user-contextual crowdsensing or social crowdsensing (e.g., design inspiration, tour planning [10], daily sharing [9]), favor the latter in event crowdsensing [8] (pictures from more angles provide more information of the events), while caring about both in sensing for urban expression and profiling [59]. To accommodate both kinds of criteria, a feasible suggestion is to enlarge the selection amount and conduct selection for aesthetics and representativeness independently and to present the photos that are either in two selections' intersection set or top-ranked in the selection subsets to the requesters.

Generalization versus Personalization: We focus on inferring the generalized aesthetic preference of common viewers on crowdsensing photos, which may have conflicts with the various tastes of individuals. In particular, the aesthetic ground truth and crowdsourcing annotations are all fitted to normal distributions for resolving the discrepancy. Using such annotated samples for adaptation will tune the aesthetic assessments to the common preference of a virtual community, inevitably conflicting with customization. We point out that the PIAA technique (e.g., [35], [36], [37]), as mentioned in Sec. II, can be taken as a remedy. Among them, user interaction is believed to be the most

promising one for selecting personalized crowdsensing photos, because crowdsensing are online tasks which are featured with and built on human involvement. In contrast with performing complex re-ranking-based interaction in [34], we consider that progressive adaptation (i.e., sending a small batch and obtaining user preference on photos in them) with user making binary judgement (i.e., favored or not) is more suitable for server-user interaction in crowdsensing and also more user-friendly.

Limitations: Although the current scale of user study is sufficient to understand the importance of aesthetic awareness in photo selection, it is quite small to capture the diverse aesthetic preferences of different user groups. Meanwhile, the crowdsourcing tasks didn't consider workers' background (e.g., expertise) differences in certain crowdsensing contexts [60]. As a result, the adaptation may be performed for a specific user group based on aesthetic knowledge learned from workers unfamiliar with the context, unfortunately incurring aesthetic bias during selection. We consider this an open issue for further improving the selection performance.

VIII. Conclusion

This work identifies the limitations of existing crowdsensing photo selection techniques on guaranteeing the subjective viewer preference with a user survey. To fill the void, a novel photo selection framework is designed by exploiting mobile crowdsourcing and domain adaptation to realize aesthetic awareness. The adaptivity issue of such a framework is carefully investigated with an aesthetic utility measuring the merits of photos in improving the adaption performance. We formalize the best-utility sampling and crowdsourcing problem, prove its NP-hardness, and provide an approximate solution. A distillation-based adaptation architecture is further designed with contextual and common knowledge jointly considered. We use three datasets to conduct experiments, whose results demonstrate both the effectiveness of the design, compared with 8 baselines, in terms of classification, ranking, and retrieval. For future work, we plan to study the trade-off between crowdsourcing overhead (e.g., response latency) and selection performance, and investigate the possibility of overcoming aesthetic uncertainty with a crowdsensing context classifier.

References

[1] A. Guo, A. Jain, S. Ghose, G. Laput, C. Harrison, and J. P. Bigham, "Crowd-ai camera sensing in the real world," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, pp. 1–20, 2018.

[2] E. Wang, Y. Yang, J. Wu, K. Lou, D. Luan, and H. Wang, "User recruitment system for efficient photo collection in mobile crowdsensing," *IEEE Trans. Human-Mach. Syst.*, vol. 50, no. 1, pp. 1–12, Feb. 2020.

[3] L. Wang, D. Zhang, Y. Wang, C. Chen, X. Han, and A. M'hamed, "Sparse mobile crowdsensing: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 161–167, Jul. 2016.

[4] Z. Yu, H. Ma, B. Guo, and Z. Yang, "Crowdsensing 2.0," *Commun. ACM*, vol. 64, no. 11, pp. 76–80, 2021.

[5] M. T. Rashid, D. Zhang, and D. Wang, "SocialDrone: An integrated social media and drone sensing system for reliable disaster response," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 218–227.

[6] Y. Shen, Y. Xu, and L. Liu, "Crowd-sourced city images: Decoding multidimensional interaction between imagery elements with volunteered photos," *ISPRS Int. J. Geo Inf.*, vol. 10, no. 11, 2021, Art. no. 740.

[7] B. Guo et al., "CrowdStory: Fine-grained event storyline generation by fusion of multi-modal crowdsourced data," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 1, pp. 55:1–55:19, 2017.

[8] Z. Xu et al., "Crowdsourcing based description of urban emergency events using social media Big Data," *IEEE Trans. Cloud Comput.*, vol. 8, no. 2, pp. 387–397, Second Quarter 2016.

[9] J.-I. Biel, N. Martin, D. Labbe, and D. Gatica-Perez, "Bites'n'Bits: Inferring eating behavior from contextual mobile data," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–33, 2018.

[10] X. Wang et al., "A picture is worth a thousand words: Share your real-time view on the road," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 2902–2914, Apr. 2017.

