DISCount: Counting in Large Image Collections with Detector-Based Importance Sampling

Gustavo Perez, Subhransu Maji[*], Daniel Sheldon[*]

University of Massachusetts Amherst
{gperezsarabi,smaji,sheldon}@cs.umass.edu

Abstract

Many applications use computer vision to detect and count objects in massive image collections. However, it may not be possible to train accurate enough counting models when the task is very difficult or requires a fast response time. For example, during disaster response, aid organizations aim to quickly count damaged buildings in satellite images to plan relief missions, but pre-trained building and damage detectors often perform poorly due to domain shifts. In such cases, there is a need for human-in-the-loop approaches that can accurately count with minimal human effort. We propose DISCount– a detector-based importance sampling framework for counting in large image collections. DISCount uses an imperfect detector and human screening to estimate low-variance unbiased counts. We propose techniques for counting over multiple spatial or temporal regions using a small amount of screening and estimate confidence intervals. This enables end-users to stop screening when estimates are sufficiently accurate, which is often the goal in real-world applications. We demonstrate our method with two applications: counting birds in radar imagery to understand responses to climate change, and counting damaged buildings in satellite imagery for damage assessment in regions struck by a natural disaster. On the technical side we develop variance reduction techniques based on control variates and prove the (conditional) unbiasedness of the estimators. DISCount leads to a 9-12× reduction in the labeling costs to obtain the same error rates compared to naive screening for tasks we consider, and surpasses alternative covariate-based screening approaches.

1 Introduction

Many applications of AI—especially to science, society, and the environment—use computer vision to detect and count objects in massive image collections. Examples include wildlife population monitoring (Wu et al. 2023) and the mapping of agriculture (Singh et al. 2022; Turkoglu et al. 2021) or poverty (Ayush et al. 2021; Yeh et al. 2020) from satellite images. We are interested in two particular applications: (1) counting bird roosts in radar to understand population responses to climate change and aid conservation, and (2) counting damaged buildings in satellite images to inform disaster response. These image collections are too large for humans to perform the counting tasks in the available time. Therefore, a common strategy is to train a computer vision detection model using labeled data and run it exhaustively on the images.

The task is interesting because the goal is not to generalize, but to achieve the scientific counting goal with sufficient accuracy for a *fixed* image collection. The best use of human effort is unclear: it could be used for model development, labeling training data, or even directly solving the counting task! A particular challenge occurs when the detection task is very difficult, so the accuracy of counts made on the entire collection is questionable even with huge investments in training data and model development. Some works resort to human screening of the detector outputs (Norouzzadeh et al. 2018; Nurkarim and Wijayanto 2023; Perez et al. 2022), which while faster than manual counting, is still very labor intensive.

These considerations motivate *statistical* approaches to counting. Instead of screening detector outputs for all images, a human can "spot-check" images to estimate accuracy, and, more importantly, use statistical techniques to obtain unbiased estimates of counts across unscreened images. In a related context, Meng et al. (2022) proposed IS-count, which employs importance sampling to estimate counts across a collection when (satellite) images are expensive to obtain by using spatial covariates.

Our work draws inspiration from two distinct applications: an environmental application to count bird roosts and a societal one assess building damage. In both cases, the cost model differs from IS-count: images are readily available, and the goal is to minimize the human effort dedicated to screening. The first application is counting bird roosts across space and time in US weather radar data (§ 4.1). These counts reveal species' response to climate change (Deng et al. 2023) and provide urgently needed information to conserve bird populations. However, the detection problem is difficult, so fully automated methods are not accurate enough even after substantial investment in training data collection and model development. The second application is building damage assessment for disaster response (§ 4.2). Aid organizations assess building damage using satellite images (Deng and Wang 2022) to plan humanitarian response after a natural disaster. However, pre-trained detection models are often not accurate enough when applied to a new dis-

[*] Equal advising contribution.

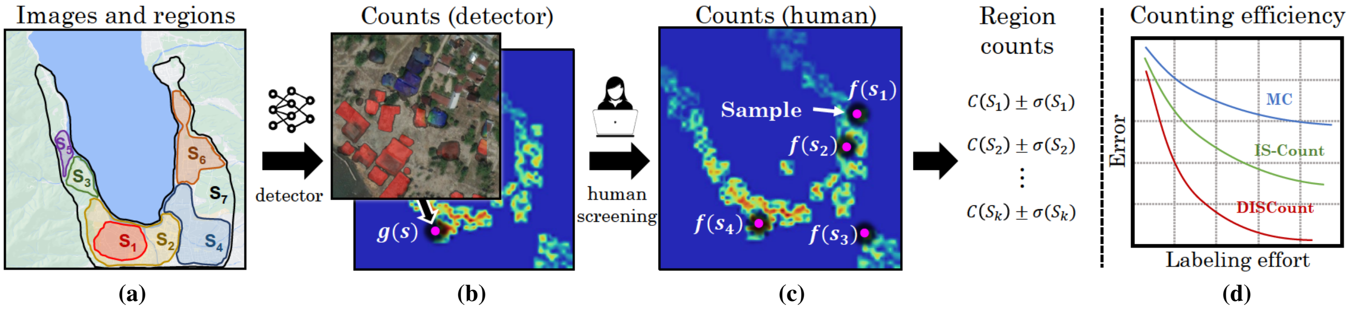

Figure 1: k-DISCount uses detector-based importance sampling to screen counts and solve multiple counting problems. **(a)** Geographical regions $S_1, S_2, \ldots, S_7$, where we want to estimate counts of damaged buildings. **(b)** Outputs of a damaged building detector on satellite imagery, which can be used to estimate counts $g(s)$ for each tile (shows as dots). **(c)** Tiles selected for human screening to obtain true counts $f(s)$, from which counts (C) and confidence intervals (σ) are estimated for all regions. **(d)** DISCount outperforms naive (MC) and covariate-based sampling (IS-Count (Meng et al. 2022)) in our experiments.

aster due to domain shift, and there is very limited time and expertise for model development, but volunteers are available to view satellite imagery and help assess damage. Both applications are well suited to human-in-the-loop counting.

