

Crowdsourcing More Effective Initializations for Single-Target Trackers Through Automatic Re-querying

Stephan J. Lemmer
lemmersj@umich.edu
University of Michigan
Ann Arbor, MI, USA

Jean Y. Song
jyskwon@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Jason J. Corso
jcorso@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

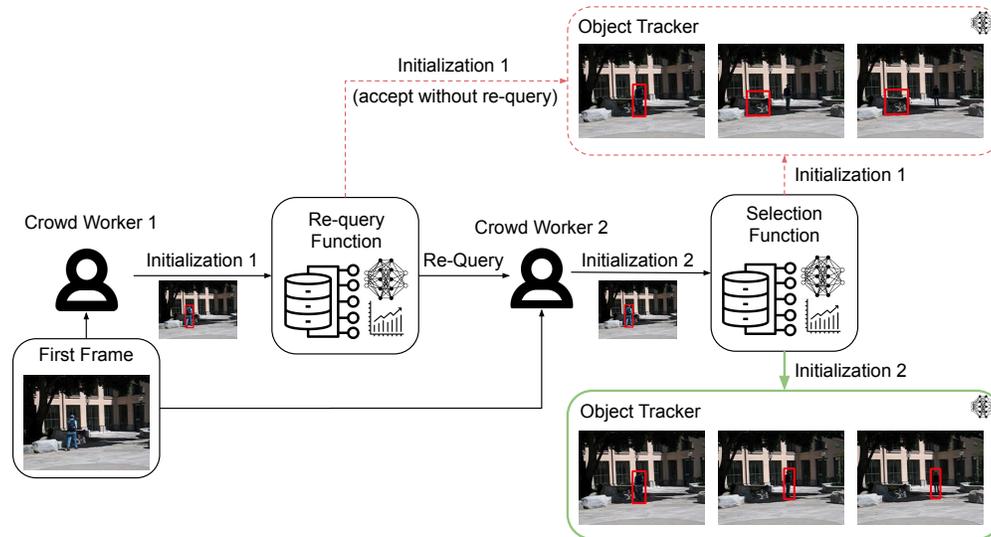


Figure 1: In single-target VOT, the tracker must be initialized with a bounding box to designate which object to follow. *Smart replacement* minimizes the number of initializations for a target accuracy by allowing the *re-query* function to accept the first initialization if it performs well, and allowing the *selection* function to choose between the first and second initialization if re-queried. We see the importance of these functions above, where nearly identical initializations result in dramatically different performances.

ABSTRACT

In single-target video object tracking, an initial bounding box is drawn around a target object and propagated through a video. When this bounding box is provided by a careful human expert, it is expected to yield strong overall tracking performance that can be mimicked at scale by novice crowd workers with the help of advanced quality control methods. However, we show through an investigation of 900 crowdsourced initializations that such quality control strategies are inadequate for this task in two major ways: first, the high level of redundancy in these methods (e.g., averaging multiple responses to reduce error) is unnecessary, as 23% of crowdsourced initializations perform just as well as the gold-standard initialization. Second, even nearly perfect initializations can lead to

degraded long-term performance due to the complexity of object tracking. Considering these findings, we evaluate novel approaches for automatically selecting bounding boxes to re-query, and introduce *Smart Replacement*, an efficient method that decides whether to use the crowdsourced replacement initialization.

CCS CONCEPTS

• **Human-centered computing** → **Interaction paradigms**; • **Information systems** → **Crowdsourcing**; • **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Tracking**.

KEYWORDS

crowd-AI collaboration, crowdsourcing, single-target video object tracking, seed rejection, smart replacement

ACM Reference Format:

Stephan J. Lemmer, Jean Y. Song, and Jason J. Corso. 2021. Crowdsourcing More Effective Initializations for Single-Target Trackers Through Automatic Re-querying. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*, May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3411764.3445181>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI '21, May 8–13, 2021, Yokohama, Japan
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8096-6/21/05.
<https://doi.org/10.1145/3411764.3445181>

1 INTRODUCTION

Inference in single-target video object tracking (VOT) [83] begins with an initialization in the form of a bounding box drawn around a semantically meaningful area. This initialization is then propagated through the remaining video frames despite occlusions, deformations, rotations, and other visual phenomena. Like many other computer vision tasks, remarkable improvement has been shown on this task in recent years due to both the transition from handcrafted features [5, 7, 68] to deep learning-based feature extractors [28, 44, 51, 79, 85] and the establishment of challenges and formal benchmarks [36, 82].

However, like a number of other problems that use both a primary input—such as a video—and a secondary *seed*—such as a bounding box initialization, evaluation assumes that all seeds match the gold-standard [2], an assumption that may be erroneous. Such problems include visual [2] and text [58] question answering, where it is assumed that the seed—the question—always requests information available in the primary input—the text document or image, keypoint-conditioned viewpoint estimation [73], where it is assumed that a given click always indicates the corresponding keypoint on the image, and others [6, 24, 26, 30, 32, 55, 60, 70, 72]. Critically, the assumption of a near gold-standard seed used in previous works neglects two important points: first, one must consider not only the model’s robustness, but also the effect of individual seeds on the output accuracy. Second, the assumption itself is incorrect in cases where the seed is provided by novice annotators (e.g., crowdsourcing), which is known to be less accurate than an expert’s responses.

Because of this, the crowdsourcing literature has studied a number of agreement-based quality control methods such as simple majority voting [2, 13, 15], expectation maximization-based aggregation [31], and POMDP-based dynamic aggregation [12] to obtain reliable responses. However, as we show in this work, these approaches are impractical when applied to inference-time seeds due to both their focus on quality in only the input space (e.g., how well the bounding box fits the target object), which does not directly correlate to model performance, and the requirement of aggregating multiple Human Intelligence Tasks (HITs) for a single sample, which is an unnecessary cost increase in the many cases where the first acquired seed is sufficient.

We demonstrate the limitations of previous approaches through an analysis of a state-of-the-art VOT model [85] initialized with 900 crowdsourced seeds—first-frame bounding boxes—across the 100 videos of the OTB-100 dataset [82]. Through this analysis, we find that the correlation between initialization (seed) accuracy and performance exists, but is not definitive, and that a number of initializations (23.3%) do not need to be re-queried. This motivates the creation of a method that maximizes the performance of a fixed model for minimum number of crowd re-queries.

We begin by evaluating the method under the assumptions of previous work in selective prediction [20] and seed rejection [41]. We then address the shortcomings of these assumptions, and shift selective prediction and seed rejection firmly into the domain of crowdsourcing by acknowledging that the replacement initialization may not be correct. To do this, we introduce the novel evaluation metrics of Replacement Mean Additional Error (RMAE) and

Area under the Replacement Mean Error-coverage Curve (ARMAE), as well as the *Smart Replacement* method shown in Figure 1. Critically, we find that re-query methods which perform best under the assumptions of previous work do not provide the best results in realistic crowdsourcing scenarios, a finding that is likely to generalize to other VOT algorithms, as well as other problems which use a crowdsourced seed when performing inference.

Explicitly stated, our contributions are as follows:

- The novel RMAE and ARMAE metrics for evaluating the effectiveness of automatic re-querying methods when a perfect re-annotation cannot be assumed.
- *Smart Replacement*: An efficient crowd-AI hybrid method that first re-queries seeds that it believes degrade model performance, then chooses whether to use the original or replacement seed.
- An analysis of the quality of crowdsourced bounding box initializations and their effect on the downstream task of single-target VOT, indicating the shortcomings of current crowdsourcing methods and motivating the need for automatic re-querying.
- An evaluation of re-query methods under assumptions of a correct second seed, a crowdsourced second seed, and smart replacement, motivating the need for our metrics and smart replacement method.

2 BACKGROUND AND RELATED WORK

While both single-target VOT and crowdsourcing bounding box annotations are well explored tasks in their respective fields, to the best of our knowledge, this is the first work to consider them jointly in the context of inference. In this section we discuss these two tasks independently, and how they relate to and differ from the problem of rejecting initializations which degrade the performance of the downstream task. Further, we discuss methods used for detecting inputs that have a high likelihood of being classified incorrectly, and how their goals differ from the goals of the presented work, as well as how our novel idea of *smart replacement* contributes to current research on crowd-AI collaboration.

2.1 Single-Target VOT

As the task of single-target VOT has a long history in the computer vision literature, a number of solutions have been proposed. Early work focused on handcrafted local features [5, 7, 68], while more recent works implement deep learning-based approaches through fine-tuning the network at inference time [44, 51, 79], asking the network to directly regress bounding box location [28], or placing template and search images in a common feature space [4, 43]. Our experiments are performed using a method from the latter category, Distractor-Aware Siamese Region Proposal Network (DaSiamRPN) [85], which won the short-term real-time category of the 2018 VOT challenge [34] by improving the negative mining strategy used to create the common feature space of the Siamese Region Proposal Network [43].

