The Challenge of Variable Effort Crowdsourcing and How Visible Gold Can Help

DANULA HETTIACHCHI*, School of Computing Technologies, RMIT University, Australia
MIKE SCHAEKERMANN, Amazon, Canada
TRISTAN MCKINNEY, Amazon, USA
MATTHEW LEASE, Amazon, USA and School of Information, University of Texas at Austin, USA

We consider a class of *variable effort* human annotation tasks in which the number of labels required per item can greatly vary (e.g., finding all faces in an image, named entities in a text, bird calls in an audio recording, etc.). In such tasks, some items require far more effort than others to annotate. Furthermore, the per-item annotation effort is not known until after each item is annotated since determining the number of labels required is an implicit part of the annotation task itself. On an image bounding-box task with crowdsourced annotators, we show that annotator accuracy and recall consistently drop as effort increases. We hypothesize reasons for this drop and investigate a set of approaches to counteract it. Firstly, we benchmark on this task a set of general *best-practice* methods for quality crowdsourcing. Notably, only one of these methods actually improves quality: the use of *visible gold* questions that provide periodic feedback to workers on their accuracy as they work. Given these promising results, we then investigate and evaluate variants of the visible gold approach, yielding further improvement. Final results show a 7% improvement in bounding-box accuracy over the baseline. We discuss the generality of the visible gold approach and promising directions for future research.

1 INTRODUCTION

Annotations (aka labels) provide the basis for training and testing supervised learning models. Consequently, ensuring the quality of annotations is important, especially in a crowdsourced setting with remote, inexpert annotators. While quality assurance for crowdsourcing is well-studied, relatively little work has studied *variable effort* annotation tasks in which the number of labels required per item can greatly vary. Examples might include labeling all faces in an image, named entities in a text, or bird calls in an audio recording. Because the number of instances to label per item can greatly vary, some items require far more effort than others to annotate. Moreover, because there is typically no natural upper-bound on the number of instances present, some individual items may require enormous effort. Finally, the annotation effort required for each item is not known until after it is annotated since determining the number of labels required is an implicit part of the annotation task itself.

In this paper, we first conceptualize the notion of variable effort tasks and how they differ from more typical annotation tasks. For example, such annotation tasks are implicitly two-step: searching the item for all instances matching a target type (e.g., "face"), then applying a labeling operation (e.g., bounding box) to each matching instance. With labeling effort proportionate to the size of search results, the variable size of search results is the key challenge. This framing also helps us relate variable effort tasks to a wider class of annotation search tasks [46].

Next, we empirically investigate the specific variable effort labeling task of *object detection*: finding and localizing human faces in Open Images [47] (via bounding boxes). Whereas many prior studies

---

Authors' addresses: Danula Hettiachchi, danula.hettiachchi@rmit.edu.au, School of Computing Technologies, RMIT University, Melbourne, Australia; Mike Schaekermann, mschaeke@amazon.com, Amazon, Toronto, Ontario, Canada; Tristan McKinney, tristamc@amazon.com, Amazon, Palo Alto, California, USA; Matthew Lease, ml@utexas.edu, Amazon, Seattle, Washington, USA, School of Information, University of Texas at Austin, Austin, Texas, USA.

on object detection report results on simpler datasets having only a few objects per image, our dataset includes as many as 14 faces per image. Our results show that crowdsourced annotator accuracy and recall on Mechanical Turk (MTurk) drops markedly as the number of faces per image increases. We hypothesize a set of key underlying issues leading contributing to this reduced quality: inconsistency of worker experience, the potential for high cognitive load, and ineffective incentive design.

To address these issues, we adapt and assess a set of general *best-practice* methods for quality crowdsourcing: financial incentives, workflow design, and visible gold (i.e., questions that provide periodic feedback to workers on their accuracy as they work). We implement five specific approaches: variable pay per instance, post-task bonuses, task decomposition, iterative improvement, and in-task visible gold with uniform frequency. Notably, only visible gold improves quality.

Motivated by this finding, we further explore the design space for effective use of visible gold questions in variable effort labeling tasks. While prior work shows that visible gold can improve data quality [24, 48], many questions remain. How should we present feedback for variable effort labeling tasks like object detection? What is the optimal strategy to issue visible gold questions? How can the effect of visible gold be strengthened by quality-related consequences (i.e., warnings and bonuses)? We explore different variants of visible gold task designs. We find that combining both upfront and regular testing sustains data quality significantly better than upfront or regular testing alone. Moreover, imposing quality-related consequences yields further improvement. Our final variant of visible gold integrates dynamic testing with tier-based consequences and significantly outperforms all other approaches.

Contributions. We make three primary contributions in this work:

(1) We conceptualize a class of *variable effort* human annotation tasks. We identify a unique set of data quality challenges they present, along with an empirical analysis of these challenges in the context of object detection.
(2) We *systematically evaluate existing methods* to address these challenges and show that providing in-task feedback through *visible gold* significantly outperforms various other baselines, including approaches that adjust pay according to effort or that standardize effort at constant pay.
(3) We contribute an in-depth analysis of different visible gold variants investigating *issuance patterns* and *consequences* for workers. Based on these investigations, we propose and evaluate an *improved visible gold design* that significantly increases bounding box accuracy by 5.7% compared to a basic visible gold variant and by 7.5% compared to a baseline without visible gold.

2 RELATED WORK

2.1 Financial Incentives and Crowd Work

Financial incentives can influence work in various ways: who chooses to accept work, how much work they perform, and the quality of work they produce. Vaughan [76] presents a valuable, succinct review of related work in this area. Early work suggested quality was not impacted by payment [9, 27, 55]. In some cases [9, 27], the difference in payments may have been too low to influence behavior. Mason and Watts [55] hypothesized an *anchoring effect*, with workers' sense of fair payment anchored by whatever was offered. Ipeirotis [36] reports a similar finding.

Whereas the early studies used crowdsourcing tasks that were relatively easy to perform, Ho et al. [32] and Ye et al. [83] instead studied "effort-responsive tasks" in which workers could improve output via more time or effort, and did see the quality improve with financial incentives. Yin and Chen [85] find that while very engaged or un-engaged workers appear insensitive to price, more middling workers improve quality with financial incentives.

Horton and Chilton [33] frame the issue wrt. the economics notion of *reservation wage*: "…the minimum wage a worker is willing to accept …for performing some task; it is the key parameter in models of labor supply." Thus as pay decreases, it could fail to match more workers' reservation

thresholds and thus potentially bias the sample of workers who choose to perform the task. However, Horton and Chilton find mixed evidence for worker behavior conforming to predictions of the rational model: "workers are clearly sensitive to price but insensitive to variations in the amount of time it takes to complete a task."

While MTurk's pay-per-task pricing model is familiar in crowdsourcing research, this model encourages work efficiency, but risks rushed work since worker earnings can be increased by completing more tasks in less time. An alternative pricing model is hourly pay. Both Mankar et al. [54] and Whiting et al. [80] proposed technical approaches making it easier for requesters to offer hourly pay jobs on MTurk. Some commercial vendor workforces[1] also set fixed hourly pay rates. For example, Amazon SageMaker GroundTruth's popular vendor iMerit[2] charges $6.12/hour per worker. While hourly pay has the potential to discourage rushed work, since all time worked is compensated, a vendor workforce may still operate internally on a call-center model, where workers may have productivity quotas that similarly encourage them to work efficiently (to enable the vendor to provide competitive pricing and ensure profitability).

2.2 Workflow Designs

A crowdsourcing task workflow defines how a single task is organized into a set of HITs that can be completed by one or more workers. Literature shows that we can improve the data quality by adopting a suitable workflow for each task. In this work, we are particularly interested in workflow designs that can be used to standardize the task effort in variable effort tasks.

Prior work highlights two main workflow paradigms, iterative and parallel [26]. In *iterative* workflows, we present the same task to multiple workers in a sequential manner where workers could see previous workers' responses. Little et al. [50] shows that an iterative workflow can improve the average data quality in writing and brainstorming tasks. However, the paper highlights that work produced through parallel workflows could still yield individual responses with higher quality. In a translation task, Ambati et al. [4] shows that a 3-phased iterative workflow can achieve higher quality than a baseline that gathers individual translations from 5 workers for each item. Iterative workflows can also allow us to engage workers with different expertise levels at each iteration [4].

Parallel workflows aim to get multiple workers to work on parts of the task at the same time [26]. Parallel work can be on the same task unit (*i.e.,* obtaining multiple answers for the same unit) or smaller sub-tasks obtained through task decomposition. Find-Fix-Verify [6] is a specific workflow pattern that facilitates task decomposition through the initial find step and works well for writing tasks such as proofreading, formatting, and shortening text [6]. Prior work by Kittur et al. [43] proposes a framework for decomposing complex crowd tasks. It shows that in a writing task, articles produced through task decomposition received higher ratings and had lower variability than individual-produced articles. Recent work has also investigated how to optimally decompose a task into atomic sub-tasks considering the desired reliability and cost [75].

Often a combination of iterative and parallel elements can be used to create a more versatile workflow. Other notable work includes tools that can help visualize and manage complex crowdsourcing workflows [42], allow workers to create workflows [45], and optimize workflows [15].

2.3 Gold Standard Questions

The use of gold standard questions (also known as control or gold questions) is a fundamental and widely used quality control mechanism in crowdsourcing [17]. By injecting gold standard questions and evaluating responses, requesters can accurately measure worker performance [34].

