IS-Count: Large-Scale Object Counting from Satellite Images with Covariate-Based Importance Sampling

Chenlin Meng*, Enci Liu*, Willie Neiswanger, Jiaming Song, Marshall Burke, David Lobell, Stefano Ermon

Stanford University

{chenlin, jesslec, neiswanger, tsong}@cs.stanford.edu, {mburke, dlobell}@stanford.edu, ermon@cs.stanford.edu

Abstract

Object detection in high-resolution satellite imagery is emerging as a scalable alternative to on-the-ground survey data collection in many environmental and socioeconomic monitoring applications. However, performing object detection over large geographies can still be prohibitively expensive due to the high cost of purchasing imagery and compute. Inspired by traditional survey data collection strategies, we propose an approach to estimate object count statistics over large geographies through sampling. Given a cost budget, our method selects a small number of representative areas by sampling from a learnable proposal distribution. Using importance sampling, we are able to accurately estimate object counts after processing only a small fraction of the images compared to an exhaustive approach. We show empirically that the proposed framework achieves strong performance on estimating the number of buildings in the United States and Africa, cars in Kenya, brick kilns in Bangladesh, and swimming pools in the U.S., while requiring as few as 0.01% of satellite images compared to an exhaustive approach.

Introduction

The quantity and location of human-made objects are key information for the measurement and understanding of human activity and economic livelihoods. Such physical capital—for instance, buildings, cars, and roads—is both an important component of current economic well-being as well as an input into future prosperity. Information on physical capital has traditionally been derived from ground-based surveys of households, firms, or communities (Bureau of Economic Analysis 2003). However, because such surveys are expensive and time consuming to conduct, key data on physical capital and related livelihood measures are lacking for much of the world, inhibiting our understanding of the patterns and determinants of economic activity (Burke et al. 2021).

Object detection in high-resolution satellite imagery has emerged as a scalable alternative to traditional survey-based approaches to gathering data on economic activity. For instance, imagery-based counts of the number of buildings at country level allows policymakers to monitor progress towards economic development (Ayush et al. 2021; Uzkent, Yeh, and Ermon 2020; Uzkent and Ermon 2020; Sheehan et al. 2019; Blumenstock, Cadamuro, and On 2015; Yeh et al. 2020), counts of the number of brick kilns allows environmental scientists to track pollution from informal industries (Lee et al. 2021), counts of multiple objects enables accurate poverty prediction (Jean et al. 2016; Ayush et al. 2021), and counts of solar panels in high-resolution imagery enables understanding of green energy adoption at broad scale (Yu et al. 2018).

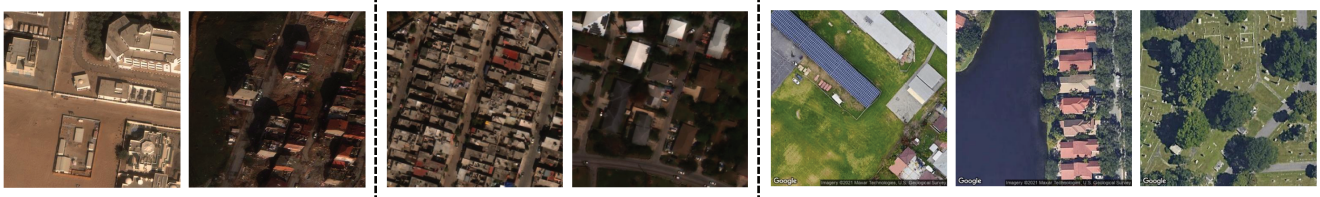
However, due to the substantial cost of purchasing high-resolution imagery, and the large amount of computation needed to estimate or apply models at scale, performing object detection over large geographies is often prohibitively expensive (Uzkent et al. 2019; Uzkent and Ermon 2020), especially if estimates need to be updated. For instance, at standard pricing for high-resolution imagery, purchasing country-wide imagery for one year would cost roughly $3 million for Uganda, $15 million for Tanzania, and $38 million in the Democratic Republic of Congo.[1] Such costs inhibit the widespread application and adoption of satellite-based approaches for livelihood measurement.