To accompany the wide variety of proposed solutions to single-target VOT, a number of evaluation methods have been proposed. Basic metrics have included center error [3], region overlap [23], tracking length [38], failure rate [33], and others [84]. Combinations

of these metrics have been brought together to provide standardized benchmarks by works such as the VOT Challenge [36] and Online Object Tracking Benchmark (OTB) [82]. Aside from standardizing comparisons between single-target VOT works, these benchmarks are notable in the context of this work, as they both acknowledge the potential for small perturbations in the initialization to have an effect on the quality of the returned track. Both test this by allowing shifts of $\pm 10\%$ of the bounding box’s size and aggregating the results across numerous samples on a baseline, with OTB additionally varying the frame used for initialization.

While these experiments provide insight into the performance of various trackers with perturbed initializations, there are two major shortcomings which indicate the need for a different evaluation: first, these works assume that the perturbations are relatively small, and become smaller as the target object becomes smaller. This ignores both the noisy process of crowdsourcing, and the fact that previous work [62] finds annotation error in pixels tends to remain constant as the size of the object changes. Further, while analyzing aggregate performance on perturbed seeds is effective for evaluating tracker robustness, selecting initializations to re-query requires focus on individual object tracks.

2.2 Crowdsourcing Strategies for Collecting Bounding-Box Annotations

The most common goal of generating bounding boxes has been the construction of datasets for training and evaluating computer vision models. Unlike acquiring initializations for inference, this goal allows high-cost, multi-stage crowdsourcing methods, as the cost is spread over many experiments and deployments. For example, the widely cited—over 11,000 times at the time of writing—Microsoft Common Objects in Context (MSCOCO) [45] dataset uses a three-stage review process: in the first stage, multiple workers are asked to determine which object classes are present in an image and label one instance of each object class. In the second stage, multiple workers are then asked to mark all instances of those classes. In the last stage, a single worker who has passed a qualification task segments each previously identified instance. These segmentations are then reviewed by multiple workers to ensure quality, and bounding boxes are created by fitting to the segmentations. A similarly complex approach was taken when producing the YouTube-BoundingBoxes dataset [59], which was used to automatically generate the TrackingNet VOT dataset [49].

Due to the difficulty of procuring high-quality bounding box data through crowdsourcing, some authors choose instead to have all annotations performed under their direct supervision. One notable instance of this is the Pascal Video Object Classes dataset [16], for which all bounding boxes were drawn at a single “annotation party”. This approach is particularly common in the VOT domain, as most works contributed relatively few videos which were later combined into large datasets [5, 18, 36, 39, 48, 56, 61, 65, 66, 82].

While the high accuracy demanded of a dataset justifies complex, expensive, multi-step review processes, a number of works have been proposed to reduce the number of review steps by ranking received annotations in the input space [64, 77] or making the annotation process itself simpler or more accurate through approaches such as a hybrid intelligence [53] or a more intuitive interfaces [54].

Critically, however, these methods are designed to maximize accuracy in the input space, where a re-query method must focus on the accuracy of the downstream task. As we show in Section 3.3, these do not always directly correspond.

2.3 Automatic Identification of Poor Quality Input

While the crowdsourcing literature does not typically evaluate the quality of crowdsourced annotations based on their effect on the downstream task, there exists a body of work called “selective prediction”, which selects difficult inputs—such as images in a classification task—to refer to an oracle annotator. This line of work has been around since the early days of automatic classification [10, 29], and has evolved alongside the various methods for automatic classification from support vector machines [19] to ensembling [76] to boosting [11] to deep neural networks [20]. Since selective prediction has a focus on provable optimality [10, 67, 75] or establishing bounds [20, 22, 76], the majority of the work has been done in the more mathematically-tractable classification domain, though Geifman & El-Yaniv empirically demonstrate performance of their SelectiveNet architecture on a regression task [21].

In parallel, some work has focused on the medical domain for selective prediction, noting the danger of misdiagnosis [40, 47, 50, 57, 74]. While this problem space does not establish the same mathematical rigor due to its focus on an application, the problems are generally still solved with a categorical classification.

We extend upon the works of selective prediction in several ways: first, we perform the task of single-target VOT, to which selective prediction has not been applied. Second, we focus not on finding the worst overall prediction, but re-querying seeds where the given initialization performs worse than the best possible performance for the primary input. Last, and most important, we remove the assumption used in these works that the potentially incorrect seed will always be replaced by a perfect one, moving these tasks firmly into the crowdsourcing domain.

2.4 Crowd-AI Collaboration

Intelligent systems designed to assist humans in decision making are a concept which is well understood by the general public, HCI practitioners, and AI practitioners alike. Such systems have been proposed in domains as varied as medical diagnosis [8], fact checking [52], text classification [37], and topic modeling [69]. The other direction of interaction, in which a human assists the AI model by providing information outside of its perceptual capability, is less well known but no less important. The human-provided information may be used to improve training of the model or provide information to assist inference.

Works in which humans provide information for training the model include AnchorViz [9], which demonstrates how users can help machines learn semantic concepts through example-based anchors, and ModelTracker [1], in which the user and an AI collaboratively examine and debug model performance through a visualization tool. Works in which the human provides information to help with inference include TARS [14], which cleans crowdsourced labels through an automatic oracle, Foureyes [71], which improves annotation accuracy in an image segmentation task by using an

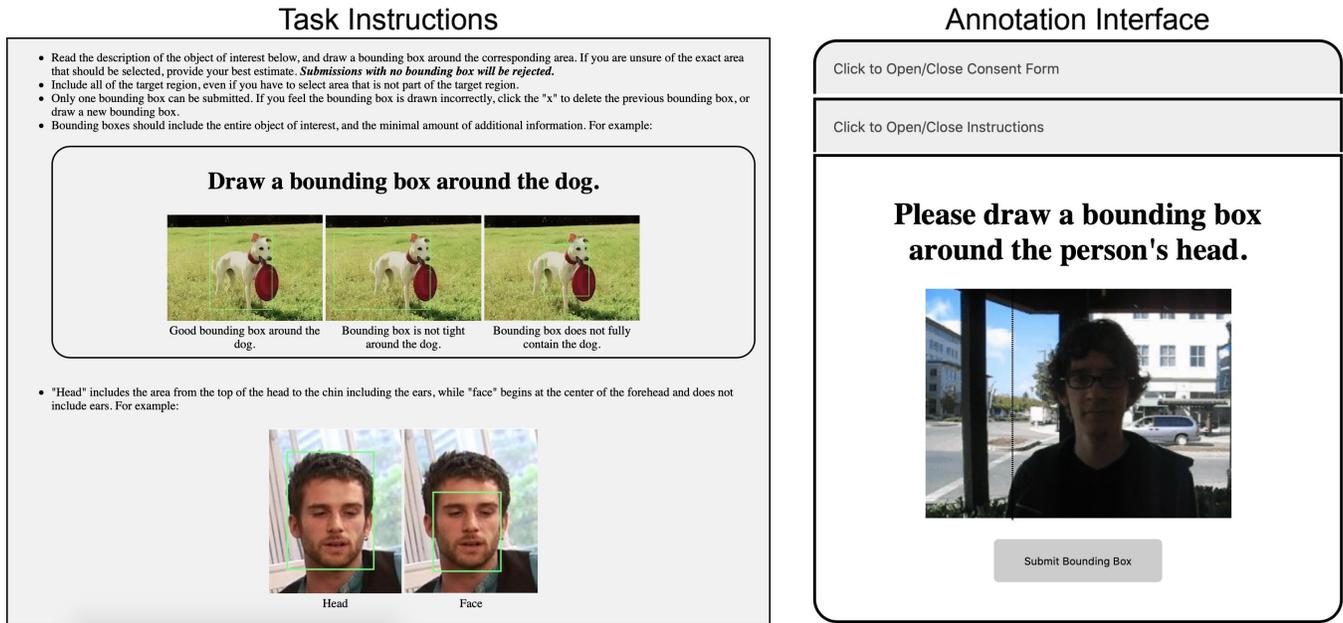


Figure 2: The instructions (left) and interface (right) provided to crowd workers for the bounding box annotation. Workers are given vertical and horizontal guidelines to assist in bounding box construction. Instruction images were taken from CelebA [46] and the ImageNet Large Scale Visual Recognition Challenge [62].

expectation-maximization based aggregation method to combine crowd responses from different interfaces, and C-reference [70], which uses the crowd’s rough estimates to narrow the search space for optimization.