---

[1]https://aws.amazon.com/sagemaker/groundtruth/pricing/
[2]https://aws.amazon.com/marketplace/pp/B07DK37Q32

Prior research has investigated how to include gold standard questions within a crowdsourcing job in a systematic way. Liu et al. [52] predict the optimum number of gold questions to include for estimation tasks such as estimating the price of a product. The paper concludes that when using a two-stage estimation (*i.e.,* estimate the worker quality only using gold data), the number of control questions should be equal to the square root of the number of labels provided by the worker as a rule of thumb. Recent work has also explored more dynamic approaches that leverage gold standard questions to select tasks for workers such that overall accuracy is maximized [8, 22, 40]. However, the problem of utilizing and assigning gold standard questions has been mainly investigated in the context of multiple-choice questions, and some solutions are not generalizable across different task types. In particular, much of the previous work has relied on worker accuracy estimation models that only work with binary outcomes or multiple-choice questions.

On the one hand, using a small pool of gold standard questions can lead to problems when gold questions are repeated and flagged by workers [12, 13]. On the other hand, crowdsourcing is typically used for problems for which sourcing ground truth data is not straightforward. Thus, creating good gold data at scale and at a low cost is essential for implementing gold standards. Oleson et al. [57] propose a programmatic approach to generate gold standard data. This study indicates that programmatic gold can increase the gold per question ratio, allowing for high-quality data without increased costs.

As opposed to creating gold questions prior to the label collection, we can also iteratively validate selected answers using experts. For example, Hung et al. [35] investigate classification tasks and proposes a probabilistic model that can find the most beneficial answer to validate in terms of result correctness and detection of faulty workers. Reliable and high-quality gold data can also be generated by using domain experts [30].

2.4   Visible Gold

Typically, workers cannot distinguish between a regular question and a gold standard question. Answers received for gold questions are used to estimate the worker quality in the post-processing step or during run-time. However, gold standard questions can also be used to provide training and feedback to workers [20, 24, 48].

Research shows that providing feedback can enhance data quality in crowdsourcing. Dow et al. [21] report that both self-assessment and external expert feedback can improve crowd work quality. The study highlights that workers who receive external assessments tend to revise their work more [21]. Similarly, feedback from peers in organized worker groups can help workers achieve high output quality [79]. In a peer-review setup, the process of reviewing others' work has also been shown to help workers elevate their own data quality [86]. While peer and expert feedback can improve data quality, it is difficult to achieve the timeliness that is critical for implementing a feedback system at scale. In addition to feedback on work, workers could also benefit from learning opportunities on how to effectively use the tools and their related metrics [67].

From prior research by Dow et al. [21], we can identify three key aspects of feedback for crowd work. 'Timeliness' is how quickly the worker receives the feedback in either a synchronous or asynchronous fashion. 'Specificity' is the level of detail in the feedback, ranging from a binary decision (*e.g.,* approve, reject) to template-based structured feedback to detailed task-specific feedback. Finally, the 'Source' of the feedback could be the requester, experts, peer workers, or the worker him- or herself.

A dedicated training phase where workers complete several training tasks and receive feedback until they reach the desired quality level has also been shown to be effective in crowd tasks that involve complex tools and interfaces [61]. Prior work also shows that training or feedback can also introduce a bias due to the specific examples selected for the training/feedback step [48]. Other work uses feedback to clarify ambiguous task instructions as opposed to improving the quality of work. For

instance, Manam and Quinn [53] propose a Q&A and Edit functionality that can be used by workers to clarify and improve task instructions or questions.

Visible gold questions allow us to provide feedback while testing for work quality. Le et al. [48] show that in a relevance categorization task, a uniform distribution of labels in visible gold standard data produces optimal peaks when considering individual worker precision, as well as majority voting aggregated results. Their study includes a dedicated pre-task training phase to qualify for the task. Visible gold questions are inserted based on a simple ratio where workers encounter 1 visible gold question for every 4 questions. Workers are also blocked from continuing on a task if their accuracy is low. Before being blocked, workers receive a warning that their accuracy is too low and that they should reread the instructions to correct mistakes.

Gadiraju et al. [24] test with two training methods with visible golds. In implicit training, workers are provided training when they provide erroneous responses to gold questions, and in explicit training, workers are required to go through a training phase before they attempt to work on the task itself. The results indicate that training provides a 5% performance gain and 40% time gain across 4 task types (Information Finding, Spam detection, Sentiment Analysis, Image transcription). However, the experiment setup doesn't define a specific gold injection strategy for implicit training. Instead, it considers all questions as gold. Using complex web search challenges as the task, Doroudi et al. [20] also show that providing expert examples upfront is an effective form of training.

2.5 Bounding Box Annotation

Early improvements to object detection include improvements to the crowdsourcing task workflow. Object annotation workflow proposed by Su et al. [72] entails three steps. First, a worker draws a bounding box around a single object instance. Second, another worker verifies the drawn box. Third, a different worker determines if there are additional instances of the object class that need to be annotated. The paper reports that 97.9% of images are correctly covered with bounding boxes.

Other approaches use computer vision methods to generate bounding boxes during the annotation process [1, 2, 58, 65]. Prior work by Papadopoulos et al. [58] using an accept/reject decision could achieve high-quality results comparable to standard manual annotation. Similarly, Adhikari and Huttunen [1] propose a semi-automated batch-wise method where a subset of images are annotated and then used to train an object detection model that can generate bounding boxes for the remaining images. As the last step, generated annotations go through a manual verification where workers add/remove boxes as required. This method can reduce the manual effort by up to 75%.

Literature has also investigated how we could use different annotation strategies instead of the standard way of drawing a bounding box through click and drag interactions. For instance, bounding boxes could be auto-generated by asking workers to annotate four edge points (points belonging to the top, bottom, left-most, and right-most parts) of the object [59]. A similar approach uses a single point that corresponds to the center of the target object as opposed to four edge points [60].

A key challenge in comparing our empirical results vs. those reported in prior studies is that they tend to report on datasets having few objects per image on average: 2.5 for PASCAL VOC 2007 (used by [58, 59]), 2.4 for 2012 (used by [1, 59]) and 1.5 for ImageNet (used by [72]) datasets [51]. So while Papadopoulos et al. [59] report 88% mIoU annotation quality on PASCAL VOC 2017, this is a much easier task than ours. In contrast, Russakovsky et al. [65] report 7 objects per image on average (similar to us), but they do not report annotator mIoU.

3   THE CHALLENGE OF VARIABLE EFFORT LABELING TASKS

3.1   Defining Variable Effort Labeling Tasks

While crowdsourced annotation is well-studied, *variable effort* tasks present three key challenges vs. more typical labeling tasks: inconsistent worker experience, the potential for high cognitive load, and effective incentive design. With regard to inconsistent experience, workers may implicitly expect all task instances to require comparable effort. Highly varying effort requirements across instances would violate such an expectation and could induce surprise or frustration. Secondly, as the cognitive load becomes excessive (e.g., labeling 1000 faces in a single image of a crowd), workers may not only be frustrated but naturally struggle to complete the task accurately. As for incentive design, the typical task-based pricing model ala MTurk assumes that all task instances are compensated at the same fixed rate. Since more effortful instances require more time to complete (accurately), this equates to a lower effective earning rate for workers. These challenges, taken separately and especially together, can have various negative impacts. Workers may choose not to accept a task or quickly abandon it. They might complete easy instances but skip over more effortful ones. They may fail to deliver quality work due to demanding cognitive load or simple lack of effort.

Such tasks exemplify the applicability of rationales to a broad class of *Where's Waldo?* [81] search problems of determining whether or not a given item contains entities of interest (e.g., does Waldo appear in a given image or video clip, do we hear his voice in a given audio recording, is he discussed in a given text, etc.). The larger the item, the greater the problem searching it. For example, imagine annotating all trees in massive satellite or aerial imagery, requiring annotators to zoom and pan around images. The search problem may be explicit – e.g., *does an audio clip contain a bird call?* – or implicit – e.g., *rate a product from its description*, where the primary task is to rate the item but the annotator must search the item for evidence to support their rating decision.

Framing this search problem lets us relate variable effort annotation tasks to a large body of related work mobilizing the crowd for distributed search of large search spaces. Classic examples include the search for extraterrestrial intelligence (SETI@Home) [70], for Jim Gray's sailboat [77] or other missing people [78], for DARPA's red balloons [73], for astronomical events of interest [49], for endangered wildlife [64] or bird species [39], etc. Attenberg et al. [5] asked the crowd to find examples on which classifiers erred. Across such examples, what is being sought must be broadly recognizable so that the crowd can accomplish the search task without the need for subject matter expertise [41]. Whereas the works above involve searching for an entity across domain instances, with variable effort labeling tasks, the challenge is searching within each instance for matching entities.

There is limited prior work that examines how crowd work quality can vary when attempting tasks that involve a variable effort. In a study where workers are asked to annotate either 5 or 10 items in each HIT, Kazai [38] shows that better results can be obtained when workers are not overloaded. Similarly, crowd workers make more errors in counting tasks that include a large number of target objects [18, 66]. Our work intends to systematically evaluate the impact of variable effort on outcome quality by using a task that involves 14 discrete effort levels and requires individual actions for each work unit in the task.

Other work that focus on task complexity or difficulty has implicitly explored the relationship between task effort and the data quality [3, 10, 56, 82]. For instance, research shows how task ordering can impact the data quality when deploying tasks with varying complexity [10]. While these attributes are closely related, task complexity is a different abstraction from the task effort. A task that requires more effort (*e.g.,* annotating an image with 15 faces) is not necessarily more complex than a task that requires less effort (*e.g.,* annotating an image with 2 faces).

3.2 Face Detection Task and Dataset

For our study, we chose the variable effort task of detecting human faces within an image and drawing bounding boxes around each face. Object detection is one of the most common tasks available on crowdsourcing platforms and is substantially more difficult and time-consuming than simpler tasks like multiple choice questions [72].