We contribute counting methods for large image collections that build on IS-count in several ways. First, since we work in a cost model where images are freely available, it is possible to train a detector and run it on all images, even though it may not be reliable enough for the final counting task. We propose to use the imperfect detector to construct a proposal distribution for human-in-the-loop count estimation, as shown in Fig. 1. Second, we consider solving multiple counting problems—for example, over disjoint or overlapping spatial or temporal regions—simultaneously, which is very common in practice. We contribute a novel sampling approach to obtain simultaneous estimates, prove their (conditional) unbiasedness, and show that the approach allocates samples to regions in a way that approximates the optimal allocation for minimizing variance. Third, we design confidence intervals, which are important practically to know how much human effort is needed. Fourth, we use variance reduction techniques based on control variates.

Our method produces unbiased estimates and confidence intervals that obtain reduced error compared to covariate-based methods. In addition, labeling effort is further reduced with DISCount as one has to verify detector predictions instead of producing annotations from scratch. On our tasks, DISCount leads to a 9-12× reduction in the labeling costs over naive screening and 6-8× reduction over IS-Count. Finally, we show that solving multiple counting problems jointly can be done more efficiently than solving them separately, demonstrating a more efficient use of samples.

2 Related Work

Computer vision techniques have been deployed for counting in numerous applications where exhaustive human-labeling is expensive due to the sheer volume of imagery involved. This includes areas such as detecting animals in camera trap imagery (Norouzzadeh et al. 2018; Tuia et al. 2022), counting buildings, cars, and other structures in satellite images (Nurkarim and Wijayanto 2023; Cavender-Bares et al. 2022; Burke et al. 2021; Leitloff, Hinz, and Stilla 2010), species monitoring in citizen science platforms (Tuia et al. 2022; Van Horn et al. 2018), monitoring traffic in videos (Won 2020; Coifman et al. 1998), as well as various medicine, science and engineering applications. For many applications the cost associated with training an *accurate* model is considerably less than that of meticulously labeling the entire dataset. Even with a less accurate model, human-in-the-loop recognition strategies have been proposed to reduce annotation costs by integrating human validation with noisy predictions (Branson et al. 2010; Wah et al. 2014).

Our approach is related to work in active learning (Settles 2009) and semi-supervised learning (Chapelle, Scholkopf, and Zien 2009), where the goal is to reduce human labeling effort to learn models that generalize on i.i.d. held out data. While these approaches reduce the cost of labels on training data, they often rely on large labeled test sets to estimate the performance of the model, which can be impractical. Active testing (Nguyen, Ramanan, and Fowlkes 2018; Kossen et al. 2021) aims to reduce the cost of model evaluation by providing a statistical estimate of the performance using a small number of labeled examples. Unlike traditional learning where the goal is performance on held out data, the goal of active testing is to estimate performance on a *fixed* dataset. Similarly, our goal is to estimate the counts on a fixed dataset, but different from active testing we are interested in estimates of the true counts and not the model's performance. In particular, we want unbiased estimates of counts even when the detector is unreliable. Importantly, since generalization is not the goal, overfitting to the dataset statistics may lead to more accurate estimates.

Statistical estimation has been widely used to conduct surveys (e.g., estimating population demographics, polling, etc.) (Cochran 1977). In IS-Count (Meng et al. 2022), the authors propose an importance sampling approach to estimate counts in large image collections using humans-in-the-loop. They showed that one can count the number of buildings at the continental scale by sampling a small number of regions based on covariates such as population density and annotating those regions, thereby reducing the cost of obtaining high-resolution satellite imagery and human labels.

However, for many applications the dataset is readily available, and running the detector is cost effective, but human screening is expensive. To address this, we propose using the detector to guide the screening process and demonstrate that this significantly reduces error rates in count estimation given a fixed amount of human effort. Furthermore, for some applications, screening the outputs of a detector can be significantly faster than to annotate from scratch, leading to additional savings.

An interesting question is what is the best way to utilize human screening effort to count on a dataset. For example, labels might be used to improve the detector, measure performance on the deployed dataset, or, as is the case in our work, to derive a statistical estimate of the counts. Our work is motivated by problems where improving the detector might require significant effort, but counts from the detector are correlated with true counts and can be used as a proposal distribution for sampling.

3 DISCount: Detector-based IS-Count

Consider a counting problem in a discrete domain Ω (usually spatiotemporal) with elements $s \in \Omega$ that represent a single unit such as an image, grid cell, or day of year. For each s there is a ground truth "count" $f(s) \geq 0$, which can be any non-negative measurement, such as the number or total size of all objects in an image. A human can label the underlying images for any s to obtain $f(s)$.

Define $F(S) = \sum_{s \in S} f(s)$ to be the cumulative count for a region S. We wish to estimate the total counts $F(S_1), \ldots, F(S_k)$ for k different subsets $S_1, \ldots, S_k \subseteq \Omega$, or *regions*, while using human effort as efficiently as possible. The regions represent different geographic divisions or time ranges and may overlap — for example, in the roost detection problem we want to estimate *cumulative* counts of birds for each day of the year, while disaster-relief planners want to estimate building damage across different geographical units such as towns, counties, and states. Assume without loss of generality that $\bigcup_{i=1}^{k} S_i = \Omega$, otherwise the domain can be restricted so this is true.

We will next present our methods; derivations and proofs of all results are found in the appendix.

3.1 Single-Region Estimators

Consider first the problem of estimating the total count $F(S)$ for a single region S. Meng et al. (2022) studied this problem in the context of satellite imagery, with the goal of minimizing the cost of purchasing satellite images to obtain an accurate estimate.