[11] A. R. Bahrehdar, B. Adams, and R. S. Purves, "Streets of london: Using flickr and openstreetmap to build an interactive image of the city," *Comput., Environ. Urban Syst.*, vol. 84, 2020, Art. no. 101524.

[12] Beautiful China, quanjingke, 2021. [Online]. Available: http://www.quanjingke.com/dest/

[13] H. Chen, B. Guo, Z. Yu, L. Chen, and X. Ma, "A generic framework for constraint-driven data selection in mobile crowd photographing," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 284–296, Feb. 2017.

[14] H. Chen, B. Guo, Z. Yu, and H. Qi, "Toward real-time and cooperative mobile visual sensing and sharing," in *Proc. IEEE 35th Annu. Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[15] B. Guo, H. Chen, Z. Yu, X. Xie, S. Huangfu, and D. Zhang, "FlierMeet: A mobile crowdsensing system for cross-space public information reposting, tagging, and sharing," *IEEE Trans. Mobile Comput.*, vol. 14, no. 10, pp. 2020–2033, Oct. 2015.

[16] Y. Hua, W. He, X. Liu, and D. Feng, "SmartEye: Real-time and efficient cloud image sharing for disaster environments," in *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 1616–1624.

[17] Y. Jiang, X. Xu, P. Terlecky, T. Abdelzaher, A. Bar-Noy, and R. Govindan, "MediaScope: Selective on-demand media retrieval from mobile devices," in *Proc. IEEE/ACM Int. Conf. Inf. Process. Sensor Netw.*, 2013, pp. 289–300.

[18] Y. Wu, Y. Wang, and G. Cao, "Photo crowdsourcing for area coverage in resource constrained environments," in *Proc. IEEE Conf. Comput. Commun.*, 2017, pp. 1–9.

[19] T. Zhou, B. Xiao, Z. Cai, and M. Xu, "A utility model for photo selection in mobile crowdsensing," *IEEE Trans. Mobile Comput.*, vol. 20, no. 1, pp. 48–62, Jan. 2021.

[20] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1338–1350, Jul. 2016.

[21] K. Lynch, *The Image of the City*. vol. 11. Cambridge, MA, USA: MIT Press, 1960.

[22] W.-T. Sun, T.-H. Chao, Y. Kuo, and W. Hsu, "Photo filter recommendation by category-aware aesthetic learning," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1870–1880, Aug. 2017.

[23] H.-J. Lee, K. Hong, H. Kang, and S. Lee, "Photo aesthetics analysis via DCNN feature encoding," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1921–1932, Aug. 2017.

[24] S. Kong, X. Shen, Z. L. Lin, R. Mech, and C. C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 662–679.

[25] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[26] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2408–2415.

[27] V.-A. Darvariu, L. Convertino, A. Mehrotra, and M. Musolesi, "Quantifying the relationships between everyday objects and emotional states through deep learning based image analysis using smartphones," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–21, 2020.

[28] Q. Xu and R. Zheng, "When data acquisition meets data analytics: A distributed active learning framework for optimal budgeted mobile crowdsensing," in *Proc. IEEE Conf. Comput. Commun.*, 2017, pp. 1–9.

[29] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[30] E. Siahaan, A. Hanjalic, and J. Redi, "Semantic-aware blind image quality assessment," *Signal Process Image Commun.*, vol. 60, pp. 237–252, 2018.

[31] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, Nov. 2015.