We selected 140 images and ground truth data from the Open Images [47] dataset. The number of faces in the images we selected ranged from 1 to 14, with 10 images per face count. For each subset, we sorted images by ID and picked the first 10 images corresponding to a pseudo-random selection. Images with potentially ambiguous human faces, such as cartoon characters or statues, were excluded to allow for definitive quality assessments. Figure 1 shows image examples with low and high face counts respectively.

Fig. 1. Example images with four (left) and ten (right) faces with ground truth data marked using blue rectangles.

Workers completed the face detection task in a standard image annotation interface supporting basic operations such as creating, adjusting and deleting annotations, and zooming. We provided consistent base instructions for the annotation task across all experimental conditions, with some condition-specific instructions added where necessary. The instructions also included three correctly annotated example images [20] in each of the conditions.

3.3 Baseline

Labeling tasks are typically crowdsourced based on MTurk's pay-per-task pricing model assumes that all task instances are compensated at the same fixed rate. We assume this standard pricing model as our baseline, though we anticipate it may not be optimal for variable effort labeling tasks in which some tasks require much more effort than others. While a stronger baseline that adjusts payments (*e.g.,* Fair Work [80]) is desired, we select a more prevalent baseline to ensure the external validity of our results.

We grouped our image set into two distinct bins based on object count and defined a static payment amount for each bin. Workers received $0.16 for completing images with an object count between 1 and 7 and $0.44 for completing images with an object count between 8 and 14. This approach assumes an amortized pay of $0.04 per individual object label. A more basic alternative would have been to administer a constant pay of $0.30 per image without binning, but we discarded this design to compare interventions with a more competitive baseline. We used the standard object detection workflow offered on the Amazon Mechanical Turk platform for our baseline, and no visible gold was administered in this condition.

3.4 Experimental Setup

We conducted our crowdsourcing experiments on the Amazon Mechanical Turk (MTurk) platform. All experiments were deployed between 2 PM and 5 PM Pacific Time. Based on our work time estimates across all conditions, workers on average received an hourly pay of $10.44, whereas the

federal hourly minimum wage in the US is $7.25. Experiments were open to a subset of MTurk workers (a cohort of more than 8000 workers) who had previously qualified for the bounding box task based on a proprietary quality assurance mechanism. Each experimental condition was available to a unique worker pool created by segmenting the worker subset. Workers were free to complete as few or as many HITs as they liked. In each condition, images were presented in a random order. We employed a basic filtering step across all the experimental conditions where we removed obvious spammers by filtering out workers who completed more than 5 HITs and had an average mIoU below 25. We only ended up removing 3 workers across all conditions.

3.4.1 Measures. We use **mean intersection over union (mIoU)** as the primary outcome to compare work quality. mIoU is a well-established quality metric for object detection tasks [47]. It is computed as the average overall IoU values for all bounding boxes in the ground truth answer key. For each unmatched ground truth box (false-negative), that box is assigned an IoU of 0. We also report **task time** as a secondary outcome defined as the time duration from when a worker accepts a task until the task is submitted. Since we keep the average pay per bounding box constant in all conditions, task payment is not reported as an outcome measure in the paper.

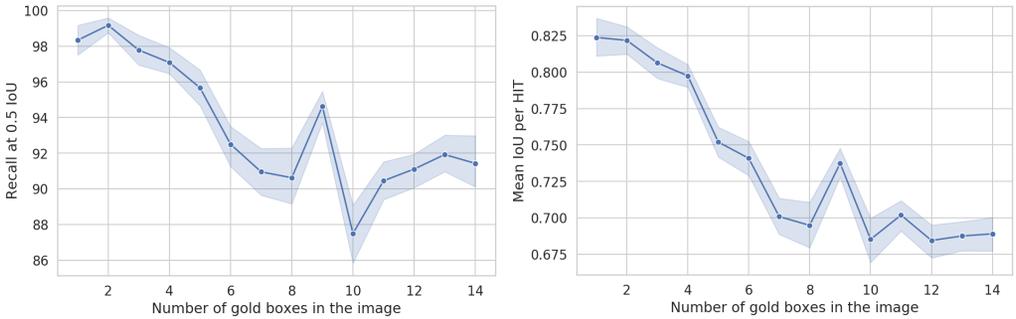## 3.5 Findings and Discussion

Fig. 2. Variation in output quality against task effort with a naïve baseline of equal pay for each task. Shaded areas correspond to standard error.

We obtained responses from 24 unique workers for the baseline. Figure 2 shows that as the number of faces per image increases, annotation quality (as measured by accuracy and recall) declines. Data quality reduction in our object detection task can attribute to either workers producing annotations of low-quality or entirely missing particular target objects when there are many target objects present in the image.

As noted above, the baseline assumes fixed-pay for all instances, despite the variable effort required. It could be argued that variation in required effort amortizes over multiple tasks to produce fair pay when workers complete enough work. However, this assumption is questionable. First, prior work shows that requesters often produce poor estimates of average effort per item and tend to underestimate the true effort [14]. This can hinder the administration of fair pay in a systematic manner. Second, work on crowdsourcing platforms typically follows a power-law distribution where only a few workers complete the majority of work, and the majority of workers abandon a task early [29]. Hence, the amortization assumption may only hold true for a small portion of the worker population while the remaining workers receive pay disproportionate to their effort. Third, even

if workers are initially motivated to complete a large number of HITs, drop-out may be encouraged if the first few tasks happen to require high effort.

While high label quality is always desired, the consistency in degradation observed in proportion to the increasing effort is noteworthy. Firstly, uniform labeling quality across all items is desirable, without any consistent biases. Secondly, if we imagine training a detection model on this data, it is particularly important to have accurately annotated images with larger object counts [71].

4 INVESTIGATING TASK DESIGNS FOR VARIABLE EFFORT LABELING TASKS

Section 3.1 suggested three key challenges with variable effort tasks: inconsistent worker experience, the potential for high cognitive load, and effective incentive design. In this section, we investigate the potential of various *best practice* task designs for crowdsourced annotation to address these challenges.

To investigate the general question of how standard quality improvement methods perform with variable effort annotation tasks. We picked several crowdsourcing data quality improvement methods that are generalizable and straightforward to implement. Since appropriate pay for effort is a key driver for quality in crowdsourcing [38], we chose to include two data collection designs—variable pay and post-task bonus—that aim to calibrate pay according to the required effort on a per-image basis. Following popular iterative and parallel design paradigms in crowdsourcing, we included two other designs—task decomposition and iterative improvement—that aim to standardize the required effort in each task unit. Finally, we add the visible gold design, which uses gold standard questions for testing and training. For each of the data collection designs outlined below, we collected three responses per image.

4.1 Variable Pay

For better per-item calibration of payment, a more sophisticated data collection design may aim to estimate the effort for each item in advance and adjust the payment on a per-item basis accordingly. These estimates can be produced either manually (e.g., via upstream crowdsourcing workflows [6]) or automatically (e.g., via pre-built object detection models [7]).

However, this data collection design introduces other challenges. First, a priori estimation of effort is a non-trivial task and can be costly when manual workflows are required. Second, since HIT payment is one of the parameters used by Amazon Mechanical Turk to group HITs in the platform, tasks with different effort levels are advertised separately, allowing workers to selectively focus on tasks with high pay and ignore low paying tasks.

We instantiated this data collection design in our study by setting the payment amount for each image proportional to the exact object count as defined in the available ground truth data. In particular, we pre-calculated pay per image by multiplying the true object count with the base pay of $0.04 per object (e.g., $0.04 for images with a single object and $0.56 for images with 14 objects).

4.2 Post-task Bonus

An alternative to a priori effort estimation is to decide an appropriate payment amount *after* a task has been completed [84]. This can be accomplished by setting up a HIT with a flat base payment and advertising a post-task bonus to compensate workers for any work completed beyond the base payment.

There are two main challenges to this data collection design pertaining to the trust relationship between workers and requesters. On the one hand, workers need to trust a requester to deliver on their promise of administering a post-hoc bonus and to choose the bonus amount fairly. On the other hand, requesters rely on good-faith execution of the task (e.g., the number of objects labeled or time spent) to produce accurate estimates of effort and fair bonus amounts. To this end, labels can be verified in a secondary process, e.g., manual verification through other workers, incurring additional cost for the requester.

We implemented this data collection design leveraging ground truth information available in our dataset. In particular, we offered workers a total payment of $0.04 for each object labeled correctly as per the ground truth data. The flat base payment was set to $0.04 for all images difference between flat base payment, and total payment was administered as the post-task bonus.

4.3 Task Decomposition

The previous two data collection designs aim at adjusting payment to variable effort on a per-item basis. An alternative approach is to decompose tasks into fixed-size units with constant effort and to administer a constant pay amount per task unit.

Prior work has shown that task decomposition not only facilitates fair payment, but also aids workers in producing high-quality answers by managing cognitive load [66]. However, the process of decomposing tasks into smaller chunks of equal size is non-trivial, and prior work suggests that some tasks depend on the context, which may not be preserved during decomposition [75]. Getting multiple individuals to attempt smaller portions of the same task can also help elevate the output quality [43, 66, 74].

We implemented this data collection design using a two-step workflow. First, we identified all target objects in a given image. Second, we created sub-tasks where workers were asked to create bounding boxes with a pre-defined set of target objects indicated through point markers. In our experiment, we created HITs with a maximum of 3 target objects and a static pay of $0.08 per HIT corresponding to 2 target objects per image on average. We implemented two variants for identifying target objects in the initial step:

- **Oracle:** In this variant, target objects were identified using the available ground truth data.
- **Manual:** Since ground truth data is generally not available in practice, we implemented a second variant in which target objects were identified manually by workers through a separate upstream point annotation task.