Like the previously mentioned works, single-target VOT requires the crowd worker to provide information that the AI model is unable to determine on its own; specifically, the object which should be tracked. This work contributes to the line of research on crowd-AI collaboration in two notable ways: first, it removes the assumption that the model will perform correctly when given a seed that is correct in the input space, which has been shown to be incorrect in some tasks [41]. Second, we do not focus on optimizing the performance of a model, but instead on how to determine if the information given by the human helps the model and if it is necessary to re-query.

3 QUALITY AND EFFECTS OF CROWDSOURCED INITIALIZATIONS

While some methods used for comparing and evaluating single-target trackers [36, 82] analyze the sensitivity to noise in the initialization, this is done to evaluate model robustness to small perturbations around the gold-standard initialization. This evaluation provides results that neither accurately reflect the characteristics of real-world initializations nor provide the specificity to understand the re-querying problem. In this section, we augment this simplistic evaluation by discussing the distribution of crowdsourced bounding box initializations, failure modes, and the initialization’s effect on the output of the DaSiamRPN model. Through this, we show

that most of the initializations are high quality, many of those of apparent poor quality are actually initializations around an incorrect object, and the fact that a seed has a high IoU with the gold-standard on the first frame does not mean that it will produce a high-quality object track (and vice-versa).

3.1 Data Collection

Using the interface and instructions shown in Figure 2, we ask workers on Amazon Mechanical Turk to provide an initial bounding box based on a text description of the target object. We request nine annotations for each of the 100 videos of the OTB-100 dataset and limit each annotator to one bounding box per video, but do not require that every annotator annotate all videos. Annotators are paid \$0.06 per bounding box drawn, which extrapolates to approximately \$12/hr based on timed data collections performed by the authors. All annotators were located in the United States.

Since bounding box annotations are a common task for crowd workers, we did not perform a qualification task. Instead, we filtered results from inattentive annotators through comparison with the known gold-standard. For this filtering method, we defined a correct annotation as one which has an IoU of greater than 0.5, consistent with designations used in object detection [16, 62, 63]. Annotators for whom more than 15% of annotations are incorrect have their annotations removed from the evaluation set, a threshold chosen based on the fact that we found all annotators had an error rate of less than 15% or greater than 49%. While filtering through a comparison with the gold-standard initialization is not possible during deployment, an analysis of annotation statistics suggests

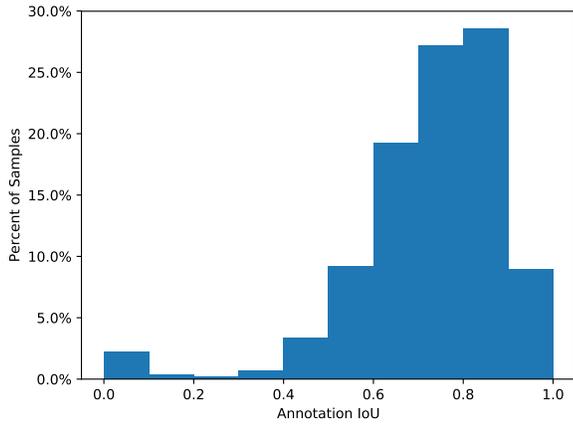


Figure 3: The histogram of first-frame IoU scores between crowdsourced and gold-standard annotations after prefiltering. 93.1% of annotations have an IoU greater than or equal to 0.5, corresponding to a successful detection in the literature [16, 62, 63].

that a simple attention check could identify a good annotator with greater than 94% accuracy with one gold-standard bounding box, and 98% with two gold-standard bounding boxes.

Overall, 26 unique annotators returned 899 of the 900 HITs with bounding boxes. Four annotators annotated more than 90 images, while 14 annotated less than 20 images. Our filtering process resulted in the elimination of four annotators who drew a combined 41 initial bounding boxes, resulting in our evaluation being performed on 858 annotations.

3.2 Image-Space Quality of Bounding-Box Annotations

While our filtering method eliminated inattentive or malicious workers, attentive workers still occasionally make mistakes, leading to the various consensus and review methods discussed in Section 2.2. In Figure 3, we show the distribution of agreement between the annotator initialization and gold standard initialization in terms of IoU. Overall, we find that 93.1% of the filtered annotations meet our definition of correct by having an IoU greater than 0.5, and most (55.4%) have an IoU with the gold-standard between 0.7 and 0.9. While a relatively small percentage (8.9%) fall within the top range of 0.9-1.0, previous work [36] has suggested that bounding boxes do not need to have an IoU near 1 to be perceptually similar, which we illustrate in Figure 4. Of the 858 accepted annotations, we examine the 59 annotations which meet our definition of incorrect to determine the failure modes of the human annotators. Broadly, we find the failure modes fit into the three categories shown in Figure 5: misidentified target, wrong region on target, and poorly fit. Overall, 15 (25.4%) were a misidentified target, 28 (47.5%) identified the wrong region of the target, and 16 (27.1%) were poorly fit.

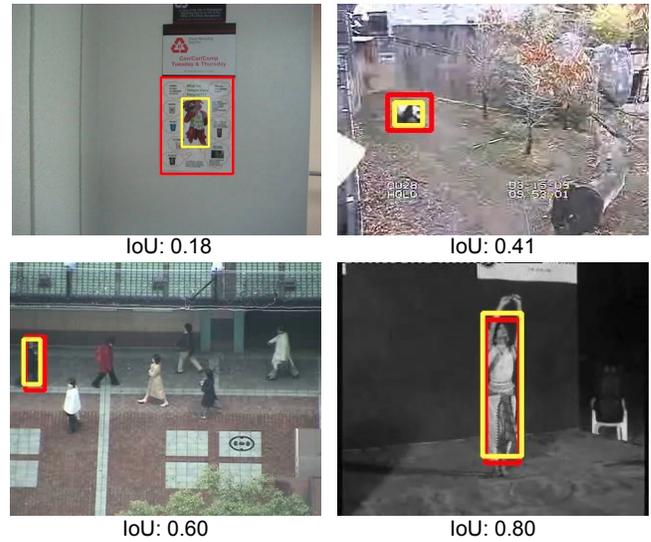


Figure 4: Example bounding boxes, showing perceptual similarity of various IoUs.

3.3 Effect of Initialization on Tracker Performance

Unlike previous crowdsourcing work which focuses solely on the quality of the initialization with respect to the image, we focus on the quality of the initialization with respect to the task of single-target VOT. This means that in addition to the input-space quality analysis performed in the previous section, we need to evaluate the relationship between the quality of the seed in the input space and the quality of the resulting object track.

One critical element in this relationship is the effect of the three types of incorrect annotations described in the previous section. In the case of a poor fit, the task is defined appropriately, but the bounding box is too small, too large, or offset in such a way that the tracker receives a different set of features to match in subsequent frames. Based on the quality of these features, the video, and the tracker, this may result in anything from a significant degradation to a significant improvement over a gold-standard initialization. An ideal re-query mechanism would be able to determine the effect of the poor fit on the downstream task. The other two failure modes

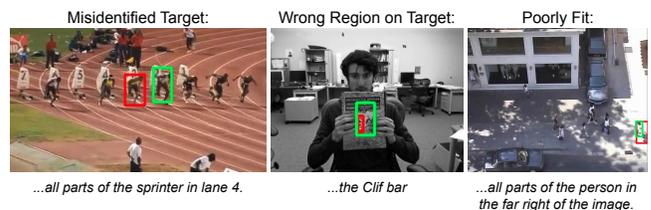


Figure 5: Examples of the three categories of annotation error. The red box represents the crowdsourced initialization, the green box represents the gold-standard initialization.

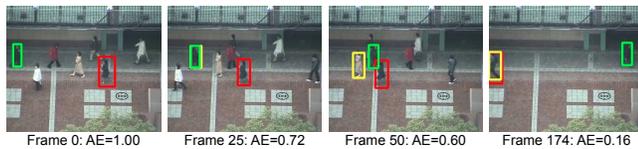


Figure 6: Calculating additional error exclusively on valid frames produces a more meaningful evaluation metric. Despite the candidate initialization (red) tracking a different object than the gold-standard initialization (yellow), additional error is low when all 174 frames are used, but high if only valid frames (frames 0-50) are used.

produce an object track that is substantially different from the target object track, but still follows a meaningful semantic object. That is, the tracker may do its task well, but it was told to do the wrong task. This makes it impossible for an automated system to detect these failure modes without prior understanding of objects that are typically tracked.

While only 16 of our 858 (1.9%) accepted annotations are incorrect in a way that is detectable, we again highlight the fact that an initialization should not be re-queried based on its quality in the input space, but instead how it affects the performance on the downstream task. As we see in Figure 1, these two are not always the same. Throughout the remainder of this section, we will discuss the agreement (or lack thereof) between these two metrics, and how it affects the task of re-querying initializations.