[32] K. Karlsson, W. Jiang, and D.-Q. Zhang, "Mobile photo album management with multiscale timeline," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1061–1064.
[33] C. Cui, W. Yang, C. Shi, M. Wang, X. Nie, and Y. Yin, "Personalized image quality assessment with social-sensed aesthetic preference," *Inf. Sci.*, vol. 512, pp. 780–794, 2020.
[34] P. Lv et al., "USAR: An interactive user-specific aesthetic ranking framework for images," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1328–1336.
[35] J. Ren, X. Shen, Z. L. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 638–647.
[36] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multitask learning for generic and personalized image aesthetics assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 3898–3910, Jan. 2020.
[37] H. Zhu, Y. Zhou, L. Li, Y. Li, and Y. Guo, "Learning personalized image aesthetics from subjective and objective attributes," *IEEE Trans. Multimedia*, vol. 25, pp. 179–190, 2021.
[38] H. Zhu, L. Li, J. Wu, S. Zhao, G. Ding, and G. Shi, "Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization," *IEEE Trans. Cybern.*, vol. 52, no. 3, pp. 1798–1811, Mar. 2022.
[39] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, "Crowdsourced data management: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2296–2319, Sep. 2016.
[40] J. Fan, M. Lu, B. C. Ooi, W.-C. Tan, and M. Zhang, "A hybrid machine-crowdsourcing system for matching web tables," in *Proc. IEEE 30th Int. Conf. Data Eng.*, 2014, pp. 976–987.
[41] B. Guo, C. Chen, D. Zhang, Z. Yu, and A. Chin, "Mobile crowd sensing and computing: When participatory sensing meets participatory social media," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 131–137, Feb. 2016.
[42] T. Zhou, Z. Cai, and F. Liu, "The crowd wisdom for location privacy of crowdsensing photos: Spear or shield?," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 3, pp. 1–23, 2021.
[43] P. Cheng, X. Lian, L. Chen, J. Han, and J. Zhao, "Task assignment on multi-skill oriented spatial crowdsourcing," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2201–2215, Aug. 2016.
[44] S. Qiu, A. Bozzon, M. V. Birk, and U. Gadiraju, "Using worker avatars to improve microtask crowdsourcing," *Proc. ACM Hum.- Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–28, 2021.
[45] D. Quercia, N. O'Hare, and H. Cramer, "Aesthetic capital: What makes london look beautiful, quiet, and happy?," in *Proc. 17th ACM Conf. Comput. Supported Cooperative Work Social Comput.*, 2014, pp. 945–955.
[46] T. Walber, A. Scherp, and S. Staab, "Smart photo selection: Interpret gaze as personal interest," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2014, pp. 2065–2074.
[47] A. Mcgregor and H. T. Vu, "Better streaming algorithms for the maximum coverage problem," *Theory Comput. Syst.*, vol. 63, pp. 1–25, 2018.
[48] J. Zhu, H. Wang, E. H. Hovy, and M. Y. Ma, "Confidence-based stopping criteria for active learning for data annotation," *ACM Trans. Speech Lang. Process.*, vol. 6, pp. 3:1–3:24, 2010.
[49] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[50] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3485–3492.
[51] Y. Wang, Z. L. Lin, X. Shen, R. Mech, G. Miller, and G. Cottrell, "Event-specific image importance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4810–4819.
[52] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[53] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2206–2213.
[54] K. Li, X. Zhang, and G. Li, "A rating-ranking method for crowdsourced top-K computation," in *Proc. Int. Conf. Manage. Data*, 2018, pp. 975–990.
[55] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tuts. Quantitative Methods Psychol.*, vol. 8, no. 1, 2012, Art. no. 23.
[56] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 945–954.
[57] A. Krause and D. Golovin, "Submodular function maximization," *Tractability*, vol. 3, pp. 71–104, 2014.
[58] N. N. Ponomarenko et al., "Color image database TID2013: Peculiarities and preliminary results," in *Proc. IEEE Eur. Workshop Vis. Inf. Process.*, 2013, pp. 106–111.
[59] M. R. Ibrahim, J. Haworth, and T. Cheng, "Understanding cities with machine eyes: A review of deep computer vision in urban analytics," *Cities*, vol. 96, 2020, Art. no. 102481.
[60] X. Gao, H. Huang, C. Liu, F. Wu, and G. Chen, "Quality inference based task assignment in mobile crowdsensing," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 10, pp. 3410–3423, Oct. 2021.

Tongqing Zhou received the BS, MS, and PhD degrees in computer science and technology from the National University of Defense Technology (NUDT), Changsha, in 2012, 2014, and 2018, respectively. He is currently an assistant researcher with the College of Computer, NUDT. His main research interests include crowdsensing and data privacy. He is the recipient of Outstanding PhD Dissertation Award and Outstanding Postdoc, both of Hunan Province, China.

Zhiping Cai (Member, IEEE) received the BS, MS, and PhD degrees in computer science and technology from the National University of Defense Technology (NUDT), Changsha, in 1996, 2002, and 2005, respectively. He is currently a full professor with the College of Computer, NUDT. His current research interests include network security and Big Data. He is a member ACM, and a senior member of the China Computer Federation. His Ph.D. dissertation received the Outstanding Dissertation Award of the Chinese PLA.

Fang Liu received the BS and PhD degrees in computer science from the College of Computer, National University of Defense Technology (NUDT), Changsha, China, in 1999 and 2005, respectively. She is currently a professor with the School of Design, Hunan University, Changsha, China. Her current research interests include distributed computing and Big Data.

Jinshu Su (Senior Member, IEEE) received the BS degree in mathematics from Nankai University, Tianjin, China, in 1985, and the MS and PhD degrees in computer science from the National University of Defense Technology, Changsha, China, in 1988 and 2000, respectively. He is a professor with the College of Computer, National University of Defense Technology. He currently leads the Distributed Computing and High Performance Router Laboratory and the Computer Networks and Information Security Laboratory, which are both key laboratories of National 211 and 985 Projects, China. He also leads the High Performance Computer Networks Laboratory, which is a key laboratory of Hunan, China. His current research interests include Internet architecture, Internet routing, security, and wireless networks.