Simple Monte Carlo (Meng et al. 2022) This is a baseline based on simple Monte Carlo sampling. Write $F(S) = \sum_{s \in S} f(s) = |S| \cdot \mathbb{E}_{s \sim \text{Unif}(S)}[f(s)]$. Then the following estimator, which draws n random samples uniformly in S to estimate the total, is unbiased:

$$\hat{F}_{\text{MC}}(S) = |S| \cdot \frac{1}{n} \sum_{i=1}^{n} f(s_i), \quad s_i \sim \text{Unif}(S).$$

IS-Count (Meng et al. 2022) Meng et al. then proposed an estimator based on importance sampling (Owen 2013). Instead of sampling uniformly, the method samples from a proposal distribution q that is cheap to compute for all $s \in S$. For example, to count buildings in US satellite imagery, the proposal distribution could use maps of artificial light intensity, which are freely available. The importance sampling estimator is:

$$\hat{F}_{\text{IS}}(S) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(s_i)}{q(s_i)}, \quad s_i \sim q.$$

DISCount IS-count assumes *images* are costly to obtain, which motivates using external covariates for the proposal distribution. However, in many scientific tasks, the images are readily available, and the key cost is that of human supervision. In this case it is possible to train a detection model and run it on all images to produce an approximate count $g(s)$ for each s. Define $G(S) = \sum_{s \in S} g(s)$ to be the approximate detector-based count for region S. We propose the *detector-based IS-count* ("DISCount") estimator, which uses the proposal distribution proportional to g on region S, i.e., with density $\bar{g}_S(s) = g(s)\,\mathbb{I}[s \in S]/G(S)$. The importance-sampling estimator then specializes to:

$$\hat{F}_{\text{DIS}}(S) = G(S) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{f(s_i)}{g(s_i)}, \quad s_i \sim \bar{g}_S. \tag{1}$$

To interpret DISCount, let $w_i = f(s_i)/g(s_i)$ be the ratio of the true count to the detector-based count for the ith sample s_i or (importance) *weight*. DISCount reweights the detector-based total count $G(S)$ by the average weight $\bar{w} = \frac{1}{n} \sum_{i=1}^{n} w_i$, which can be viewed as a correction factor based on the tendency to over- or under-count, on average, across all of S.

DISCount is unbiased as long as $\bar{g}(s) > 0$ for all $s \in S$ such that $f(s) > 0$. Henceforth, we assume detector counts are pre-processed if needed so that $g(s) > 0$ for all relevant units, for example, by adding a small amount to each count.

3.2 k-DISCount

We now return to the multiple region counting problem. A naive approach would be to run DISCount separately for each region. However, this is suboptimal. First, it allocates samples equally to each region, regardless of their size or predicted count. Intuitively, we want to allocate more effort to regions with higher predicted counts. Second, if regions overlap it is wasteful to repeatedly draw samples from each one to solve the estimation problems separately.

k-DISCount We propose estimators based on n samples drawn from all of Ω with probability proportional to g. Then, we can estimate $F(S)$ for any region using only the samples from S. Specifically, the k-DISCount estimator is

$$\hat{F}_{k\text{DIS}}(S) = \begin{cases} G(S) \cdot \bar{w}(S) & n(S) > 0 \\ 0 & n(S) = 0 \end{cases}, \quad s_i \sim \bar{g}_\Omega, \tag{2}$$

where $n(S) = |\{i : s_i \in S\}|$ is the number of samples in region S and $\bar{w}(S) = \frac{1}{n(S)} \sum_{i: s_i \in S} w_i$ is the average importance weight for region S.

Claim 1. *The k-DISCOUNT estimator $\hat{F}_{k\text{DIS}}(S)$ is conditionally unbiased given at least one sample in region S. That is, $\mathbb{E}[\hat{F}_{k\text{DIS}}(S) \mid n(S) > 0] = F(S)$.*

The unconditional bias can also be analyzed (see Appendix). Overall, bias has negligible practical impact. It occurs only when the sample size $n(S)$ is zero, which is an event that is both observable and has probability $(1-p(S))^n$ that decays exponentially in n, where $p(S) = G(S)/G(\Omega)$.

In terms of variance, k-DISCOUNT behaves similarly to DISCOUNT run on each region S with sample size equal to $\mathbb{E}[n(S)] = np(S)$. To first order, both approaches have variance $\frac{G(S)^2 \cdot \sigma^2(S)}{np(S)}$ where $\sigma^2(S)$ is the importance-weight variance. In the case of *disjoint* regions, running DISCOUNT on each region is the same as *stratified importance sampling* across the regions, and the allocation of $np(S)$ samples to region S is optimal in the following sense:

Claim 2. *Suppose $S_1, \ldots, S_k$ partition Ω and the importance weight variance $\sigma^2(S_i) = \sigma^2$ is constant across regions. Assume* DISCOUNT *is run on each region S_i with n_i samples. Given a total budget of n samples, the sample sizes that minimize $\sum_{i=1}^{k} \mathrm{Var}(\hat{F}_{\text{DIS}}(S_i))$ are given by $n_i = np(S_i) = nG(S_i)/G(\Omega)$.*

The analysis uses reasoning similar to the *Neyman allocation* for stratified sampling (Cochran 1977), and shows that k-DISCOUNT approximates the optimal allocation of samples to (disjoint) regions under the stated assumptions. One key difference is that k-DISCOUNT draws samples from all of Ω and then assigns them to regions, which is called "post-stratification" in the sampling literature (Cochran 1977). An exact variance analysis in the Appendix reveals that, if the expected sample size $np(S)$ for a region is very small, k-DISCOUNT may have up to 30% "excess" variance compared to stratification due to the random sample size, but the excess variance disappears quickly and both approaches have the same asymptotic variance. A second key difference to stratification is that regions can overlap; k-DISCOUNT's approach of sampling from all of Ω and then assigning samples to regions extends cleanly to this setting.

3.3 Control Variates

Control variates are functions $h(s)$ whose integrals $H(S) = \sum_{s \in S} h(s)$ are known and can be combined with importance sampling using the following estimator:

$$\hat{F}_{k\text{DIS}cv}(S) = \begin{cases} G(S) \cdot \bar{w}_h(S) + H(S) & n(S) > 0 \\ 0 & n(S) = 0 \end{cases}, \quad s_i \sim \bar{g}_\Omega, \tag{3}$$

where $\bar{w}_h(S) = \frac{1}{n(S)} \sum_{i: s_i \in S} w_{h,i}$ and $w_{h,i} = (f(s_i) - h(s_i))/g(s_i)$. It is clear that $\hat{F}_{k\text{DIS}cv}(S)$ has the same expectation as $\hat{F}_{k\text{DIS}(S)}$, but $\hat{F}_{k\text{DIS}cv}(S)$ might have a lower variance if f and h are sufficiently correlated (Owen 2013). For bird counting, estimated counts from previous years could be used as control variates as migration is periodic to improve count estimates (see § 4 for details).