In addition to the above variants, for scalability, task decomposition can also be achieved through automated estimation methods. For example, as the decomposition step does not require high accuracy, we can use a generic automated object detection model [7] to generate the object markers and decompose tasks.

4.4 Iterative Improvement

We also included a task design informed by iterative crowdsourcing workflows [26] where several workers contribute to the same task while each worker can see the results from the previous worker. Prior work shows that iteration can increase the response quality in specific tasks like writing [50], brainstorming [50], and translation [4]. Quality increase typically attributes to corrective actions taken by subsequent workers. In addition, as our task involves variable effort, we also leverage an iterative workflow to regulate the amount of work that each worker needs to complete in a single iteration.

In our experimental design, multiple workers iteratively annotate the same image. In each iteration, we ask workers to either annotate a maximum of N=3 new objects, adjust existing work or mark the task as completed indicating that there is nothing left to annotate. Each iteration was deployed as a single HIT, and we set a static pay of $0.08 per HIT corresponding to an average number of two target objects per image.

4.5 Visible Gold Questions

Finally, we also explored a scenario where gold standard answers are available for a small subset of items in the dataset. These items with known answers are injected into workers' task queues to assess annotation quality on an intermittent basis. In addition to assessing quality, gold standard answers can

be used to provide near real-time feedback to workers for each object they labeled (or failed to label), e.g., for each face in an image. We refer to this feedback mechanism as "visible gold". Prior works [20, 24, 48] have investigated the use of visible gold for simple tasks with binary outcomes or multiple choices and focused on presenting visible gold questions upfront or with a static gold-to-task ratio.

Figure 3 shows our interface to provide feedback to workers using visible gold for the variable effort annotation task used in our study. Workers encounter the interface immediately after submitting their answer to a visible gold question. The feedback provided to workers included the number of objects missed (i.e., false negatives), the number of bounding boxes drawn by the worker that did not match any objects in the answer key (i.e., false positives), the accuracy for each bounding box correctly matching an object (i.e., for each true positive) and the average accuracy across all bounding boxes in the answer key (using 0% accuracy for false negatives). Accuracy for individual bounding boxes was calculated as "intersection over union" (IoU), i.e., the ratio between the area of overlap and the area of union of the worker annotation and the ground truth bounding box. IoU is a standard accuracy measure in object detection tasks. Finally, all ground truth bounding boxes were displayed to workers along with their own annotations as part of the feedback interface. While gold standard annotations were available for all images in our study, the interface was designed to dynamically decide whether a particular question should be a visible gold for a given worker based on the experimental condition.
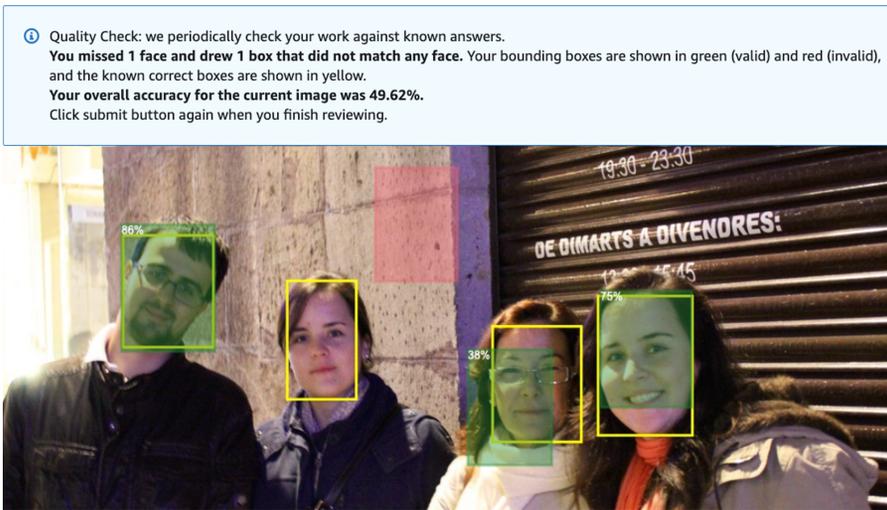

Fig. 3. Visible Gold Feedback Interface

As workers get immediate performance metrics through visible golds, they are motivated to produce high-quality work even when the task involves more effort. In addition, through the detailed feedback mechanism, visible golds can provide increased task clarity and help workers understand and improve their task performance. We expect that these characteristics will help set accurate expectations and motivate workers to carefully attempt tasks that involve an increased effort.

5 EVALUATION

We conducted our crowdsourcing experiments on the Amazon Mechanical Turk (MTurk) platform and used a consistent experimental setup as described in Section 3.4. On average, workers received a payment of $0.04 per bounding box (e.g., a worker receives a payment of $0.4 for a HIT that includes an image with 10 bounding boxes).

5.1 Results

Table 1 presents a summary of results, including the mean and standard error for mIoU values and average task time for each condition. In each condition, mIoU values follow a non-normal distribution. A Mann Whitney test with Bonferroni correction for multiple comparisons shows that mIoU values are significantly lower in iterative improvement ($M = 55.6$), $U = 499197.0$, $p < 0.001$, post-task bonus ($M = 59.5$), $U = 221215.0$, $p < 0.001$, task decomposition oracle ($M = 67.5$), $U = 286194.0$, $p < 0.001$, variable pay ($M = 69.1$), $U = 289741$, $p < 0.001$ and not significantly different to task decomposition manual ($M = 71.7$), $U = 330296.0$, $p = 0.21$ compared to the baseline ($M = 73.7$). However, mIoU values in visible gold regular ($M = 75.5$), $U = 397034.0$, $p < 0.01$ are significantly higher compared to the baseline.

Table 1. Mean IoU across conditions.

Condition	Mean (mIoU)	SE (mIoU)	Average Task Time (sec)
Baseline	73.7	0.46	182.6
Iterative Improvement	55.6	0.80	**164.4**
Post-task Bonus on work load	59.5	1.08	243.4
Task Decomposition Manual	71.7	0.74	613.8
Task Decomposition Oracle	67.5	0.90	557.7
Variable Pay on work load	69.1	0.79	227.2
Visible Gold - Regular	**75.5**	0.58	168.4

Figure 4a shows how task time per bounding box varies according to the number of ground truth target objects available in the image.

5.1.1 Impact of Task effort. Complex crowdsourcing tasks also include variable efforts within HITs. In object annotation, the number of bounding boxes that needs to be annotated in each image can range from 1 to many. In Figure 4, we examine how the output quality varies when the task effort increases. We see that our improved visible gold method consistently outperforms other methods in terms of mIoU (Figure 4c) and recall at mIoU>0.5 (Figure 4b).
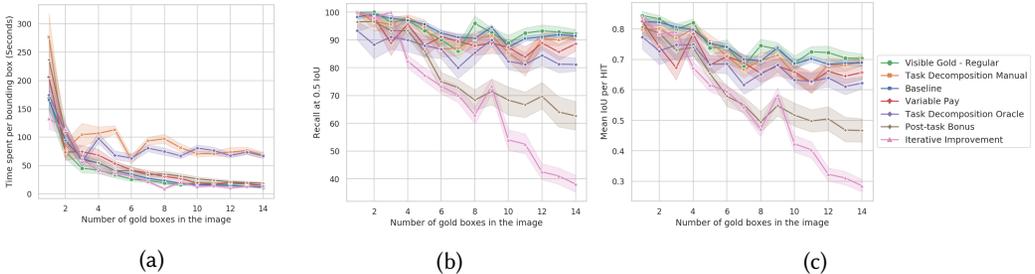

(a)　　　　　　　　　　　(b)　　　　　　　　　　　(c)

Fig. 4. Variation in task time and output quality against task effort across conditions. Shaded areas correspond to standard error.

Figure 5 shows how mIoU varies depending on the size of each ground truth bounding box. The worker output quality is relatively low for smaller objects. While this trend is visible across all the

methods, the improved visible gold method performs well above the other methods regardless of the target object size.
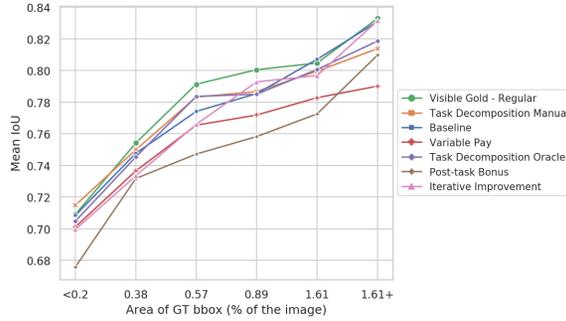

Fig. 5. Variation in mIoU against object size across conditions.

5.1.2 Task Completion. Task completion patterns based on the task submit time for variable pay and baseline conditions are given in Figure 6.
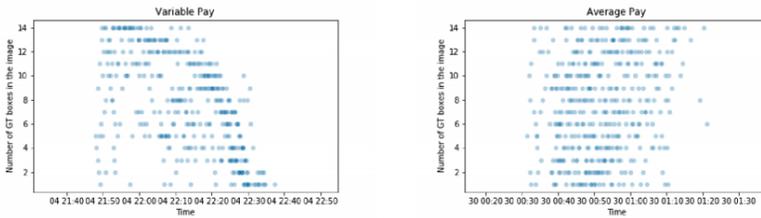

Fig. 6. Variation in task completion between the baseline (average pay) and the variable pay condition.

5.2 Analysis of Findings

Based on the results presented in Table 1, visible gold method results in the highest mIoU across all the attempted quality improvement methods. All the other quality improvement methods fail to surpass the baseline when considering the mIoU value.