3.3.1 Metric. In order to quantify the effect of the initialization on subsequent frames, we use the additional error [41]. This metric compares the performance of the model when given the gold-standard initialization with the performance of the model when given a candidate—in this case, crowdsourced—initialization. Notably, it addresses the need to separate inferences that are performed poorly due to a bad initialization from inferences that are performed poorly due to a difficult video by designating that the gold-standard initialization provides the best possible performance. This is enforced by constraining additional error to be greater than or equal to zero. We further enforce this constraint in the single-target VOT task by finding the frame where a tracker initialized with the gold-standard seed no longer has any overlap with the gold-standard frame-by-frame annotation. Similar to the designation of valid frames used by Kristan et al. [36], frames prior to this zero-intersect frame are considered valid and used in our additional error calculation. We can see the effect of designating frames after a lost gold-standard track as invalid in Figure 6, where processing on invalid frames results in a relatively low additional error, despite the candidate initialization having zero IoU with the gold-standard.

Put together, this gives the formula for additional error in single-target VOT:

$$AE(T_t, T_{gs}, T_c) = \max \left(\frac{1}{N_v} \sum_{n=1}^{N_v} \frac{t_t^n \cap t_{gs}^n}{t_t^n \cup t_{gs}^n} - \frac{1}{N_v} \sum_{n=1}^{N_v} \frac{t_t^n \cap t_c^n}{t_t^n \cup t_c^n}, 0 \right). \quad (1)$$

In this equation, t^n represents the bounding box around the target at the n^{th} frame of the object track given by T . The object track may

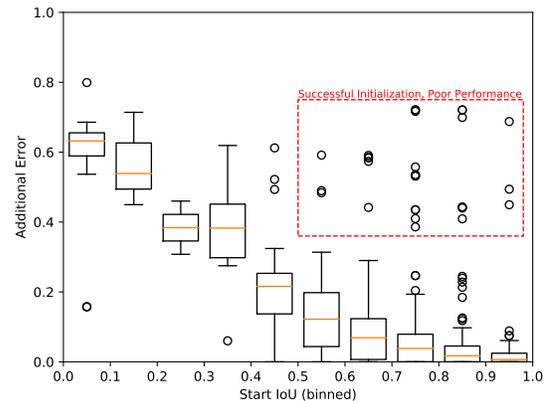


Figure 7: The IoU of the crowdsourced initialization with the gold-standard initialization is not the sole determining factor in an initialization’s performance, as evidenced by successful initializations with poor performance.

be obtained by using the gold-standard per-frame target annotation (T_t), running the object tracker with the gold standard initialization (T_{gs}), or running the object tracker with the crowdsourced candidate initialization (T_c). N_v represents the number of valid frames in the video.

3.3.2 Effect of Initialization on Performance. In Figure 7, we show the relationship between the IoU of the crowdsourced and gold-standard initializations and additional error, and draw two conclusions: first, following current HCI methods which accept or reject an initialization based entirely on the input space (i.e., first-frame IoU) will re-query many initializations that are acceptable and miss many that should be re-queried. This observation is due not only to outliers at high initialization IoUs, but also to substantial overlaps within inliers. For example, the additional error at the third quartile of the 0.7-0.8 IoU bin is near the median of the 0.6-0.7 IoU bin, meaning that it is better to re-query the worst quarter of the 0.7-0.8 bin before the best half of the 0.6-0.7 bin.

Second, we see that the second quartile begins at zero additional error for the highest three start IoU bins, and at approximately 0.005 for the fourth highest start IoU bin. In total, 200 of our 858 (23.3%) annotations result in no additional error. We also calculate this value using one randomly chosen initialization per video, repeated for 1000 trials. Under this scenario, an average of 22.93% ($\sigma = 2.75\%$) of initializations result in no additional error.

These two findings support the idea that using crowdsourcing strategies which focus on the quality of the initialization in the input space will not perform optimally when the goal is to produce the best possible output. This is for two reasons: first, *annotations that appear good in the input space may still cause significant error on the downstream task*. Second, requiring multiple touches for every datapoint is an unnecessary expense, as *about 23% of initializations cause no additional error*.

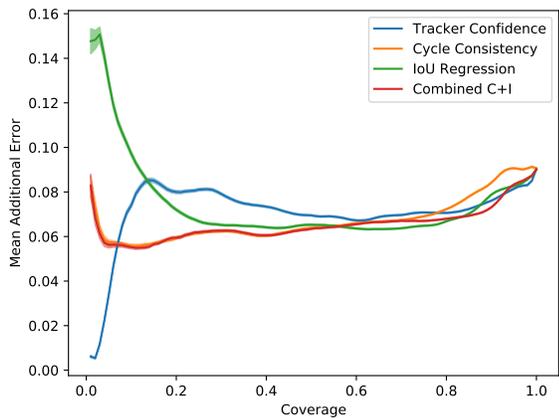


Figure 8: The mean additional error for all methods and coverages. The shaded region represents one standard error.

4 RE-QUERYING WITH ORACLE REPLACEMENT

In the previous section, we examined the characteristics of crowdsourced initializations, and showed that methods which focus on the accuracy in the input space—analogue to current HCI methods—will both re-query initializations that should not be re-queried, and accept initializations that should. This motivates the development of a novel method to determine when to re-query an initialization in single-target VOT. In this work, we assume that the decision to re-query an initialization is made by thresholding on a scoring function and consider four scoring functions:

- (1) **Tracker Confidence:** The Distractor-Aware Siamese Region Proposal Network architecture [85] that we use for our experiments returns a confidence score, which corresponds to the confidence that a bounding box contains the object which is being tracked. In the tracking algorithm, this score is used to select between potential bounding boxes, as well as to determine when a track has been lost and re-acquired for long-term tracking. We use the mean of this tracking score over all valid video frames to produce a single metric.
- (2) **Regression of per-frame IoU:** While it is not tractable to train a model to predict the additional error directly due to the large number of non-parallelizable inferences that would need to be performed for a train step, it is possible to predict the IoU of a bounding box on a given frame, similar to the work of Gurari et al. [25]. We formulate this regression as a classification problem that attempts to predict which of 10 evenly spaced IoU bins the bounding box has with the gold-standard, which is hidden from the classifier. We use a pretrained ResNet-18 [27] backbone, add a fourth input channel which accepts a binary mask representing the candidate bounding box, and train the model using perturbed bounding boxes on the MSCOCO [45] dataset. The predicted IoU on a single frame is the mean of the classifier’s output distribution, while the re-query score is the mean of predictions across all valid frames.

	AMAE	Standard Error
Combined C+I	0.06541	0.000388
IoU Regression	0.08178	0.000480
Cycle Consistency	0.06735	0.000379
Tracker Confidence	0.07070	0.000357

Table 1: AMAE of evaluated re-query functions.

- (3) **Cycle Consistency:** Proposed by Wu et al. [81] to evaluate trackers without a ground-truth annotation, cycle consistency appends the reversed video to the end of the forward video, and runs the tracker across this forward-backward video. The final bounding box is compared with the initial bounding box, and the more they agree, the higher the predicted quality of the track. The original work used the distance between the prior and posterior densities as the comparison, but since this is not compatible with modern methods, we use the IoU between the initial and final bounding boxes.
- (4) **Cycle Consistency + IoU Regression (Combined C+I):** The cycle consistency score discussed above has a lower bound of zero, which occurs for any initialization for which the track is lost. This means that no distinction is made between tracks that are lost early in the video and tracks that are lost late. To compensate for this, we separate cycle consistency scores into “hit” and “miss” bins based on the previously defined threshold of 0.5. The hit bin is accepted in order of the cycle consistency score, then the miss bin is accepted in the order given by the IoU regressor.

To calculate the performance of a re-query function we use the mean additional error [41], which corresponds to the mean of the additional error for a set of accepted samples from dataset \mathcal{D} :

$$MAE(g, \mathcal{D}) = \frac{\frac{1}{|\mathcal{D}|} \sum_{(x, T_t, T_{gs}, T_c) \in \mathcal{D}} g(x, T_c) AE(T_t, T_{gs}, T_c)}{\frac{1}{|\mathcal{D}|} \sum_{(x, T_c) \in \mathcal{D}} g(x, T_c)} . \quad (2)$$

Here, the additional error, AE , is given by Equation 1. g is our re-query function, which uses the video, x , and object track generated using the candidate seed, T_c , to decide whether to accept the initialization by returning a one or select it for re-query by returning a zero. Since the re-query function is implemented by applying a threshold to a continuous scoring function, we also have the measurement of *coverage*, which is the proportion of seeds which are accepted. Due to the relationship between coverage and MAE, our analysis shows the mean additional error at every coverage, and uses the area under that curve—the area under the MAE-Coverage curve (AMAE)—to find a single value for comparison between different methods.