3.4 Confidence Intervals

Confidence intervals for k-DISCOUNT can be constructed in a way similar to standard importance sampling. For a region S, estimate the importance weight variance $\sigma^2(S)$ as:

$$\hat{\sigma}^2(S) = \frac{1}{n(S)} \sum_{i: s_i \in S} \left(\frac{f(s_i)}{g(s_i)} - \frac{\hat{F}_{k\text{DIS}}(S)}{G(S)} \right)^2. \tag{4}$$

An approximate $1 - \alpha$ confidence interval is then given by $\hat{F}_{k\text{DIS}}(S) \pm z_{\alpha/2} \cdot G(S) \cdot \hat{\sigma}(S)/\sqrt{n(S)}$, where z_γ is the $1 - \gamma$ quantile of the standard normal distribution, e.g., $z_{0.025} = 1.96$ for a 95% confidence interval. The theoretical justification is subtle due to scaling by the *random* sample size $n(S)$. It is based on the following asymptotic result, proved in the Appendix.

Claim 3. *The k-DISCOUNT estimator with scaling factor $G(S)\hat{\sigma}(S)/\sqrt{n(S)}$ is asymptotically normal, that is, the distribution of $\frac{\hat{F}_{k\text{DIS}}(S) - F(S)}{G(S) \cdot \hat{\sigma}(S)/\sqrt{n(S)}}$ converges to $\mathcal{N}(0, 1)$ as $n \to \infty$.*

In preliminary experiments we observed that for small expected sample sizes the importance weight variance $\sigma^2(S)$ can be underestimated leading to intervals that are too small — as an alternative, we propose a practical heuristic for smaller sample sizes where $\hat{\sigma}^2(\Omega)$ is used instead of $\hat{\sigma}^2(S)$; that is, *all* samples are used to estimate variability of importance weights for each region S.

4 Experimental Setup

In this section we describe the counting tasks and detection models (§ 4.1–4.2) and the evaluation metrics (§ 4.3) we will use to evaluate different counting methods. We focus on two applications: counting roosting birds in weather radar images and counting damaged buildings in satellite images of a region struck by a natural disaster.

4.1 Roosting Birds from Weather Radar

Many species of birds and bats congregate in large numbers at nighttime or daytime roosting locations. Their departures from these "roosts" are often visible in weather radar, from which it's possible to estimate their numbers (Winkler 2006; Buler et al. 2012; Horn and Kunz 2008). The US "NEXRAD" weather radar network (Oceanic and NOAA) has collected data for 30 years from 143+ stations and provides an unprecedented opportunity to study long-term and wide-scale biological phenomenon such as roosts (Rosenberg et al. 2019; Sánchez-Bayo and Wyckhuys 2019). However, the sheer volume of radar scans (>250M) prevents manual analysis and motivates computer vision approaches (Chilson et al. 2018; Lin et al. 2019; Cheng et al. 2020; Perez et al. 2022). Unfortunately, the best computer vision models (Perez et al. 2022; Cheng et al. 2020) for detecting roosts have average precision only around 50% and are not accurate enough for fully automated scientific analysis, despite using state-of-the-art methods such as Faster R-CNNs (Ren et al. 2015) and training on thousands of human annotations — the complexity of the task suggests substantial labeling and model development efforts would be needed to improve accuracy, and may be impractical.

Previous work (Belotti et al. 2023; Deng et al. 2023) used a roost detector combined with manual screening of the detections to analyze more than 600,000 radar scans spanning a dozen stations in the Great Lakes region of the US to reveal patterns of bird migration over two decades. The vetting of nearly 64,000 detections was orders of magnitude faster than manual labeling, yet still required a substantial 184 hours of manual effort. Scaling to the entire US network would require at least an order of magnitude more effort, thus motivating a statistical approach.

We use the exhaustively screened detections from the Great Lakes analysis in (Belotti et al. 2023; Deng et al. 2023) to systematically analyze the efficiency of sampling based counting. The data is organized into domains $\Omega^{\texttt{sta,yr}}$ corresponding to 12 stations and 20 years (see Fig. 7 in Appendix B). Thus the domains are disjoint and treated separately. Counts are collected for each day s by running the detector using all radar scans for that day to detect and track roost signatures and then mapping detections to bird counts using the measured radar "reflectivity" within the tracks. For the approximate count $g(s)$ we use the automatically detected tracks, while for the true count $f(s)$ we use the manually screened and corrected tracks. For a single domain, i.e., each station-year, we divide a complete roosting season into temporal regions in three different scenarios: (1) estimating bird counts up to each day in the roosting season (i.e., regions are nested prefixes of days in the entire season), (2) the end of each quarter of (i.e., regions are nested prefixes of quarters in the entire season), and (3) estimating each quarter's count (each region is one quarter). We measure error using the fully-screened data and average errors across all domains and regions. Fig. 2 shows the counts and confidence intervals estimated using k-DISCOUNT for the first scenario on four station-years.

4.2 Damaged Buildings from Satellite Images

Building damage assessment from satellite images (Kim and Yoon 2018; Deng and Wang 2022) is often used to plan humanitarian response after a natural disaster strikes. However, the performance of computer vision models degrades when applied to new regions and disaster types. Our approach can be used to quickly vet the data produced by the detector to correctly estimate counts in these scenarios.

We use the building damage detection model by (DIUx-xView 2020), the winner of the xView2 challenge (The Defense Innovation Unit 2020). The model is based on U-Net (Ronneberger, Fischer, and Brox 2015) to detect buildings in the pre-disaster image, followed by a "siamese network" that incorporates at pre- and post-disaster images to estimate damage. The model is trained on the xBD dataset (Gupta et al. 2019) that contains building and damage annotations spanning multiple geographical regions and disaster types (e.g., earthquake, hurricane, tsunami, etc.). While the dataset contains four levels of damage (i.e., 0: no-damage, 1: minor-damage, 2: major-damage, and 3: destroyed), in this work we combine all damage levels (i.e., classes 1-3) into a single "damage" class.