To produce high-quality annotations, we also want crowd workers to carefully complete the task by spending ample time. As seen in Figure 4a, time spent on a work unit or a single object annotation diminishes when the number of objects in the image increases. On the contrary, task decomposition conditions that appear as outliers in Figure 4a can ensure consistent work time on each unit when compared to other conditions. However, we did not obtain the desired quality improvement. As given in Table 1, mIoU values from task decomposition conditions are lower than the baseline condition.

As shown in Figure 4, output quality measured in mIoU and the percentage of annotated ground truth boxes drops as the number of target objects in images increases. While this general trend is present in all experiment conditions, Visible Gold, Baseline, and Task Decomposition (Manual and Oracle) perform relatively better compared to other methods. Output quality drastically drops with effort in iterative improvement and post-task bonus conditions. For post-task bonus, this is mainly due to the low task base-price. For instance, when a worker accurately labels an image with 12 target objects, they receive a base pay of $0.08 and a bonus of $0.40. Although we use a reputed requester account

with a 99% approval rate, we can argue that workers are still unwilling to commit to a task with low specified payment and do additional work without a payment guarantee. The reason behind the observation in iterative improvement condition is not explicit. Prior work also notes that data quality can be reduced when subsequent workers are led down the wrong path in an iterative workflow of tasks with high difficulty [50]. Another possible cause can be workers failing to understand the task/instructions fully and confusing it with standard verification jobs, and prematurely marking the task as completed.

As a general trend, mIoU in our object detection task decreases when the target object is smaller. There are two possible causes for this observation. First, workers could completely miss smaller target objects during the annotation process when there are many target objects present in the image (as seen in Figure 4). Second, when the object is smaller, the impact on error is also higher (e.g., the error caused by missing the margin by a single pixel is higher for relatively smaller target objects).

In contrast to other conditions, workers could pick tasks with specific object counts under variable pay condition. In Figure 6, we observe that workers prioritized tasks with higher effort and higher pay.

6 IMPROVING VISIBLE GOLD

Our initial evaluation shows how visible gold method can result in annotations with higher quality in object detection task when compared to other quality improvement methods. Also, visible gold method that relies on gold standard questions is potentially more generalizable across many task types as opposed to other methods like task decomposition and iterative improvement. The applicability of visible gold is also not limited to tasks that involve a variable effort. These factors led us to further investigate visible gold as a promising generic quality improvement method for crowdsourcing.

In this section, we detail how we refined our visible gold method. First, we explore different visible gold issuing patterns. Second, we evaluate how bonuses and warnings work as consequences when using visible gold. Finally, we present an improved visible gold method that incorporates tier-based consequences and dynamic visible gold issuing.

6.1 Visible Gold Issuing Pattern

We evaluate three ways of issuing visible gold questions and obtained five responses per image in each condition.

- *Upfront*: Workers complete a fixed number of visible gold questions at the beginning. This condition is comparable to the explicit training in previous work by Gadiraju et al. [24]. Upfront condition is straightforward to implement and can be considered as a variant of a qualification test.
- *Regular*: Workers encounter a visible gold regularly with a visible gold to task ratio of 1:4 similar to the prior work by Le et al. [48]. We also included an offset such that workers are not able to anticipate a visible gold by following the issuing pattern.
- *Fib+Regular*: We propose a new strategy that combines the characteristics of Upfront and Regular conditions. We seek to present more visible gold questions at the beginning such that we can test the workers reliably while providing ample training examples. But as workers continue, we want to test less frequently. We achieve this by following the Fibonacci sequence. Under this condition, up to 50 questions, we follow the Fibonacci sequence (*i.e.,* 1,2,3,5,8,13..) to issue visible golds and then falls to a more infrequent regular visible gold to task ratio of 1:19.

6.2 Bonus vs. Warning as a consequence

For Upfront, Regular and Fib+Regular conditions, we used warning as the consequence where we warned workers that they would not be able to attempt future tasks if their outcome measured via quality checks (i.e., visible gold questions) does not meet the expected quality standard. However, we did not block any workers during the task or remove any contributions from our analysis. To

examine if incentivizing workers to produce high-quality annotations works better than the warning, we deployed the following additional condition.

- *Regular Bonus*: Similar to the 'Regular' condition, workers encounter a visible gold regularly with a visible gold to task ratio of 1:4. For the consequence, instead of the warning, workers receive a bonus payment of $0.08 (face count less than 8) or $0.22 (face count greater than 7) per image if they maintain an accuracy above a pre-specified threshold. Bonus amounts and thresholds were specified in the task description.

6.3 Dynamic visible gold and tier-based consequences

Based on the literature and results obtained in our first round of experiments detailed above, we further improved our visible gold mechanism by dynamically adjusting the visible gold issuing pattern and by adding a performance metric display element.

Fig. 7. Improved bounding box task interface

During our initial experiments, a handful of workers ignored the training and guidance provided by visible gold questions and continued to produce low-quality work. To counter this, we altered the Fib+Regular visible gold issuing pattern by adding bonusing and blocking conditions with predefined quality thresholds. Bonus thresholds $T_{BONUS-A}$ and $T_{BONUS-B}$ determined whether a worker qualified for a bonus payment. Blocking threshold T_{MIN} was the minimum mIoU that a worker needed to achieve to pass a given visible gold. When a worker completed a visible gold with their mIoU for the current image falling below T_{MIN}, we overrode the standard visible gold pattern and issued another visible gold as the next HIT. We continued to issue visible golds until either the worker passed a visible gold or they met the blocking condition by failing three consecutive visible golds with an overall average accuracy below the blocking threshold T_{MIN}.

To increase transparency, we added a dedicated performance metric banner at the top of the task interface as seen in Figure 7. A worker could see their current average accuracy and the relevant quality tier in a simplified manner. The content of the banner was updated as workers completed visible golds.

We ran additional experiments with the improved visible gold interface. We collected five responses per image and kept all other parameters regarding the experiment consistent with previous deployments detailed in Crowdsourcing Experiments (Section 3.4).

7 EVALUATION II

7.1 Results

A summary of results for visible gold conditions, including the mean and standard error for mIoU values and average task time is given in Table 2.

Table 2. Mean IoU across conditions.

Condition	Mean (mIoU)	SE (mIoU)	Average Task Time (sec)
Baseline	73.7	0.46	182.6
Visible Gold - Regular	75.5	0.58	168.4
Visible Gold - Upfront	75.5	0.53	**165.6**
Visible Gold - Fib+Regular	75.7	0.58	177.0
Visible Gold - Regular Bonus	74.7	0.59	193.9
Visible Gold - Improved	**79.3**	0.41	168.0
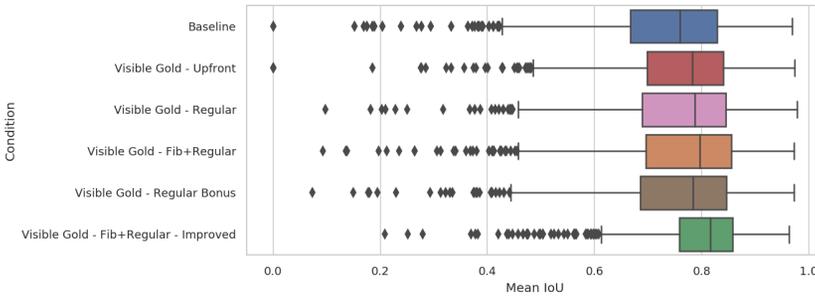
Fig. 8. Distribution of mIoU for each submission across conditions

7.1.1 Visible Gold Execution. From our initial experiments to identify the suitable visible gold execution strategy, all strategies produced better outcomes when compared to the baseline. Obtained mIoU values in each condition follow a non-normal distribution. A Mann Whitney test with Bonferroni correction for multiple comparisons shows that mIoU visible-gold-upfront ($M = 75.447$), $U = 483319$, $p < 0.01$, visible-gold-regular ($M = 75.48$), $U = 397034$, $p < 0.01$, and visible-gold-fib+regular ($M = 75.650$), $U = 451168$, $p < 0.001$ have significantly higher mIoU values compared to the baseline ($M = 73.739$).

In order to identify the most suitable visible gold execution strategy, we separated answers into three bins based on the hit completion order and plotted the mIoU metric in Figure 9a. In Figure 9b, we also show the variation in the distribution of the total number of tasks completed by each worker.

7.1.2 Visible Gold Improved. A Mann Whitney test indicated that mean mIoU in improved visible gold condition ($M = 79.301$) is significantly higher than mean mIoU in baseline ($M = 73.739$),
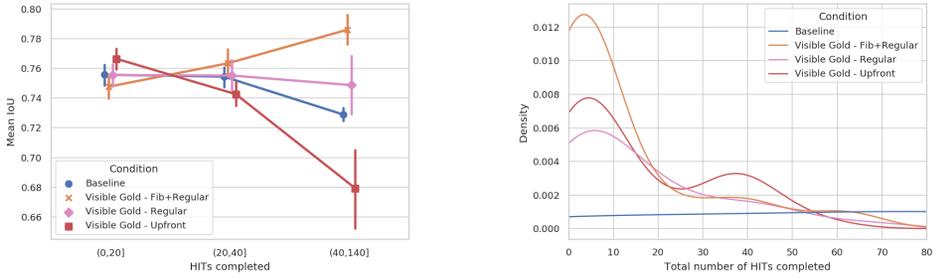
(a) Variation in mIOU against HIT completion. Error bars show standard error.

(b) Distribution of the total number of HITs completed by each worker

Fig. 9. HIT completion across conditions

$U = 393812$, $p < 0.001$. Results in improved condition were also significantly higher than visible-gold-fib+regular condition ($M = 75.650$), $U = 172731$, $p < 0.001$, which provided the best outcome in the first round of experiments.