Since we collect nine initializations for every video, and the AMAE metric requires one initialization for each of the 100 videos, we take 1,000 samples of the approximately 9^{100} potential combinations of initializations. These samples are then used to calculate the mean and standard error of the evaluated re-query functions.

We see in Table 1 that the Combined C+I method outperforms all other methods due to the complementary behavior of the IoU

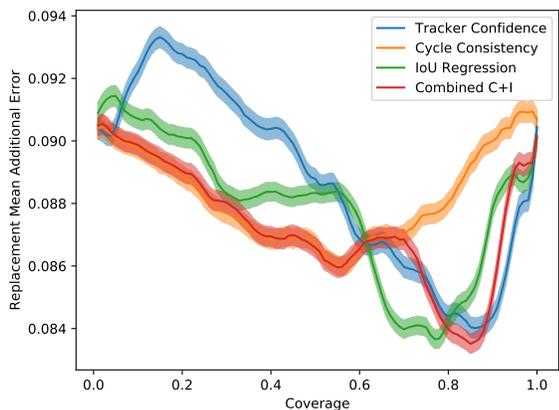


Figure 9: The RMAE when the second initialization is always used (Naive Replacement). Note the low RMAE near 85% coverage, where the worst seeds are being replaced, and the replacements have a high likelihood of improving the result.

regressor and the cycle consistency metric, which can be seen in Figure 8. The high performance of the cycle consistency metric at low coverages means that it can effectively determine when a track is good, but it does not maintain the high level of performance through higher coverages. In short, if the IoU between the initialization and final bounding boxes is very high, the track is guaranteed to be high quality, but if it is not, the quality of the track is estimated poorly.

In contrast, the IoU regression can not distinguish well between tracks which are high quality and tracks which are medium quality, leading to poor performance low coverages. However, this metric does have the ability to detect the exact frame at which the track is lost, or if the track is lost and recovered. This translates to a strong ability to discriminate between the high-error initializations which the cycle consistency metric groups together.

5 SMART REPLACEMENT: RE-QUERYING WITH CROWD REPLACEMENT

The previous section makes the implicit assumption that once a poor initialization has been found, it will be corrected perfectly by an oracle annotator. While this assumption is used in previous work (Section 2.3), it is inappropriate for the case where the re-query may result in worse performance than the initialization it was meant to replace. In this section, we introduce *smart replacement* and the related Replacement Mean Additional Error (RMAE) and Area under the Replacement Mean Additional Error curve (ARMAE) metrics. We first use these metrics to highlight the importance of smart replacement via an evaluation of model performance when the re-queried seed is always used (naive replacement). Next, we compare several potential scoring functions under the assumption that the same scoring function is used for both the re-query and selection functions. Last, we compare the performance of different re-query functions when the same scoring function is used for selection and find that despite the apparent complementary abilities of the

	ARMAE	Standard Error
Combined C+I	0.08710	0.000261
IoU Regression	0.08792	0.000252
Cycle Consistency	0.08839	0.000265
Tracker Confidence	0.08896	0.000263

Table 2: ARMAE of re-query methods when the second initialization is always used (Naive Replacement).

different methods, the same method still provides the best overall smart replacement performance.

Smart Replacement and RMAE. In order to produce a system that performs well when the re-query returns a noisy seed, we introduce *smart replacement*, which adds a selection function to the re-query function. Like the re-query function, the selection function makes a binary decision based on the output of a scoring function. However, while the re-query function applies a threshold to the output of the scoring function, the selection function directly compares scores to choose between two potential initializations. We evaluate the combined performance of our re-query and selection functions using a novel metric, *replacement mean additional error* (RMAE). Unlike the MAE, the RMAE is calculated using additional error from all videos regardless of the output of the re-query function, and takes into account the performance of both the first seed and the seed received after re-query.

More precisely, we have a re-query function, g , and a selection function h , which operate on the object track created from the first initialization, T_{c1} , and the object track created from the second initialization, T_{c2} . The RMAE is then:

$$RMAE(g, h, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x, T_t, T_{g_s}, T_{c1}, T_{c2}) \in \mathcal{D}} g(x, T_{c1})AE(T_t, T_{g_s}, T_{c1}) + (1 - g(x, T_{c1}))(h(x, T_{c1}, T_{c2})AE(T_t, T_{g_s}, T_{c1}) + (1 - h(x, T_{c1}, T_{c2}))AE(T_t, T_{g_s}, T_{c2})) , \quad (3)$$

Intuitively, this metric represents a choice between three options for each datapoint in \mathcal{D} : if the first seed is accepted, $g(x, T_{c1}) = 1$ and the additional error from the first seed contributes to the RMAE. If the first seed is re-queried, $1 - g(x, T_{c1}) = 1$ and the selection function is used to choose which additional error contributes to the RMAE. If $h(x, T_{c1}, T_{c2}) = 1$, the first initialization is used in the RMAE calculation, whereas if $1 - h(x, T_{c1}, T_{c2}) = 1$ the replacement initialization is used.

Naive Replacement. To illustrate the importance of this metric and the selection function, we evaluate our re-query functions using our RMAE-based metrics and compare the results to those obtained when an oracle re-query is assumed. We do not implement a selection function for this experiment, meaning $h(x, T_{c1}, T_{c2}) = 0 \forall (x, T_{c1}, T_{c2}) \in \mathcal{D}$. Similar to the previous evaluation, we perform our analysis on 1,000 randomly drawn sets of initializations. However, instead of selecting one initialization per video, we select two initializations per video without replacement corresponding to the original query (T_{c1}) and the re-query (T_{c2}). The results of this evaluation are shown in Figure 9 and Table 2.

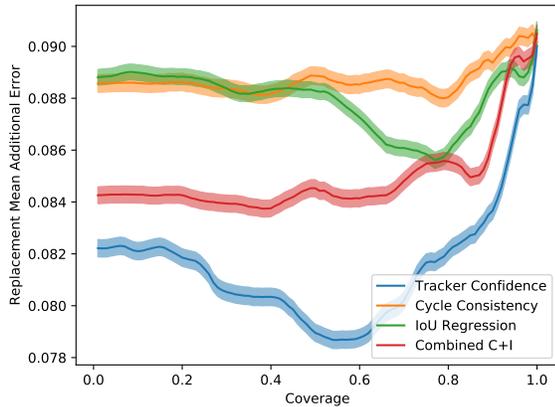


Figure 10: RMAE-Coverage curves for the case where the re-query and selection functions are the same.

As expected, the ARMAE is substantially higher than the AMAE, as there is additional error contributed by the replacement seeds that is not considered when calculating the AMAE. Under the new assumptions, the combined C+I scoring function still performs best, though the IoU regressor now outperforms the cycle consistency score on the mean. We speculate that this is due to the fact that cycle consistency is excellent at detecting lost tracks, which are often related to qualities of the video, such as occlusion, motion blur, fast motion, and others. Due to this, there is a high likelihood that replacing an initialization that causes a lost track will still result in a lost track. The IoU regressor, on the other hand, detects lost tracks less reliably than the cycle consistency metric, but is more sensitive to failure modes that are video independent, meaning the samples it re-queries first are highly likely to benefit from being replaced.

We also note that three of our four methods have minima near 85% coverage. At this high coverage, these methods identify poorly performing initializations well, which results in replacement initializations having a high likelihood of outperforming the first initialization. As the coverage decreases (we replace more and better initializations), it is no longer a near guarantee that the replacement initialization will provide better performance, and the replacement initializations may perform worse than the initialization they were meant to replace. This is further supported by the fact that cycle consistency is the sole re-query function which does not have this minimum, since it does not have any discriminative power within the 20% of samples for which it returns a score of zero.

	ARMAE	Standard Error
Combined C+I	0.08600	0.000296
IoU Regression	0.08796	0.000282
Cycle Consistency	0.08843	0.000305
Tracker Confidence	0.08205	0.000291

Table 3: ARMAE of various smart replacement methods, where the re-query and selection functions are the same.

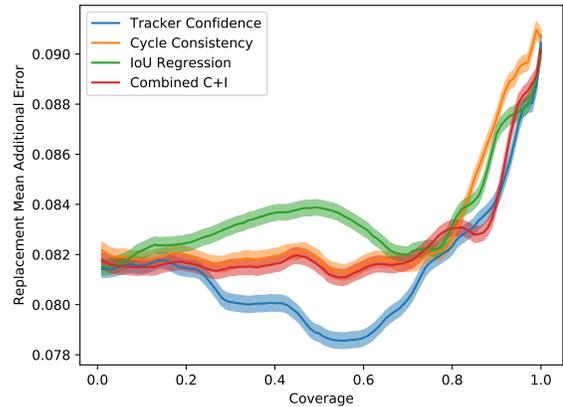


Figure 11: RMAE-Coverage curves for the case where tracker confidence is used as the selection function.