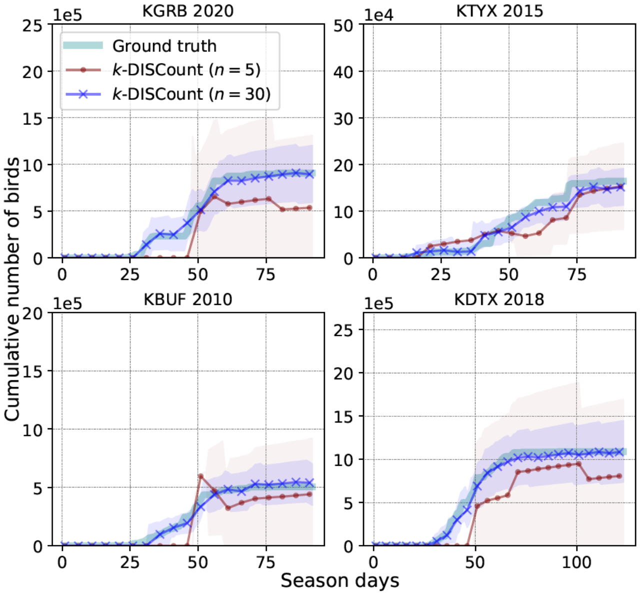

Figure 2: Count estimates with confidence intervals for two station years (i.e., KGRB 2020 and KBUF 2010) using different numbers of samples.

We consider the Palu Tsunami from 2018; the data consists of 113 high-resolution satellite images labeled with 31,394 buildings and their damage levels. We run the model on each tile s to estimate the number of damaged buildings $g(s)$, while the ground-truth number of damaged buildings is used as $f(s)$. Our goal is to estimate the cumulative damaged building count in sub-regions expanding from the area with the most damaged buildings as shown in Fig. 9 in the Appendix C. To define the sub-regions, we sort all m images by their distance from the epicenter (defined as the image tile with the most damaged buildings) and then divide into chunks or "annuli" $A_1, \ldots, A_7$ of size $m/7$. The task is to estimate the cumulative counts $S_j = \bigcup_{i=1}^{j} A_i$ of the first j chunks for j from 1 to 7.

4.3 Evaluation

We measure the fractional error between the *true* and the estimated counts averaged over all regions in a domain $S_1, \ldots, S_k \subseteq \Omega$ as:

$$\texttt{Error}(\Omega) = \frac{1}{k}\sum_{i=1}^{k}\frac{|F(S_i) - \hat{F}(S_i)|}{F(\Omega)}.$$

For the bird counting task, for any given definition of regions within one station-year Ω (i.e., cumulative days or quarters defined in § 4.1) we report the error averaged across all station-years corresponding to 12 stations and $\approx$ 20 years. For the damaged building counting problem there is only a single domain corresponding to the Palu Tsunami region. In addition, we calculate the average confidence interval width normalized by $F(\Omega)$. We run 1000 trials and plot average metrics $\pm 1.96 \times$ std. error over the trials. We also evaluate confidence interval coverage, which is the fraction of confidence intervals that contain the true count over all domains, regions, and trials.

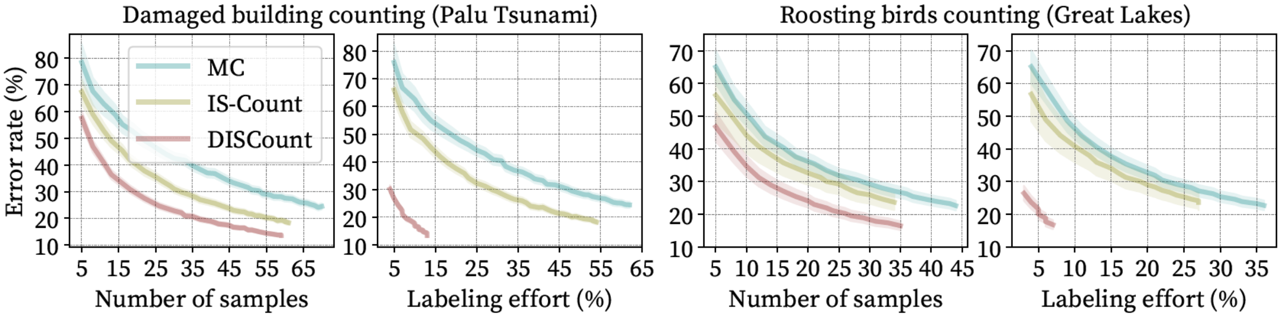

Figure 3: Detector-based sampling. Estimation error of damaged building counts in the Palu Tsunami region from the xBD dataset (left) and counting roosting birds from the Great Lakes radar stations in the US from NEXRAD data (right). We get lower error with DISCOUNT compared to IS-Count and simple Monte Carlo sampling (MC). The labeling effort is further reduced with DISCOUNT since the user is not required to label an image from scratch but only to verify outputs from the detector (See § 5 for details). The estimation errors are averaged over 1000 runs.

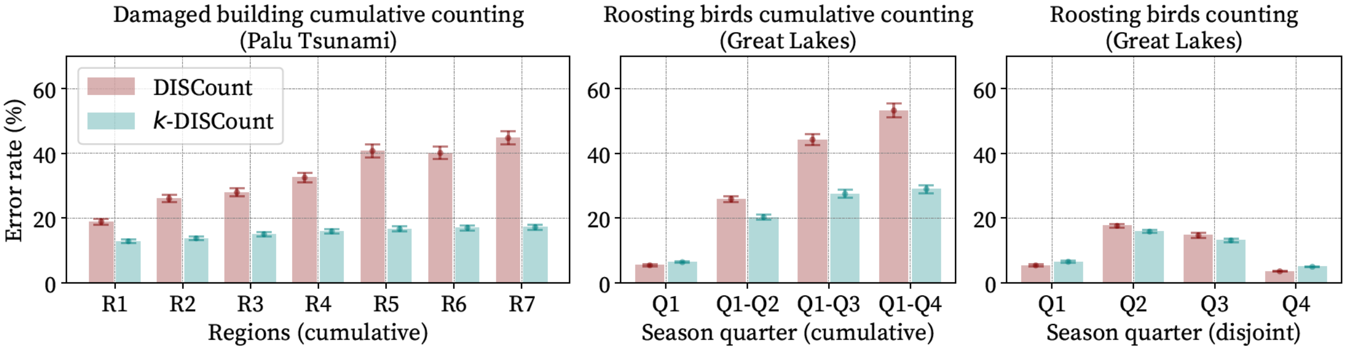

Figure 4: Solving multiple counting problems jointly. Estimation error of counting damaged buildings in the Palu Tsunami region from the xBD dataset (left) and counting roosting birds from the Great Lakes radar stations in the US from NEXRAD data (right). We compare solving the counting problems jointly (k-DISCOUNT) against solving the counting problems separately (DISCOUNT). We use 10 samples for both these tests. The estimation errors are averaged over 1000 runs.