7.1.3 Bonus vs. Warning. We compare between using warnings and bonuses as a consequence for failing visible gold questions. Our results show that there is no significant difference between bonus ($M = 74.670$) and warning ($M = 75.483$) conditions, $U = 145062$, $p = 0.26$.

7.1.4 Impact of Task effort. In Figure 10, we revisit how the output quality varies when the task effort increases. We see that our improved visible gold method consistently outperforms other visible gold methods and the baseline in terms of mIoU (Figure 10b) and recall at mIoU>0.5 (Figure 10a).

(a)

(b)
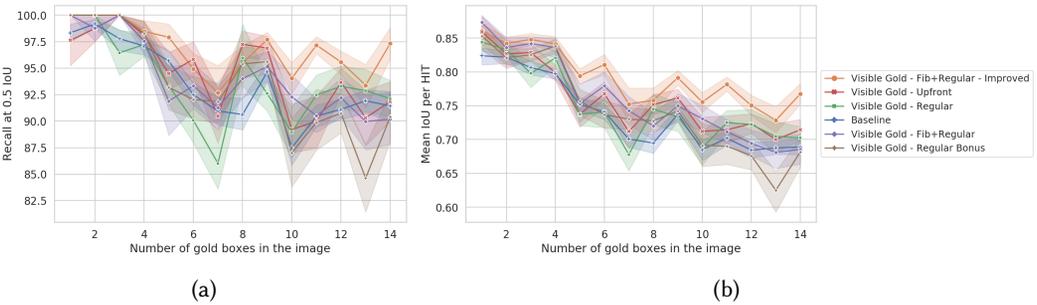
Fig. 10. Variation in output quality against task effort for different visible gold conditions. Shaded areas correspond to standard error.

Figure 11 shows how mIoU varies depending on the size of each ground truth bounding box. The worker output quality is relatively low for smaller objects. While this trend is visible across all the methods, the improved visible gold interface outperformed other methods regardless of the target object size.

7.2 Analysis of Findings

We first examined the optimum visible gold execution pattern. Our analysis revealed that visible-gold-fib+regular is superior to standard upfront [24] or regular [48] methods. In addition to producing
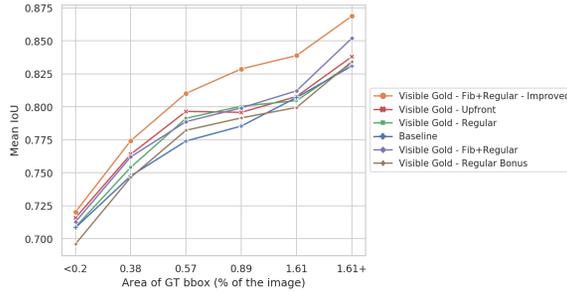
Fig. 11. Variation in mIoU against object size for different visible gold conditions.

a marginally better overall mIoU score, the fib+regular method helps us preserve the data quality as workers continue to complete tasks. As seen in Figure 9a, work quality declines as workers complete more tasks in both baseline and visible-gold-upfront conditions. However, in visible-gold-regular and visible-gold-fib+regular work quality remains steady as work continues. Fib+regular method is also more advantageous for jobs with a large number of HITs. For example, when a worker completes 100 HITs, regular pattern issues 20% visible golds, whereas fib+regular issues only 11%. In Figure 9b, we also observe that a large portion of workers tend to leave the task after completing several HITs in visible-gold conditions. This positive observation confirms that certain workers left the task as they were confronted with quality checks.

Concerning bonus and warning, our results indicate no significant difference. We used these findings to inform the design of improved method where we used the fib+regular as the base visible gold issuing pattern and incorporated both bonus and warning with a tier-based design.

Table 2 presents the impact of different visible gold presentation strategies on the quality of responses for our object detection task. Note that there is little variation in average time per task among the different visible gold presentation strategies when compared with the other quality-improvement strategies (Table 1). On the other hand, the overall quality of annotations increased markedly once we implemented the improved task interface with different "tiers" of performance, suggesting that dynamic feedback with clear and transparent communication about penalties and rewards incentivizes higher quality. Variation in the cadence of visible gold presentation appears to have had less impact than the improvement in the task interface.

Figure 10 demonstrates that the improvement from the improved interface can be primarily attributed to better performance on the tasks requiring the most effort as measured by the number of ground truth boxes in the image. The dynamic tracking and reporting of the worker's running accuracy score on gold data may have made the impact of a small number of poor annotations on a worker's overall performance more clear, incentivizing increased attention to detail on the more difficult tasks.

8 DISCUSSION

In this paper, we investigated how data quality improvement mechanisms perform for tasks that involve variable effort. We evaluated common quality enhancement methods and showed that the visible gold method produced annotations of significantly higher quality. We further refined and evaluated the visible gold method, demonstrating the effectiveness of combining upfront and regular testing patterns. Our results also suggest that workers produce better annotations when confronted with consequences via the visible gold feedback interface. However, there was no detectable difference in how bonuses (for high-quality work) and warnings (for low-quality work) affected output quality in the context of visible gold.

8.1 Data Quality and Variable Effort Tasks

Our systematic evaluation shows that both object count and object size can impact the annotation quality in object detection task. Our results are consistent with prior work that shows data quality suffers when tasks involve increased effort [18, 38]. While there are numerous other crowdsourcing data quality improvement methods (*e.g.,* work strategies [28], task assignment [31]), in this work, we are primarily interested in methods that can potentially support variable effort crowdsourcing.

Initially, we hypothesized that data quality improvement methods that aim to either standardize the effort (e.g., task decomposition and iterative improvement) or match the pay according to the effort (e.g., variable pay and post-task bonus) should work better when compared to a baseline average pay scheme. However, as detailed in Table 1 and Figure 4, none of these methods were successful in surpassing the baseline. While literature [6, 50] highlights that these methods can improve data quality in specific tasks and scenarios, our work shows that such improvements may not hold when the task effort varies. Prior work also highlights that static workflows perform poorly for complex crowd tasks [62]. Therefore, it is important to consider the task effort when using crowdsourcing to obtain annotations as well as when evaluating methods related to crowd work on tasks like bounding box annotation.

However, we show that out of the detailed methods, visible gold is the most useful method in preserving data quality in variable effort tasks.

8.2 Visible Gold for Training

When creating the visible gold mechanism for object detection, we considered three dimensions highlighted in the crowdsourcing literature on worker feedback [21]. In terms of timeliness, we designed our feedback to be synchronous. However, to prevent workers from guessing which tasks were visible gold questions, they were shown feedback only *after* completing the entire task, i.e., drawing all the bounding boxes for a particular image. Regarding specificity, we provided automated yet detailed feedback (Figure 3), including aggregate metrics on the image level and fine-grained metrics for each bounding box.

Our study extends prior work on visible gold [24, 48] by integrating existing testing patterns into a hybrid pattern (Fib+Regular) with both upfront and regular testing. Our results demonstrate that this hybrid pattern is more effective at maintaining annotation quality over the course of large amounts of tasks. The capacity to sustain data quality is particularly important as crowd work typically follows a power-law distribution where only a few workers self-select to complete the majority of work, whereas the remainder of the workforce abandons a task early on [63]. Further, our results in Figures 11 and 10 show that the improved visible gold mechanism is robust when the task involves variable effort with respect to object count and object size. Our findings are in line with prior work that uses periodic bonus payments [19] and achievement priming [23] to motivate high quality crowd work.

In addition to improving annotation quality, we argue that our visible gold method possesses a variety of positive attributes from a worker's perspective. First, our method provides transparency around the expected annotation *accuracy*. Second, it provides task *clarity* [25] to workers by demonstrating how a task should be done by means of concrete examples. Third, visible gold provides *feedback* to workers, helping them understand their individual task performance, correct specific mistakes, and improve their subsequent work. These three factors contribute to a better understanding of expected task outcomes and reduce the possibility of workers abandoning tasks [29], leading to unpaid work. Finally, visible gold provides the opportunity for workers to give feedback to the requester if gold standard annotations are faulty or if evaluation results seem to be incorrect (e.g., if there is a bug in the evaluation metric code), which is not possible with hidden gold.

8.3   Implementing Visible Gold

Our visible gold method can be easily implemented by task requesters or crowdsourcing platform itself. We discuss several factors that should be carefully considered in a practical implementation.

Generalizability: We anticipate that it is possible to develop visible gold templates for many other tasks. Prior work demonstrates how this can be achieved for text-based classification tasks [24, 48] and content creation tasks [24]. It would be straightforward to extend the current work for certain tasks like semantic segmentation and keypoint annotation [44] but would require additional effort if the task has no objective evaluation metric like accuracy or mIoU. While it is trivial to present feedback based on visible golds for object detection, additional explanations would be desired for certain complex tasks. Future work can explore how to augment visible golds with explanations or rationales.

In the experiment, we picked threshold values for bonuses and blocks based on percentiles in the baseline data. When implementing visible gold in a crowdsourcing platform, requesters can either set these values based on absolute quality expectations or calibrate thresholds based on initial insights from pilot jobs.

Generating Gold: As visible golds serve as training examples, it is also important to create a reliable set of visible gold standard questions. Like hidden gold questions, visible gold should be representative of the dataset and should sufficiently cover edge cases and ambiguous tasks. While our study assumes that there exists a single objective ground truth answer for any given task, in practice, many tasks are ambiguous [69]. Future work may test evaluation strategies that accept multiple valid ground truth answers [11, 68]. The problem of generating high-quality gold standards is a non-trivial process and remains an open research challenge. An important challenge for future work is how visible gold mechanisms can work when available ground truth data are imperfect or noisy.