Matching Re-query and Selection Functions. For the first evaluation of smart replacement, we use the same scoring function for both the re-query (g) and selection (h) functions. We see the results of this in Figure 10 and Table 3. We note that IoU regression and cycle consistency do not perform significantly better using a selection function, while the C+I and Tracker Confidence methods do, demonstrating that a scoring function that can effectively differentiate between absolute performance of initializations on different videos—its inter-video discrimination—may not effectively differentiate between performances of initializations on a single video—its intra-video discrimination—and vice-versa.

In the case of the two methods that do not improve their ARMAE significantly using smart replacement, we see that they do outperform naive replacement at zero coverage, where every initialization has two options. However, their intra-video discrimination is so weak that the cases which cause the dips seen at high coverages in Figure 9 may not necessarily be replaced by the new annotation, leading to a higher overall ARMAE. In contrast, while the tracker confidence is a poor performer under the assumptions of Section 4 due to its poor inter-video discriminative ability, its strong intra-video discriminative ability makes it the best performing scoring function under the assumption that the second seed will be provided by a crowd worker.

Tracker Confidence as Universal Selection Function. After observing that tracker confidence has the best intra-video discriminative ability, we remove the assumption that the re-query and selection functions are identical and use tracker confidence as our selection

	ARMAE	Standard Error
Combined C+I	0.08269	0.000294
IoU Regression	0.08351	0.000290
Cycle Consistency	0.08294	0.000289
Tracker Confidence	0.08205	0.000291

Table 4: ARMAE of various re-query functions where the tracker confidence is used as the selection function.

function for all re-query functions. Ideally, this combines the best inter-video performance (combined C+I) with the best intra-video performance (tracker confidence) to create the best overall performance. However, as we see in Figure 11 and Table 4, this is not the case.

While all of the methods are significantly improved by using tracker confidence as the selection function, using the tracker confidence as both the re-query and selection functions outperforms the other combinations despite the combination's ostensibly poorer inter-video performance. This suggests that the tracker confidence not only determines the better of two object tracks on a video, but also re-queries initializations on videos for which it can effectively differentiate between good and bad initializations.

The fact that the best re-query and selection function under the more realistic assumptions of this section was a poor performer under the assumptions of previous works highlights the importance of the RMAE-based metrics in a crowdsourcing scenario. In short, if we use the Combined C+I method which is best under the assumptions of previous works, we receive suboptimal performance when replacing bad initializations with other crowdsourced initializations.

6 DISCUSSION

In this work, we addressed the task of re-querying crowdsourced seeds, as motivated by single-target VOT. To demonstrate the shortcomings of current crowdsourcing methods on this task, we performed an evaluation of 900 crowdsourced initializations and found that while most initializations met our definition of good in the input space, a good initialization does not guarantee a good outcome. This shows that current methods used for crowdsourcing are insufficient, as they will both re-query initializations that shouldn't be re-queried, and accept initializations that should.

On this task, we also show that current evaluation methods are unable to effectively capture the reality that a crowdsourced replacement seed may not improve performance, an assumption that is used extensively in previous work. To address this, we develop a novel evaluation method based on the Replacement Mean Additional Error (RMAE) metric, and use it to evaluate four re-query and selection functions: tracker confidence, regression of per-frame IoU, cycle consistency, and cycle consistency combined with IoU regression. Critically, we find that the method which provides the best performance under the more realistic assumptions of our work does not provide the best performance under the assumptions of previous work, illustrating the importance of our novel RMAE and ARMAE metrics.

The findings of this work have applications to both other tracking algorithms and other tasks which rely on a crowdsourced seed. With respect to other object tracking algorithms, the best re-query and selection functions are specific to the DaSiamRPN architecture [85] which has a number of derivatives [17, 42, 78, 80] that in general produce state-of-the-art results on recent object tracking challenges [35]. For significantly different architectures, the re-query and selection functions will respond to the same failure modes, but these architectures may be more susceptible to different failure modes. It follows that these experiments outlined in

this paper can be repeated with the different tracking algorithm to determine the best scoring function.

In addition, as other work [41] shows similarly complex relationships between the input seed and output accuracy in other tasks, we believe three of the fundamental findings of this work will extend outside of the task of video object tracking: the high redundancy of current crowdsourcing methods when applied to this task is an unnecessary cost, that it is insufficient to optimize in the input space only, and that the assumptions of a clean replacement, and thus metrics, used in seed rejection and selective prediction are not valid in the crowdsourcing domain. It follows that by defining appropriate scoring functions and a measure of additional error, smart replacement can be used on a variety of tasks, including visual [2] and text [58] question answering, keypoint-conditioned viewpoint estimation [73], hierarchical scene classification [32], and pose estimation with dimension lines [72].

One major limitation of this study is the assumption that the gold-standard initialization provides optimal performance. As in previous work [41], we compare performance of the candidate initialization to the gold-standard initialization to encourage the re-query mechanism to select videos which would be improved by re-query, as opposed to videos for which there exists no initialization which performs well. Unlike previous work, however, single-target VOT has no single correct seed, which means that the gold-standard initialization might not match the optimal initialization. This means that some of the videos evaluation assumes cannot be improved may still have room for improvement.

7 CONCLUSION

In this work, we addressed re-querying crowdsourced seeds, as motivated by the task of single-target video object tracking. The need for this was shown through an evaluation of 900 crowdsourced initializations and their effect on the quality of the corresponding track, which demonstrated that current crowdsourcing methods would both re-query initializations that should not be re-queried, and accept initializations that should. We then explored methods for selecting seeds to re-query, and introduced novel metrics that reflect a more realistic crowdsourcing scenario. Critically, we find that methods which perform the best under previous assumptions—namely that the re-query results in optimal performance—do not necessarily perform best under metrics which account for the quality of the re-initialization. Further, while the evaluation focuses on one implementation of the task of single-target VOT, we discuss the extensibility of our procedures to other tasks. Overall, we believe these findings will be critical in reducing the number of queries—and therefore the cost—for an acceptable performance, making applications which use such tasks accessible to a substantially larger audience.

ACKNOWLEDGMENTS

Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1I1A1A0107238511).