5 Results

In this section, we present the results comparing detector-based to covariate-based sampling. Also, we show reductions in labeling effort and demonstrate the advantages of estimating multiple counts jointly. Finally, we show confidence intervals and control variates results.

Detector-based sampling reduces error We first compare DISCOUNT (detector-based sampling) to IS-Count and simple Monte Carlo sampling for estimating $F(\Omega)$, that is, the total counts of birds in a complete roosting season for a given station year, or damaged buildings in the entire disaster region. Fig. 3 shows the error rate as a function of number of labeled samples (i.e., the number of *distinct* s_i sampled, since each s is labeled at most once). In the buildings application, a sample refers to an image tile of size 1024 × 1024 pixels, while for the birds a sample refers to a single day.

Using the detector directly without *any* screening results in high error rates — roughly 136% and 149% for estimating the total count for the damaged buildings and bird counting tasks respectively. Meng et al. (2022) show the advantages of using importance sampling with screening to produce count estimates with base covariates as opposed to simple Monte Carlo sampling (MC vs. IS-Count). For the bird counting task, we construct a non-detector covariate g_{IS} by fitting a spline to $f(s)$ with 10% of the days from an arbitrarily selected station-year pair (station KBUF in 2001). For the damaged building counting task, the covariate g_{IS} is the true count of all buildings (independent of the damage) obtained using the labels provided with the xBD dataset.

IS-Count leads to significant savings over Monte Carlo sampling (MC), but DISCOUNT provides further improvements. In particular, to obtain an error rate of 20% DISCOUNT requires $\approx 1.6\times$ fewer samples than IS-Count and $\approx 3\times$ fewer samples than MC for both counting problems.

Screening leads to a further reduction in labeling effort DISCOUNT alleviates the need for users to annotate an image from scratch, such as identifying an object and drawing a bounding box around it. Instead, users only need to verify the detector's output, which tends to be a quicker process. In a study by Su, Deng, and Fei-Fei (2012) on the ImageNet dataset (Deng et al. 2009), the median time to draw a bounding-box was found to be 25.5 seconds, whereas verification took only 9.0 seconds (this matches the screening time of ≈10s per bounding-box in (Deng et al. 2023; Be-

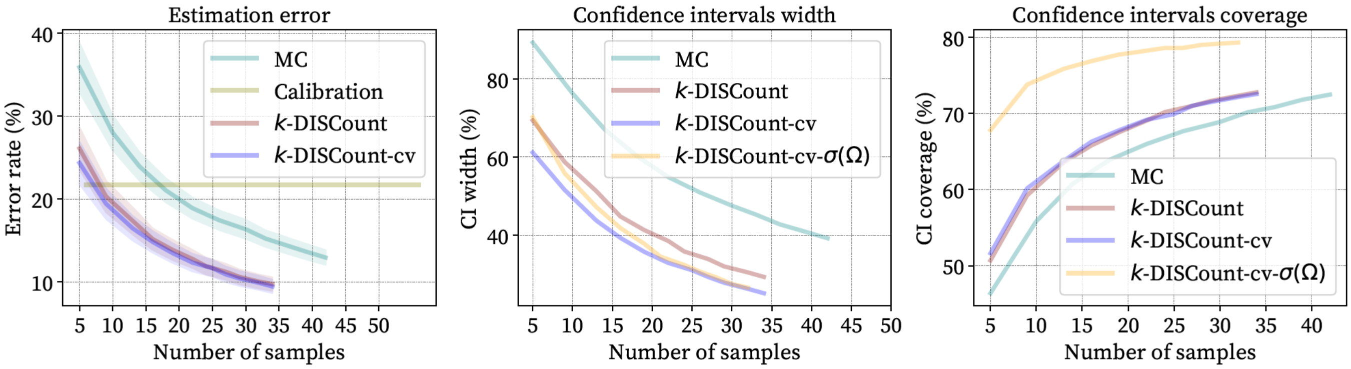

Figure 5: Control variates and confidence intervals on bird counting. We compare simple Monte Carlo (MC), calibration with isotonic regression, and variations of k-DISCount that include control variates (-cv) and improved variance estimates $(-\sigma(\Omega))$. (left) Error rates using k-DISCount are significantly smaller than MC and calibration. (middle) Confidence intervals' width. (right) Confidence intervals' coverage. The error and the confidence intervals' width are slightly reduced when control variates are used while maintaining the coverage. Furthermore, k-DISCount-cv-$\sigma(\Omega)$ improves the coverage. The results are averaged over all station-years and over 1,000 runs.

lotti et al. 2023)). The right side of Fig. 3 presents earlier plots with the x-axis scaled based on labeling effort, computed as $100 \cdot c \cdot n/|\Omega|$, where n denotes the number of screened samples and $c \in [0, 1]$ represents the fraction of time relative to labeling from scratch. For instance, the labeling effort is 100% when all elements must be labeled from scratch ($c = 1$ and $n = |\Omega|$). For DISCount, we estimate $c_{\text{DIS}} = 9.0/(25.5+9.0) = 0.26$, since annotating from scratch requires both drawing and verification, while screening requires only verification. To achieve the same 20% error rate, DISCount requires 6× less effort than IS-Count and 9× less effort than MC for the bird counting task, and 8× less effort than IS-Count and 12× less effort than MC for building counting.

Multiple counts can be estimated efficiently (k-DISCount) To solve multiple counting problems, we compared k-DISCount to using DISCount separately on each region. For bird counting, the task was to estimate four quarterly counts (cumulative or individual) as described in § 4.1. For k-DISCount, we sampled $n = 40$ days from the complete season to estimate the counts simultaneously. For DISCount, we solved each of the four problems separately using $n/4 = 10$ samples per region for the same total number of samples. For building damage counting, the task was to estimate seven cumulative counts as described in § 4.2. For k-DISCount, we used $n = 70$ images sampled from the entire domain, while for DISCount we used $n/7 = 10$ sampled images per region.