Presenting Feedback: The visible gold presentation in the current work includes detailed feedback for each visible gold answer, providing an overall accuracy score along with metrics and visual feedback for each work unit. We also added continuous feedback on quality checks during further refinements. However, we identify several future improvements for visible gold interfaces. First, future designs may introduce a 'revise and resubmit' mechanism. The design implemented in this study did not let workers adjust their original annotation after being presented with feedback on a visible gold question. Correcting their original annotation according to the gold answer could help workers understand how to achieve higher task accuracy through experiential learning. Second, the interface could be enhanced by adding interactive feedback. In the current implementation, workers only receive feedback on their work once they submit the answer. It is also possible to explore whether interactive feedback for partially completed tasks is helpful for the workers. This can be more meaningful for highly complex tasks such as bounding box or 3D point annotation tasks with a large number of target objects in the same image/task. Third, similar to worker-led instruction refinement [37] or workflow creation [45], we could encourage workers to provide feedback on visible gold. If a worker encounters a flawed visible gold question, there should be a way to flag it or provide detailed feedback. This way, requesters can remove the reported visible gold questions from the task. As worker quality is measured through visible golds, this is an important enhancement when implementing visible gold at scale. In addition to the information provided regarding work quality, we could explore interventions such as micro-diversions [16], dedicated training sub-tasks [20], mandatory instruction documents, etc.

Testing Patterns: The adaptive gold execution strategy introduced in this work can be expanded by considering worker quality metrics outside the current job. For instance, if we already know a particular worker is doing well in object detection based on previous jobs, we can reduce the visible gold frequency. It is also possible to draw from prior work that investigates how to issue hidden gold questions optimally [8]. Finally, the positive impact of the tiering system in the improved task interface suggests two interesting directions for future study: (1) whether we can expand it to a

platform-level system and (2) how we can estimate/dynamically vary threshold quality values and rewards to improve performance.

The use of visible golds also has several inherent limitations that practitioners should be aware of. With visible golds, workers can easily flag the HITs as Gold Questions, and with the help of third-party plugins, other workers may be able to detect in advance whether a particular HIT is gold or not [12, 13].

8.4 Limitations

We acknowledge several limitations of our study. First, in our crowdsourcing experiments, workers were allowed to attempt an arbitrary number of questions instead of being assigned a fixed quota. We made this design decision to match the typical workflow in crowdsourcing platforms and therefore ensure the ecological validity of our work. As a result, the distribution of work between workers was uneven. Second, we did not specifically collect worker demographic information that may impact the output quality. However, we utilize a curated worker pool that excludes workers who would intentionally spam the task or produce low-quality data.

9 CONCLUSION

In this paper, we systematically evaluated the impact of existing quality improvement methods for tasks involving variable effort. Our results from a series of crowdsourced experiments in the context of object detection show that providing feedback to human annotators via visible gold can produce better quality outcomes than methods aiming to balance effort and pay at the item level through adjusting pay per item or decomposing the task into chunks of similar effort. We further designed and empirically evaluated variants of the visible gold method testing different issuance patterns and quality-related consequences. Our final design iteration of visible gold combined dynamic testing patterns with tier-based consequences and significantly improved bounding box accuracy by 5.7% compared to a basic visible gold variant and by 7.5% compared to a baseline without visible gold. Our work broadens the understanding of quality assurance processes for variable effort annotation tasks and emphasizes the value of visible gold-based feedback mechanisms in this process.

ACKNOWLEDGMENTS

REFERENCES

[1] Bishwo Adhikari and Heikki Huttunen. 2020. Iterative Bounding Box Annotation for Object Detection. In *International Conference on Pattern Recognition (ICPR)*.

[2] Bishwo Adhikari, Jukka Peltomaki, Jussi Puura, and Heikki Huttunen. 2018. Faster Bounding Box Annotation for Object Detection in Indoor Scenes. In *2018 7th European Workshop on Visual Information Processing (EUVIP)*. 1–6. https://doi.org/10.1109/EUVIP.2018.8611732

[3] Alan Aipe and Ujwal Gadiraju. 2018. SimilarHITs: Revealing the Role of Task Similarity in Microtask Crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media (HT '18)*. ACM, New York, NY, USA, 115–122. https://doi.org/10.1145/3209542.3209558

[4] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2012. Collaborative Workflow for Crowdsourcing Translation. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. ACM, New York, NY, USA, 1191–1194. https://doi.org/10.1145/2145204.2145382

[5] Josh Attenberg, Panagiotis G. Ipeirotis, and Foster Provost. 2011. Beat the Machine: Challenging Workers to Find the Unknown Unknowns. In *Proceedings of the AAAI Conference on Human Computation (HCOMP)*. AAAI Press, 2–7.

[6]   Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 313–322. https://doi.org/10.1145/1866029.1866078

[7]   Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. 2019. Salient object detection: A survey. *Computational Visual Media* 5, 2 (2019), 117–150. https://doi.org/10.1007/s41095-019-0149-9

[8]   Jonathan Bragg, Mausam, and Daniel S. Weld. 2016. Optimal testing for crowd workers. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* (2016), 966–974.

[9]   Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5.

[10] Carrie J Cai, Shamsi T Iqbal, and Jaime Teevan. 2016. Chain Reactions: The Impact of Order on Microtask Chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3143–3154. https://doi.org/10.1145/2858036.2858237

[11] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[12] Alessandro Checco, Jo Bates, and Gianluca Demartini. 2018. All That Glitters is Gold – An Attack Scheme on Gold Questions in Crowdsourcing. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. AAAI Press.

[13] Alessandro Checco, Jo Bates, and Gianluca Demartini. 2020. Adversarial Attacks on Crowdsourcing Quality Control. *Journal of Artificial Intelligence Research* 67 (2020), 375–408. https://doi.org/10.1613/jair.1.11332

[14] Justin Cheng, Jaime Teevan, and Michael S. Bernstein. 2015. Measuring Crowdsourcing Effort with Error-Time Curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 1365–1374. https://doi.org/10.1145/2702123.2702145

[15] Peng Dai, Christopher H. Lin, Mausam, and Daniel S. Weld. 2013. POMDP-based control of workflows for crowdsourcing. *Artificial Intelligence* 202 (2013), 52 – 85. https://doi.org/10.1016/j.artint.2013.06.002

[16] Peng Dai, Jeffrey M. Rzeszotarski, Praveen Paritosh, and Ed H. Chi. 2015. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. ACM, New York, NY, USA, 628–638. https://doi.org/10.1145/2675133.2675260

[17] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *Comput. Surveys* 51, 1 (4 2018), 1–40. https://doi.org/10.1145/3148148

[18] Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2016. Towards Globally Optimal Crowdsourcing Quality Management. In *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16 (SIGMOD '16, Vol. 26-June-20)*. ACM Press, New York, New York, USA, 47–62. https://doi.org/10.1145/2882903.2882953

[19] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Philippe Cudr, and Philippe Cudré-Mauroux. 2014. Scaling-up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement. *Second AAAI Conference on Human Computation and Crowdsourcing* Hcomp (2014), 50–58.

[20] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 2623–2634. https://doi.org/10.1145/2858036.2858268

[21] Steven P. Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1013–1022. https://doi.org/10.1145/2145204.2145355

[22] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. 2015. iCrowd: An Adaptive Crowdsourcing Framework. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15)*. ACM, New York, NY, USA, 1015–1030. https://doi.org/10.1145/2723372.2750550

[23] Ujwal Gadiraju and Stefan Dietze. 2017. Improving Learning through Achievement Priming in Crowdsourced Information Finding Microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*. ACM, New York, NY, USA, 105–114. https://doi.org/10.1145/3027385.3027402

[24] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. 2015. Training Workers for Improving Performance in Crowdsourcing Microtasks. In *Design for Teaching and Learning in a Networked World*. Springer, Cham, 100–114. https://doi.org/10.1007/978-3-319-24258-3_8

[25] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media - HT '17 (HT '17)*. ACM Press, New York, New York, USA, 5–14. https://doi.org/10.1145/3078714.3078715

[26] Shinsuke Goto, Toru Ishida, and Donghui Lin. 2016. Understanding crowdsourcing workflow: modeling and optimizing iterative and parallel processes. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP, Vol. 4)*. AAAI Press.

[27] Catherine Grady and Matthew Lease. 2010. Crowdsourcing Document Relevance Assessment with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, Los Angeles, 172–179.

[28] Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. 2020. Crowd worker strategies in relevance judgment tasks. In *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3336191.3371857

[29] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. ACM, New York, NY, USA, 321–329. https://doi.org/10.1145/3289600.3291035

[30] Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining Crowdsourcing and Google Street View to Identify Street-Level Accessibility Problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 631–640. https://doi.org/10.1145/2470654.2470744

[31] Danula Hettiachchi, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (10 2020), 1–22. https://doi.org/10.1145/3415181

[32] Chien Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 419–429. https://doi.org/10.1145/2736277.2741102

[33] John Joseph Horton and Lydia B. Chilton. 2010. The Labor Economics of Paid Crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce* (Cambridge, Massachusetts, USA) *(EC '10)*. ACM, New York, NY, USA, 209–218. https://doi.org/10.1145/1807342.1807376

[34] Shih-Wen Huang and Wai-Tat Fu. 2013. Enhancing Reliability Using Peer Consistency Evaluation in Human Computation. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. ACM, New York, NY, USA, 639–648. https://doi.org/10.1145/2441776.2441847

[35] Nguyen Quoc Viet Hung, Duong Chi Thang, Matthias Weidlich, and Karl Aberer. 2015. Minimizing Efforts in Validating Crowd Answers. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (Melbourne, Victoria, Australia) *(SIGMOD '15)*. ACM, New York, NY, USA, 999–1014. https://doi.org/10.1145/2723372.2723731

[36] Panos Ipeirotis. 2011. Pay Enough or Don't Pay at All. May 13. https://www.behind-the-enemy-lines.com/2011/05/pay-enough-or-dont-pay-at-all.html.