REFERENCES

- [1] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM Press, Seoul, Republic of Korea, 337–346. <https://doi.org/10.1145/2702123.2702509>
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile, 2425–2433. <https://doi.org/10.1109/ICCV.2015.279>
- [3] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. 2009. Visual tracking with online Multiple Instance Learning. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, 983–990. <https://doi.org/10.1109/CVPR.2009.5206737> ISSN: 1063-6919.
- [4] Luca Bertinetto, Jack Valmader, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In *European conference on computer vision*. Springer International Publishing, Amsterdam, The Netherlands, 850–865.
- [5] S. Birchfield. 1998. Elliptical head tracking using intensity gradients and color histograms. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, Santa Barbara, California, USA, 232–237. <https://doi.org/10.1109/CVPR.1998.698614>
- [6] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. Visual Recognition with Humans in the Loop. In *European Conference on Computer Vision*, Vol. 6314. Springer Berlin Heidelberg, Berlin, Heidelberg, 438–451. https://doi.org/10.1007/978-3-642-15561-1_32
- [7] C. Cafforio and F. Rocca. 1976. Methods for measuring small displacements of television images. *IEEE Transactions on Information Theory* 22, 5 (Sept. 1976), 573–579. <https://doi.org/10.1109/TIT.1976.1055602>
- [8] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland UK, 14.
- [9] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating classifier error discovery through interactive semantic data exploration. In *23rd International Conference on Intelligent User Interfaces*. ACM Press, Tokyo, Japan, 269–280.
- [10] C. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16, 1 (Jan. 1970), 41–46. <https://doi.org/10.1109/TIT.1970.1054406>
- [11] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. 2016. Boosting with Abstention. In *Advances in Neural Information Processing Systems*. Curran Associates, Barcelona, Spain, 9 pages.
- [12] Peng Dai, Christopher H. Lin, Mausam, and Daniel S. Weld. 2013. POMDP-based control of workflows for crowdsourcing. *Artificial Intelligence* 202 (Sept. 2013), 52–85. <https://doi.org/10.1016/j.artint.2013.06.002>
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, Miami, FL, 8 pages.
- [14] Mohamad Dolatshah. 2018. *Cleaning crowdsourced labels using oracles for statistical classification*. Ph.D. Dissertation. Applied Sciences: School of Computing Science.
- [15] I. Endres, A. Farhadi, D. Hoiem, and D. A. Forsyth. 2010. The benefits and challenges of collecting richer object annotations. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. IEEE, San Francisco, CA, 1–8. <https://doi.org/10.1109/CVPRW.2010.5543183>
- [16] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (June 2010), 303–338. <https://doi.org/10.1007/s11263-009-0275-4>
- [17] Heng Fan and Haibin Ling. 2019. Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 7944–7953. <https://doi.org/10.1109/CVPR.2019.00814>
- [18] Robert B Fisher. 2004. The PETS04 Surveillance Ground-Truth Data Sets. In *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*. IEEE, Prague, Czech Republic, 5.
- [19] Giorgio Fumera and Fabio Roli. 2002. Support Vector Machines with Embedded Reject Option. In *Pattern Recognition with Support Vector Machines*, Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Seong-Whan Lee, and Alessandro Verri (Eds.), Vol. 2388. Springer Berlin Heidelberg, Berlin, Heidelberg, 68–82. https://doi.org/10.1007/3-540-45665-1_6
- [20] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Long Beach, CA, USA, 4878–4887.
- [21] Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th international conference on Machine learning*. ACM Press, Long Beach, CA, USA, 2151–2159.
- [22] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. 2019. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers. In *Proceedings of the Seventh International Conference on Learning Representations*. OpenReview.net, New Orleans, LA, 14 pages.
- [23] M. Godec, P.M. Roth, and H. Bischof. 2013. Hough-based tracking of non-rigid objects. *Computer Vision and Image Understanding* 117, 10 (Oct. 2013), 1245–1256. <https://doi.org/10.1016/j.cviu.2012.11.005>
- [24] Brent Griffin, Victoria Florence, and Jason J. Corso. 2020. Video Object Segmentation-based Visual Servo Control and Object Depth Estimation on a Mobile Robot Platform. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, Snowmass Village, CO, 1647–1657. arXiv:1903.08336 <http://arxiv.org/abs/1903.08336>
- [25] Danna Gurari, Suyog Dutt Jain, Margrit Betke, and Kristen Grauman. 2016. Pull the Plug? Predicting If Computers or Humans Should Segment Images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 382–391. <https://doi.org/10.1109/CVPR.2016.48>
- [26] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 3608–3617. <https://doi.org/10.1109/CVPR.2018.00380>
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Las Vegas, NV, 770–778.
- [28] David Held, Sebastian Thrun, and Silvio Savarese. 2016. Learning to Track at 100 FPS with Deep Regression Networks. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9905. Springer International Publishing, Cham, 749–765. https://doi.org/10.1007/978-3-319-46448-0_45
- [29] M. E. Hellman. 1970. The Nearest Neighbor Classification Rule with a Reject Option. *IEEE Transactions on Systems Science and Cybernetics* 6, 3 (July 1970), 179–185. <https://doi.org/10.1109/TSSC.1970.300339>
- [30] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. Segmentation from Natural Language Expressions. In *Proceedings of the 2016 European Conference on Computer Vision*. Springer International Publishing, Amsterdam, The Netherlands, 108–124.
- [31] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*. ACM Press, Washington DC, 64–67. <https://doi.org/10.1145/1837885.1837906>
- [32] Michal Koperski, Tomasz Konopczynski, Rafal Nowak, Piotr Semberecki, and Tomasz Trzcinski. 2020. Plugin Networks for Inference under Partial Evidence. In *Proceedings of The 2020 IEEE Winter Conference on Applications of Computer Vision*. IEEE, Snowmass Village, CO, 2883–2891.
- [33] Matej Kristan, Stanislav Kovacic, Aleš Leonardis, and Janez Pers. 2010. A Two-Stage Dynamic Model for Visual Tracking. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40, 6 (Dec. 2010), 1505–1520. <https://doi.org/10.1109/TSMCB.2010.2041662> Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).
- [34] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman El-desokey, Gustavo Fernandez, Alvaro Garcia-Martin, Alvaro Iglesias-Arias, A. Aydin Alatan, Abel Gonzalez-Garcia, Alfredo Petrosino, Alireza Memarmoghdam, Andrea Vedaldi, Andrej Muhic, Anfeng He, Arnold Smeulders, Asanka G. Perera, Bo Li, Boyu Chen, Changick Kim, Changsheng Xu, Changzhen Xiong, Cheng Tian, Chong Luo, Chong Sun, Cong Hao, Daijin Kim, Deepak Mishra, Deming Chen, Dong Wang, Dongyoon Wee, Efstratios Gavves, Erhan Gundogdu, Erik Velasco-Salido, Fahad Shahbaz Khan, Fan Yang, Fei Zhao, Feng Li, Francesco Battistone, George De Ath, Gorthi R. K. S. Subrahmanyam, Guilherme Bastos, Haibin Ling, Hamed Kiani Galoogahi, Hankyeol Lee, Haojie Li, Haojie Zhao, Heng Fan, Honggang Zhang, Horst Possegger, Houqiang Li, Huchuan Lu, Hui Zhi, Huiyun Li, Hyemin Lee, Hyung Jin Chang, Isabela Drummond, Jack Valmadre, Jaime Spencer Martin, Javean Chahl, Jin Young Choi, Jing Li, Jinqiao Wang, Jinqing Qi, Jinyoung Sung, Joakim Johannander, Joao Henriques, Jongwon Choi, Joost van de Weijer, Jorge Rodriguez Herranz, Jose M. Martinez, Josef Kittler, Junfei Zhuang, Junyu Gao, Klemen Grm, Lichao Zhang, Lijun Wang, Lingxiao Yang, Litu Rout, Liu Si, Luca Bertinetto, Lutao Chu, Manqiang Che, Mario Edoardo Maresca, Martin Danelljan, Ming-Hsuan Yang, Mohamed Abdel-pakey, Mohamed Shehata, Myunggu Kang, Namhoon Lee, Ning Wang, Ondrej Miksik, P. Moallem, Pablo Vicente-Monivar, Pedro Senna, Peixia Li, Philip Torr, Priya Mariam Raju, Qian Ruihe, Qiang Wang, Qin Zhou, Qing Guo, Rafael Martin-Nieto, Rama Krishna Gorthi, Ran Tao, Richard Bowden, Richard Everson, Runling Wang, Sangdoon Yun, Seokwon Choi, Sergio Vivas, Shuai Bai, Shuangping Huang, Sihang Wu, Simon Hadfield, Siwen Wang, Stuart Golodetz, Tang Ming, Tianyang