Fig. 4 shows that solving multiple counting problems jointly (k-DISCount) is better than solving them separately (DISCount). For the cumulative tasks, k-DISCount makes much more effective use of samples from overlapping regions. For single-quarter bird counts, k-DISCount has slightly higher error in Q1 and Q4 and lower errors in Q2 and Q3. This can be understood in terms of sample allocation: k-DISCount allocates in proportion to predicted counts, which provides more samples and better accuracy in Q2-Q3, when many more roosts appear, and approximates the optimal allocation of Claim 2. DISCount allocates samples equally, so has slightly lower error for the smaller Q1 and Q4 counts. In contrast, for building counting, k-DISCount has lower error even for the smallest region R1, since this has the most damaged buildings and thus gets more samples than DISCount. Fig. 5 (left) shows k-DISCount outperforms simple Monte Carlo (adapted to multiple regions similarly to k-DISCount) for estimating cumulative daily bird counts as in Fig. 2.

Confidence intervals We measure the width and coverage of the estimated confidence intervals (CIs) per number of samples for cumulative daily bird counting; see examples in Fig. 2. We compare the CIs of k-DISCount, k-DISCount-cv (control variates), k-DISCount-cv-$\sigma(\Omega)$ (using all samples to estimate variance), and simple Monte Carlo sampling in Fig. 5. When using control variates, the error rate and the CI width are slightly reduced while keeping the same coverage. CI coverage is lower than the nominal coverage (95%) for all methods, but increasing with sample size and substantially improved by k-DISCount-cv-$\sigma(\Omega)$, which achieves up to $\approx 80\%$ coverage. Importance weight distributions can be heavily right-skewed and the variance easily underestimated (Hesterberg 1996).

DISCount improves over a calibration baseline We implement a calibration baseline where the counts are estimated as $\hat{F}_{\text{CAL}}(S) = \sum_{s \in S} \hat{\phi}(g(s))$, where we learn an isotonic regression model $\hat{\phi}$ between the predicted and true counts trained for each station using 15 uniformly selected samples from one year from that station. Results are shown as the straight line in Fig. 5 (left). DISCount outperforms calibration with less than 10 samples per station suggesting the difficulties in generalization across years using a simple calibration approach.

Control variates (k-DISCount-cv) We perform experiments adding control variates to k-DISCount in the roosting birds counting problem. We use the calibrated detector counts $\hat{\phi}(g(s))$ defined above as the control variate for each station year. Fig. 5 shows that control variates reduce the confidence interval width (middle: k-DISCount vs. k-DISCount-cv) without hurting coverage (right). In addition, the error of the estimate is reduced slightly, as shown in Fig. 5 (left). Note that this is achieved with a marginal increase in the labeling effort.

6 Discussion and Conclusion

We contribute methods for counting in large image collections with a detection model. When the task is complex and the detector is imperfect, allocating human effort to estimate the scientific result directly might be more efficient than improving the detector. For instance, performance gains from adding more training data may be marginal for a mature model. Our proposed solution produces accurate and unbiased estimates with a significant reduction in labeling costs from naive and covariate-based screening approaches. We demonstrate this in two real-world open problems where data screening is still necessary despite large investments in model development. Our approach is limited by the availability of a good detector, and confidence interval coverage is slightly low; possible improvements are to use bootstrapping or corrections based on importance-sampling diagnostics (Hesterberg 1996).

Acknowledgements

We thank Wenlong Zhao for the deployment of the roost detector, Maria Belotti, Yuting Deng, and our Colorado State University AeroEco Lab collaborators for providing the screened data of the Great Lakes radar stations, and Yunfei Luo for facilitating the building detections on the Palu Tsunami region. This work was supported by the National Science Foundation award #2017756.

References

Ayush, K.; Uzkent, B.; Tanmay, K.; Burke, M.; Lobell, D.; and Ermon, S. 2021. Efficient Poverty Mapping from High Resolution Remote Sensing Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1): 12–20.

Belotti, M. C. T.; Deng, Y.; Zhao, W.; Simons, V. F.; Cheng, Z.; Perez, G.; Tielens, E.; Maji, S.; Sheldon, D.; Kelly, J. F.; et al. 2023. Long-term analysis of persistence and size of swallow and martin roosts in the US Great Lakes. *Remote Sensing in Ecology and Conservation*.

Branson, S.; Wah, C.; Schroff, F.; Babenko, B.; Welinder, P.; Perona, P.; and Belongie, S. 2010. Visual recognition with humans in the loop. In *European Conference on Computer Vision (ECCV)*.

Buler, J. J.; Randall, L. A.; Fleskes, J. P.; Barrow Jr, W. C.; Bogart, T.; and Kluver, D. 2012. Mapping wintering waterfowl distributions using weather surveillance radar. *PloS one*, 7(7): e41571.

Burke, M.; Driscoll, A.; Lobell, D. B.; and Ermon, S. 2021. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535): eabe8628.

Cavender-Bares, J.; Schneider, F. D.; Santos, M. J.; Armstrong, A.; Carnaval, A.; Dahlin, K. M.; Fatoyinbo, L.; Hurtt, G. C.; Schimel, D.; Townsend, P. A.; et al. 2022. Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nature Ecology & Evolution*, 6(5): 506–519.

Chapelle, O.; Scholkopf, B.; and Zien, A. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3): 542–542.

Cheng, Z.; Gabriel, S.; Bhambhani, P.; Sheldon, D.; Maji, S.; V, A.; and Winkler, D. 2020. Detecting and Tracking Communal Bird Roosts in Weather Radar Data. *Association for the Advancement of Artificial Intelligence (AAAI)*.

Chilson, C.; Avery, K.; McGovern, A.; Bridge, E.; Sheldon, D.; and Kelly, J. 2018. Automated Detection of Bird Roosts using NEXRAD Radar Data and Convolutional Neural Networks. *Remote Sensing in Ecology and Conservation*.

Cochran, W. G. 1977. *Sampling techniques*. John Wiley & Sons.