Here we propose an importance-sampling approach to efficiently generate object count statistics over large geographies, and validate the approach across multiple continents and object types. Our approach draws inspiration from traditional approaches to large-scale ground survey data collection, which use information from prior surveys (*e.g.*, a prior census) to draw sample locations with probability proportionate to some covariate of interest (*e.g.*, village population). In our setting, we expect most objects of interest (*e.g.*, cars) to have close to zero density in certain regions (*e.g.* forested areas). In this case, sampling locations uniformly at random (*i.e.*, with a uniform proposal) would have a high variance and require a large number of samples. We therefore propose to use importance sampling (IS) to select locations from important regions (*e.g.* regions where the counts of cars are expected to be non-zero) by sampling from a proposal distribution. While a good proposal can significantly reduce variance, the optimal proposal distribution is unknown, often complicated, and object-specific. We therefore propose to learn the proposal distribution by relating the

*Joint first authors.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] We assume a price per sq km of $17 for 3-band imagery, consistent with current industry rates.



(a) Evaluated on training set (b) Evaluated on test set (c) Evaluated on different data source

Figure 1: Predicted building counts of a YOLO-v3 model (Redmon and Farhadi 2017) trained on xView (Lam et al. 2018). Ground truth building counts are highlighted in green and provided in the parenthesis next to the model prediction. We observe that the model has an almost perfect prediction on the training set (see Figure (a)). However, the performance degrades on the test set (see Figure (b)) and satellite imagery from other data sources (see Figure (c)). Additional training details of the object counting model are provided in Appendix.

presence (or absence) of objects to widely available covariates such as population density, nightlights, etc.

Given a cost budget, our method, IS-Count, selects a small number of informative areas, using an object detector or gold-standard human annotations for object counting. It largely reduces the number of satellite images as well as human annotations compared to an exhaustive approach used by object detectors in many real-world counting tasks, while achieving a high accuracy. We show with experiments on counting buildings, cars, brick kilns, and swimming pools in Africa, United States, and Bangladesh, that IS-Count is able to achieve an accurate estimation while requiring as few as 0.01% images compared to an exhaustive approach, thus potentially reducing the cost of generating large-scale object counts by four orders of magnitude. Our code is available at https://github.com/sustainlab-group/IS-Count.

Problem Setup

Given a region $R \subseteq \mathbb{R}^2$ (*e.g.* a country) and a class of target object (*e.g.* buildings) demonstrated with a small number of labeled images, we want to estimate the total number of objects of the desired class that are located within the target region R. We denote $f(\mathbf{x}, l)$ the total number of objects within a $l \times l$ bounding box centered at point $\mathbf{x}$ (see Figure 2). Given a partition of R into non-overlapping $l \times l$ images with centers $\mathcal{S}$, the goal is to estimate

$$C = \sum_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x}, l). \tag{1}$$

While the object count C is less informative than the precise location of all the objects in R, object counts are often sufficient in many downstream tasks such as regression analyses, e.g. to estimate poverty levels from satellite images (Ayush et al. 2021, 2020). Additionally, we will demonstrate that object counts can be obtained much more efficiently, allowing us to scale over large regions where exhaustive mapping would be infeasible.

A naive solution to estimate C is to acquire all image tiles covering the entire region R, identify the objects in each image (*e.g.*, using a machine learning model), and sum the counts (Crowther et al. 2015; Yu et al. 2018; Yi et al. 2021b). However, this approach has the following limitations.

Figure 2: Given a location $\mathbf{x}$, $f(\mathbf{x}, l)$ denotes the total number of objects (e.g. buildings shown here), within the bounding box with size $l \times l$ centered at $\mathbf{x}$.

Satellite Images are Costly In many applications, we need to use high-resolution images for the target objects (*e.g.*, cars) to be visible. Although low-resolution images are publicly available, high-resolution images typically need to be purchased from private vendors. For instance, it costs approximately $164 million for purchasing high-resolution satellite images that cover the entire United States[2] , which is often infeasible for researchers.

Labeling is Expensive If the target region is large, it would be impractical for human annotators to manually count objects in all images. For instance, we estimate it would take approximately 115,000 hours for human annotators to count the total number of cars in Kenya using satellite images[3] (see Appendix). A more efficient approach is to use an algorithm (typically a machine learning model) to estimate counts (Gao, Liu, and Wang 2020; Bazi, Al-Sharari, and Melgani 2009) or detect objects (Crowther et al. 2015; Yu et al. 2018; Salamí et al. 2019; Mubin et al. 2019; Yi et al. 2021b) within each image. However, training such a model often requires very large amounts (*e.g.* 360k) of labeled data (Yu et al. 2018), whose labels eventually come from human annotators. As the distribution of satellite im-

[2]https://g3f3z9p8.rocketcdn.me/wp-content/uploads/2018/04/landinfo.com-LAND_INFO_Satellite_Imagery_Pricing.pdf

[3]Estimation based on Amazon Mechanical Turk

ages often changes drastically across data sources (*e.g*. Digital Globe and Google Static Map), objects of interest (*e.g*. buildings and farmland), and regions (*e.g*. U.S. and Africa), an object detector pre-trained on one dataset could fail easily on another due to covariate shifts (Yi et al. 2021a) (see Figure 1). This makes it hard to directly apply a pre-trained object detector to a new task where sources of satellite images are different even if large-scale labeled datasets are available for the object of interest.