- Xu, Tianzhu Zhang, Tobias Fischer, Vincenzo Santopietro, Vitomir Struc, Wang Wei, Wangmeng Zuo, Wei Feng, Wei Wu, Wei Zou, Weiming Hu, Wengang Zhou, Wenjun Zeng, Xiaofan Zhang, Xiaohe Wu, Xiao-Jun Wu, Xinmei Tian, Yan Li, Yan Lu, Yee Wei Law, Yi Wu, Yiannis Demiris, Yicai Yang, Yifan Jiao, Yuhong Li, Yunhua Zhang, Yuxuan Sun, Zheng Zhang, Zheng Zhu, Zhen-Hua Feng, Zhihui Wang, and Zhiqun He. 2018. The sixth Visual Object Tracking VOT2018 challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. Springer International Publishing, Munich, Germany, 52 pages.
- [35] Matej Kristan, J Matas, A Leonardis, Michael Felsberg, et al. 2020. The eighth visual object tracking VOT2020 challenge results. In *Proceedings of the 16th European Conference Computer Vision Workshops*. Springer International Publishing, Virtual, 56 pages.
- [36] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojtíš, Roman Pflugfelder, Gustavo Fernández, Georg Nebel, Fatih Porikli, and Luka Čehovin. 2016. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 11 (Nov. 2016), 2137–2155. <https://doi.org/10.1109/TPAMI.2016.2516982>
- [37] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naive Bayes text classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 1 (2011), 1–31.
- [38] Junseok Kwon and Kyoung Mu Lee. 2009. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive Basin Hopping Monte Carlo sampling. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, 1208–1215. <https://doi.org/10.1109/CVPR.2009.5206502>
- [39] Junseok Kwon and Kyoung Mu Lee. 2010. Visual tracking decomposition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, San Francisco, CA, 1269–1276. <https://doi.org/10.1109/CVPR.2010.5539821> ISSN: 1063-6919.
- [40] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports* 7, 1 (Dec. 2017), 17816. <https://doi.org/10.1038/s41598-017-17876-z>
- [41] Stephan J. Lemmer and Jason J. Corso. 2020. DAER to Reject Seeds with Dual-loss Additional Error Regression. arXiv:2009.07414 [cs.CV]
- [42] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. 2019. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 4277–4286. <https://doi.org/10.1109/CVPR.2019.00441>
- [43] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. 2018. High Performance Visual Tracking with Siamese Region Proposal Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, 8971–8980. <https://doi.org/10.1109/CVPR.2018.00935>
- [44] Hanxi Li, Yi Li, and Fatih Porikli. 2016. DeepTrack: Learning Discriminative Feature Representations Online for Robust Visual Tracking. *IEEE Transactions on Image Processing* 25, 4 (April 2016), 1834–1848. <https://doi.org/10.1109/TIP.2015.2510583> Conference Name: IEEE Transactions on Image Processing.
- [45] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the 2014 European Conference on Computer Vision*. Springer International Publishing, Zurich, Switzerland, 740–755.
- [46] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile, 3730–3738.
- [47] Alireza Mehrtash, William M Wells, Clare M Tempny, Purang Abolmaesumi, and Tina Kapur. 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging* 39, 12 (2020), 3868–3878.
- [48] Kourosh Meshgi and Shin Ishii. 2015. Expanding histogram of colors with gridding to improve tracking accuracy. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, Tokyo, Japan, 475–479. <https://doi.org/10.1109/MVA.2015.7153234>
- [49] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. 2018. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11205. Springer International Publishing, Cham, 310–327. https://doi.org/10.1007/978-3-030-01246-5_19
- [50] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis* 59 (2020), 101557.
- [51] Hyeonseob Nam and Bohyung Han. 2016. Learning Multi-domain Convolutional Neural Networks for Visual Tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 4293–4302. <https://doi.org/10.1109/CVPR.2016.465>
- [52] An T Nguyen, Aditya Kharosekar, Saumya Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM Press, Berlin, Germany, 189–199.
- [53] 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)*. IEEE, Las Vegas, NV, USA, 854–863. <https://doi.org/10.1109/CVPR.2016.99>
- [54] 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)*. IEEE, Venice, 4940–4949. <https://doi.org/10.1109/ICCV.2017.528>
- [55] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, 724–732. <https://doi.org/10.1109/CVPR.2016.85>
- [56] P.J. Phillips, Hyeonjoon Moon, S.A. Rizvi, and P.J. Rauss. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 10 (Oct. 2000), 1090–1104. <https://doi.org/10.1109/34.879790> Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [57] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct Uncertainty Prediction for Medical Second Opinions. In *Proceedings of the 2019 International Conference on Machine Learning*. ACM Press, Long Beach, CA, 10. arXiv:1807.01771 <http://arxiv.org/abs/1807.01771>
- [58] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. <https://doi.org/10.18653/v1/P18-2124>
- [59] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. 2017. YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, 7464–7473. <https://doi.org/10.1109/CVPR.2017.789>
- [60] N Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. 2018. CarFusion: Combining Point Tracking and Part Detection for Dynamic 3D Reconstruction of Vehicles. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 1906–1915. <https://doi.org/10.1109/CVPR.2018.00204>
- [61] David A. Ross, Jongwoo Lim, Rwei-Sung Lin, and Ming-Hsuan Yang. 2008. Incremental Learning for Robust Visual Tracking. *International Journal of Computer Vision* 77, 1-3 (May 2008), 125–141. <https://doi.org/10.1007/s11263-007-0075-7>
- [62] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (Dec. 2015), 211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- [63] 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)*. IEEE, Boston, MA, USA, 2121–2131. <https://doi.org/10.1109/CVPR.2015.7298824>
- [64] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision* 77, 1-3 (May 2008), 157–173. <https://doi.org/10.1007/s11263-007-0090-8>
- [65] Juan C. SanMiguel, Andrea Cavallaro, and José M. Martínez. 2012. Adaptive Online Performance Evaluation of Video Trackers. *IEEE Transactions on Image Processing* 21, 5 (May 2012), 2812–2823. <https://doi.org/10.1109/TIP.2011.2182520> Conference Name: IEEE Transactions on Image Processing.
- [66] Jakob Santner, Christian Leistner, Amir Saffari, Thomas Pock, and Horst Bischof. 2010. PROST: Parallel robust online simple tracking. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, San Francisco, CA, 723–730. <https://doi.org/10.1109/CVPR.2010.5540145>
- [67] Carla M. Santos-Pereira and Ana M. Pires. 2005. On optimal reject rules and ROC curves. *Pattern Recognition Letters* 26, 7 (May 2005), 943–952. <https://doi.org/10.1016/j.patrec.2004.09.042>
- [68] Jianbo Shi and Tomasi. 1994. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, WA, 593–600. <https://doi.org/10.1109/CVPR.1994.323794> ISSN: 1063-6919.
- [69] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *23rd International Conference on Intelligent User Interfaces*. ACM Press, Tokyo, Japan, 293–304.

- [70] Jean Y Song, John Joon Young Chung, David F Fouhey, and Walter S Lasecki. 2020. C-Reference: Improving 2D to 3D Object Pose Estimation Accuracy via Crowdsourced Joint Object Estimation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.
- [71] Jean Y. Song, Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, and Walter S. Lasecki. 2018. Two Tools are Better Than One: Tool Diversity as a Means of Improving Aggregate Crowd Performance. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval - IUI 18*. ACM Press, Tokyo, Japan, 559–570. <https://doi.org/10.1145/3172944.3172948>
- [72] Jean Y. Song, Stephan J. Lemmer, Michael Xieyang Liu, Shiyun Yan, Juho Kim, Jason J. Corso, and Walter S. Lasecki. 2019. Popup: reconstructing 3D video using particle filtering to aggregate crowd responses. In *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19*. ACM Press, Marina del Rey, California, 558–569. <https://doi.org/10.1145/3301275.3302305>
- [73] Ryan Szeto and Jason J. Corso. 2017. Click Here: Human-Localized Keypoints as Guidance for Viewpoint Estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, 1604–1613. <https://doi.org/10.1109/ICCV.2017.177>
- [74] Ryutaro Tanno, Daniel Worrall, Enrico Kaden, Aurobrata Ghosh, Francesco Grussu, Alberto Bizzi, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. 2019. Uncertainty quantification in deep learning for safer neuroimage enhancement. , 37 pages. arXiv:1907.13418 [eess.IV]
- [75] Francesco Tortorella. 2000. An Optimal Reject Rule for Binary Classifiers. In *Advances in Pattern Recognition*, Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Francesc J. Ferri, José M. Iñesta, Adnan Amin, and Pavel Pudil (Eds.). Vol. 1876. Springer Berlin Heidelberg, Berlin, Heidelberg, 611–620. https://doi.org/10.1007/3-540-44522-6_63
- [76] K. R. Varshney. 2011. A risk bound for ensemble classification with a reject option. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, Nice, France, 769–772. <https://doi.org/10.1109/SSP.2011.5967817>
- [77] Sirion Vittayakorn and James Hays. 2011. Quality Assessment for Crowdsourced Object Annotations. In *Proceedings of the British Machine Vision Conference 2011*. British Machine Vision Association, Dundee, 109.1–109.11. <https://doi.org/10.5244/C.25.109>
- [78] Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. 2019. SPM-Tracker: Series-Parallel Matching for Real-Time Visual Object Tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 3638–3647. <https://doi.org/10.1109/CVPR.2019.00376>
- [79] Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. 2015. Visual Tracking with Fully Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile, 3119–3127. <https://doi.org/10.1109/ICCV.2015.357>
- [80] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. 2019. Fast Online Object Tracking and Segmentation: A Unifying Approach. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, CA, 1328–1338.
- [81] Hao Wu, Aswin C. Sankaranarayanan, and Rama Chellappa. 2010. Online Empirical Evaluation of Tracking Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 8 (Aug. 2010), 1443–1458. <https://doi.org/10.1109/TPAMI.2009.135> Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [82] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2013. Online Object Tracking: A Benchmark. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Portland, OR, USA, 2411–2418. <https://doi.org/10.1109/CVPR.2013.312>
- [83] Alper Yilmaz, Omar Javed, and Mubarak Shah. 2006. Object tracking: A survey. *Acm computing surveys (CSUR)* 38, 4 (2006), 13–es.
- [84] Luka Čehovin Zajc, Matej Kristan, and Aleš Leonardis. 2014. Is my new tracker really better than yours?. In *WACV 2014: IEEE Winter Conference on Applications of Computer Vision*. IEEE, Steamboat Springs, CO, 540–547. <http://prints.vicos.si/publications/302>
- [85] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. 2018. Distractor-aware Siamese Networks for Visual Object Tracking. In *Proceedings of the 2018 European Conference on Computer Vision*. IEEE, Munich, Germany, 101–117.