Coifman, B.; Beymer, D.; McLauchlan, P.; and Malik, J. 1998. A real-time computer vision system for vehicle tracking and traffic surveillance. *Transportation Research Part C: Emerging Technologies*, 6(4): 271–288.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Deng, L.; and Wang, Y. 2022. Post-disaster building damage assessment based on improved U-Net. *Scientific Reports*, 12(1): 2045–2322.

Deng, Y.; Belotti, M. C. T.; Zhao, W.; Cheng, Z.; Perez, G.; Tielens, E.; Simons, V. F.; Sheldon, D. R.; Maji, S.; Kelly, J. F.; et al. 2023. Quantifying long-term phenological patterns of aerial insectivores roosting in the Great Lakes region using weather surveillance radar. *Global Change Biology*, 29(5): 1407–1419.

DIUx-xView. 2020. Diux-xView/xview2_first_place: 1st place solution for ”Xview2: Assess building damage” challenge.

Gupta, R.; Hosfelt, R.; Sajeev, S.; Patel, N.; Goodman, B.; Doshi, J.; Heim, E.; Choset, H.; and Gaston, M. 2019. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Hesterberg, T. C. 1996. Estimates and confidence intervals for importance sampling sensitivity analysis. *Mathematical and computer modeling*, 23(8-9): 79–85.

Horn, J. W.; and Kunz, T. H. 2008. Analyzing NEXRAD doppler radar images to assess nightly dispersal patterns and population trends in Brazilian free-tailed bats (Tadarida brasiliensis). *Integrative and Comparative Biology*, 48(1): 24–39.

Kim, K.; and Yoon, S. 2018. Assessment of Building Damage Risk by Natural Disasters in South Korea Using Decision Tree Analysis. *Sustainability*, 10(4).

Kossen, J.; Farquhar, S.; Gal, Y.; and Rainforth, T. 2021. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*.

Leitloff, J.; Hinz, S.; and Stilla, U. 2010. Vehicle detection in very high resolution satellite images of city areas. *IEEE transactions on Geoscience and remote sensing*, 48(7): 2795–2806.

Lin, T.-Y.; Winner, K.; Bernstein, G.; Mittal, A.; Dokter, A. M.; Horton, K. G.; Nilsson, C.; Van Doren, B. M.; Farnsworth, A.; La Sorte, F. A.; et al. 2019. MistNet: Measuring historical bird migration in the US using archived weather radar data and convolutional neural networks. *Methods in Ecology and Evolution*, 10(11): 1908–1922.

Meng, C.; Liu, E.; Neiswanger, W.; Song, J.; Burke, M.; Lobell, D.; and Ermon, S. 2022. Is-count: Large-scale object counting from satellite images with covariate-based importance sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Nguyen, P.; Ramanan, D.; and Fowlkes, C. 2018. Active Testing: An Efficient and Robust Framework for Estimating Accuracy. In *International Conference on Machine Learning (ICML)*.

Norouzzadeh, M. S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M. S.; Packer, C.; and Clune, J. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725.

Nurkarim, W.; and Wijayanto, A. 2023. Building footprint extraction and counting on very high-resolution satellite imagery using object detection deep learning framework. *Earth Sci Inform*.

Oceanic, N.; and (NOAA), A. A. 2023. Next generation weather radar (NEXRAD). *National Centers for Environmental Information (NCEI)*.

Owen, A. B. 2013. *Monte Carlo theory, methods and examples*.

Perez, G.; Zhao, W.; Cheng, Z.; T. D. Belotti, M. C.; Deng, Y.; Simons, V. F.; Tielens, E.; Kelly, J. F.; Horton, K. G.; Maji, S.; and Sheldon, D. 2022. Using spatio-temporal information in weather radar data to detect and track communal bird roosts. *bioRxiv*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Rosenberg, K. V.; Dokter, A. M.; Blancher, P. J.; Sauer, J. R.; Smith, A. C.; Smith, P. A.; Stanton, J. C.; Panjabi, A.; Helft, L.; Parr, M.; and Marra, P. P. 2019. Decline of the North American avifauna. *Science*, 366(6461): 120–124.

Sánchez-Bayo, F.; and Wyckhuys, K. A. 2019. Worldwide decline of the entomofauna: A review of its drivers. *Biological conservation*, 232: 8–27.

Settles, B. 2009. Active learning literature survey.

Singh, G.; Singh, S.; Sethi, G.; and Sood, V. 2022. Deep Learning in the Mapping of Agricultural Land Use Using Sentinel-2 Satellite Data. *Geographies*, 2(4): 691–700.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

The Defense Innovation Unit. 2020. The xview2 AI challenge.

Tuia, D.; Kellenberger, B.; Beery, S.; Costelloe, B. R.; Zuffi, S.; Risse, B.; Mathis, A.; Mathis, M. W.; van Langevelde, F.; Burghardt, T.; et al. 2022. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1): 792.

Turkoglu, M. O.; D'Aronco, S.; Perich, G.; Liebisch, F.; Streit, C.; Schindler, K.; and Wegner, J. D. 2021. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment*, 264: 112603.

van der Vaart, A. W.; and Wellner, J. A. 1996. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The Inaturalist species classification and detection dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Žnidarič, M. 2009. Asymptotic Expansion for Inverse Moments of Binomial and Poisson Distributions. *The Open Mathematics, Statistics and Probability Journal*, 1(1).

Wah, C.; Van Horn, G.; Branson, S.; Maji, S.; Perona, P.; and Belongie, S. 2014. Similarity comparisons for interactive fine-grained categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Winkler, D. W. 2006. Roosts and migrations of swallows. *Hornero*, 21(2): 85–97.

Won, M. 2020. Intelligent traffic monitoring systems for vehicle classification: A survey. *IEEE Access*, 8: 73340–73358.

Wu, Z.; Zhang, C.; Gu, X.; Duporge, I.; Hughey, L.; Stabach, J.; Skidmore, A.; Hopcraft, G.; Lee, S.; Atkinson, P.; McCauley, D.; Lamprey, R.; Ngene, S.; and Wang, T. 2023. Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature Communications*, 14: 3072.

Yeh, C.; Perez, A.; Driscoll, A.; Azzari, G.; Tang, Z.; Lobell, D.; Ermon, S.; and Burke, M. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11: 2583.