[37] V. K. Chaithanya Manam, Dwarakanath Jampani, Mariam Zaim, Meng-Han Wu, and Alexander J. Quinn. 2019. TaskMate: A Mechanism to Improve the Quality of Instructions in Crowdsourcing. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. ACM, New York, NY, USA, 1121–1130. https://doi.org/10.1145/3308560.3317081

[38] Gabriella Kazai. 2011. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *Advances in Information Retrieval*. Springer, Berlin, Heidelberg, 165–176. https://doi.org/10.1007/978-3-642-20161-5_17

[39] Steve Kelling, Jeff Gerbracht, Daniel Fink, Carl Lagoze, Weng-Keen Wong, Jun Yu, Theodoros Damoulas, and Carla Gomes. 2012. A Human/Computer Learning Network to Improve Biodiversity Conservation and Research. *AI Magazine* 34, 1 (Dec. 2012), 10. https://doi.org/10.1609/aimag.v34i1.2431

[40] Asif R. Khan and Hector Garcia-Molina. 2017. CrowdDQS: Dynamic Question Selection in Crowdsourcing Systems. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17)*. ACM, New York, NY, USA, 1447–1462. https://doi.org/10.1145/3035918.3064055

[41] Kenneth A. Kinney, Scott B. Huffman, and Juting Zhai. 2008. How Evaluator Domain Expertise Affects Search Result Relevance Judgments. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA) *(CIKM '08)*. ACM, New York, NY, USA, 591–598. https://doi.org/10.1145/1458082.1458160

[42] Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. 2012. CrowdWeaver: Visually Managing Complex Crowd Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. ACM, New York, NY, USA, 1033–1036. https://doi.org/10.1145/2145204.2145357

[43] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: Crowdsourcing Complex Work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. ACM, New York, NY, USA, 43–52. https://doi.org/10.1145/2047196.2047202

[44] Adriana Kovashka, Olga Russakovsky, and Li Fei-Fei. 2016. *Crowdsourcing in Computer Vision*. Now Publishers Inc., Hanover, MA, USA.

[45] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively Crowdsourcing Workflows with Turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. ACM, New York, NY, USA, 1003–1012. https://doi.org/10.1145/2145204.2145354

[46] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. 2020. Annotator Rationales for Labeling Tasks in Crowdsourcing. *Journal of Artificial Intelligence Research (JAIR)* 69 (2020), 143–189. https://doi.org/10.1613/jair.1.12012

[47] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* 128, 7 (2020), 1956–1981. https://doi.org/10.1007/s11263-020-01316-z

[48] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*. 21–26.

[49] Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, et al. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (2008), 1179–1189.

[50] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*. 68–76.

[51] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2020. Deep learning for generic object detection: A survey. *International journal of computer vision* 128, 2 (2020), 261–318. https://doi.org/10.1007/s11263-019-01247-4

[52] Qiang Liu, Alexander T Ihler, and Mark Steyvers. 2013. Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?. In *Advances in Neural Information Processing Systems*, Vol. 26. Curran Associates, Inc., 1914–1922.

[53] V.K. Chaithanya Manam and Alexander J. Quinn. 2018. WingIt: Efficient refinement of unclear task instructions. In *Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP, Vol. 6)*. AAAI Press.

[54] Akash Mankar, Riddhi J. Shah, and Matthew Lease. 2017. Design Activism for Minimum Wage Crowd Work. In *Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP): Works-in-Progress Track*.

[55] Winter Mason and Duncan J. Watts. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, 77–85.

[56] Edward Newell and Derek Ruths. 2016. How One Microtask Affects Another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3155–3166. https://doi.org/10.1145/2858036.2858490

[57] David Oleson, Alexander Sorokin, Greg Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation (HCOMP)*. AAAI Press.

[58] Dim P. Papadopoulos, Jasper R.R. Uijlings, Frank Keller, and Vittorio Ferrari. 2016. We Don't Need No Bounding-Boxes: Training Object Class Detectors Using Only Human Verification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 854–863. https://doi.org/10.1109/CVPR.2016.99

[59] Dim P. Papadopoulos, Jasper R.R. Uijlings, Frank Keller, and Vittorio Ferrari. 2017. Extreme Clicking for Efficient Object Annotation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 4940–4949. https://doi.org/10.1109/ICCV.2017.528

[60] Dim P. Papadopoulos, Jasper R.R. Uijlings, Frank Keller, and Vittorio Ferrari. 2017. Training Object Class Detectors with Click Supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 180–189. https://doi.org/10.1109/CVPR.2017.27

[61] Sunghyun Park, Philippa Shoemark, and Louis-Philippe Morency. 2014. Toward Crowdsourcing Micro-Level Behavior Annotations: The Challenges of Interface, Training, and Generalization. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (Haifa, Israel) *(IUI '14)*. ACM, New York, NY, USA, 37–46. https://doi.org/10.1145/2557500.2557512

[62] Daniela Retelny, Michael S. Bernstein, and Melissa A. Valentine. 2017. No Workflow Can Ever Be Enough: How Crowdsourcing Workflows Constrain Complex Work. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 89 (Dec. 2017), 23 pages. https://doi.org/10.1145/3134724

[63] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. AAAI Press.

[64] Holly Rosser and Andrea Wiggins. 2019. Crowds and Camera Traps: Genres in Online Citizen Science Projects. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[65] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. 2015. Best of both worlds: human-machine collaboration for object annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2121–2131. https://doi.org/10.1109/CVPR.2015.7298824

[66] Akash Sarma, Ayush Jain, Arnab Nandi, Aditya Parameswaran, and Jennifer Widom. 2015. Surpassing humans and computers with jellybean: Crowd-vision-hybrid counting algorithms. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP, Vol. 3)*. AAAI Press.

[67] Saiph Savage, Chun Wei Chiang, Susumu Saito, Carlos Toxtli, and Jeffrey Bigham. 2020. Becoming the Super Turker:Increasing Wages via a Strategy from High Earning Workers. In *Proceedings of The Web Conference 2020 (WWW '20)*. ACM, New York, NY, USA, 1241–1252. https://doi.org/10.1145/3366423.3380200

[68] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-Aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376506

[69] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 154 (Nov. 2018), 19 pages. https://doi.org/10.1145/3274423

[70] SetiHome. 2021. SetiHome. https://setiathome.berkeley.edu

[71] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE international conference on computer vision*. 8430–8439.

[72] Hao Su, Jia Deng, and Li Fei-Fei. 2012. Crowdsourcing Annotations for Visual Object Detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press.

[73] John C Tang, Manuel Cebrian, Nicklaus A Giacobe, Hyun-Woo Kim, Taemie Kim, and Douglas Beaker Wickert. 2011. Reflecting on the DARPA red balloon challenge. *Commun. ACM* 54, 4 (2011), 78–85.

[74] Jaime Teevan, Shamsi T. Iqbal, and Curtis von Veh. 2016. Supporting Collaborative Writing with Microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 2657–2668. https://doi.org/10.1145/2858036.2858108

[75] Yongxin Tong, Lei Chen, Zimu Zhou, H. V. Jagadish, Lidan Shou, and Weifeng Lv. 2018. SLADE: A Smart Large-Scale Task Decomposer in Crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* 30, 8 (8 2018), 1588–1601. https://doi.org/10.1109/TKDE.2018.2797962

[76] Jennifer Wortman Vaughan. 2017. Making better use of the crowd: How crowdsourcing can advance machine learning research. *The Journal of Machine Learning Research* 18, 1 (2017), 7026–7071.

[77] Werner Vogels. 2007. *Help Find Jim Gray*. https://www.allthingsdistributed.com/2007/02/help_find_jim_gray.html.

[78] Bing Wang, Bonan Hou, Yiping Yao, and Laibin Yan. 2009. Human flesh search model incorporating network expansion and gossip with feedback. In *2009 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*. IEEE, 82–88. https://doi.org/10.1109/DS-RT.2009.36

[79] Mark E. Whiting, Dilrukshi Gamage, Snehalkumar (Neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan, Teogenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A. Mullings, Yoni Dayan, Kristy Milland, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, and Michael S. Bernstein. 2017. Crowd Guilds: Worker-Led Reputation and Feedback on Crowdsourcing Platforms. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 1902–1913. https://doi.org/10.1145/2998181.2998234

[80] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP, Vol. 7)*. 197–206.

[81] Wikipedia. 2020. Where's Wally? https://en.wikipedia.org/wiki/Where%27s_Wally%3F.

[82] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling Task Complexity in Crowdsourcing. *The Fourth AAAI Conference on Human Computation and Crowdsourcing* October, 249–258. https://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/viewFile/14039/13653

[83] Teng Ye, Sangseok You, and Lionel Robert Jr. 2017. When Does More Money Work? Examining the Role of Perceived Fairness in Pay on the Performance Quality of Crowdworkers. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1.

[84] Ming Yin and Yiling Chen. 2015. Bonus or Not? Learn to Reward in Crowdsourcing. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 201–207.

[85] Ming Yin and Yiling Chen. 2016. Predicting crowd work quality under monetary interventions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP, Vol. 4)*. AAAI Press.

[86] Haiyi Zhu, Steven P. Dow, Robert E. Kraut, and Aniket Kittur. 2014. Reviewing Versus Doing: Learning and Performance in Crowd Assessment. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1445–1455. https://doi.org/10.1145/2531602.2531718