Sampling Denote S_R the area of R, and $U(R)$ the uniform distribution on R. When $l \ll R$, the total number of objects of interest C in the region R can be computed as

$$C = \frac{S_R}{l^2}\mathbb{E}_{\mathbf{x}\sim U(R)}[f(\mathbf{x}, l)]. \tag{2}$$

The following unbiased estimator is often used to evaluate Equation (2)

$$\hat{C} = \frac{S_R}{l^2}\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i, l),\ \mathbf{x}_i \sim U(R), \tag{3}$$

where $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. samples from the uniform distribution $U(R)$. This can drastically reduce cost if n is small, even allowing for (gold-standard) human annotations for evaluating $f(\mathbf{x}_i, l)$.

In real-world applications, however, it is expected that the object of interest (*e.g*., buildings) can have a close to zero density in certain regions (*e.g*., forest). In this case, estimating object counts directly via uniform sampling would have a high variance and thus require a huge number of samples as we will show in the experimental section.

IS-Count: Large-scale Object Counting with Importance Sampling

In this paper, we alleviate the above challenges by proposing an efficient object counting framework that incorporates prior knowledge from socioeconomic indicators into importance sampling (IS). Our method, IS-Count, provides an unbiased estimation for object counts and requires as few as 0.01% high-resolution satellite images compared to an exhaustive approach (see Figure 3). IS-Count can be easily adapted to different target regions or object classes using a small number of labeled images, while achieving strong empirical performance.

Importance Sampling

Importance sampling (IS) introduces a proposal distribution to choose representative samples from important regions (*e.g*., regions where the counts are non-zero) with the goal of reducing the variance of uniform sampling. Given a proposal distribution $q(\mathbf{x})$ with a full support on R, we can estimate the total object count using importance sampling

$$C = \frac{1}{l^2}\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\frac{f(\mathbf{x}, l)}{q(\mathbf{x})}\right]. \tag{4}$$

Equation (4) can be approximated by the following unbiased estimator

$$\hat{C}_n = \frac{1}{l^2}\frac{1}{n}\sum_{i=1}^{n}\frac{f(\mathbf{x}_i, l)}{q(\mathbf{x}_i)},\ \mathbf{x}_i \sim q(\mathbf{x}), \tag{5}$$

Algorithm 1: Object counting with IS-Count

Input: Region R, object class, budget n, covariate, and a small number of labeled examples
Output: Estimated object count

1: $q(\mathbf{x}) \leftarrow$ covariate distribution
2: Fine-tune $q(\mathbf{x})$ with labeled examples
3: Sample $\{\mathbf{x}_i\}_{i=1}^n$ from $q(\mathbf{x})$
4: $\hat{C}_n \leftarrow$ Estimate C using Equation (5)
5: **return** $\hat{C}_n$

where $\{\mathbf{x}_i\}_{i=1}^N$ are i.i.d. samples from $q(\mathbf{x})$. The optimal proposal distribution $q^*(\mathbf{x})$ which has the smallest variance should be proportional to $f(\mathbf{x}, l)$ (Owen 2013).

We therefore want to design a proposal distribution that is as close as possible to the object density. Although high-resolution images are costly, socioeconomic covariates such as nightlight intensity are globally available, free of charge, and correlate strongly human activities (Jean et al. 2016).

In the following, we assume that we always have access to certain covariates that are cheap and publicly available for the target region. We treat the covariate as the base distribution for designing the proposal distribution $q(\mathbf{x})$. In order for the base distribution to capture information specialized to the task, we propose to fine-tune the base distribution using a small number of labeled satellite images, where the labels are count statistics (see Figure 3). We also provide the pseudocode for the framework in Algorithm 1.

Our key insight is that the base covariate distribution can provide good prior knowledge for a given task, and therefore we only need a small number of labeled images for fine-tuning to obtain a task-specific proposal distribution that reduces the variance for sampling. As this framework only requires a small amount of labeled images for each task and always provides an unbiased estimate, it can be easily adapted to different counting tasks, providing a general framework for large-scale counting.

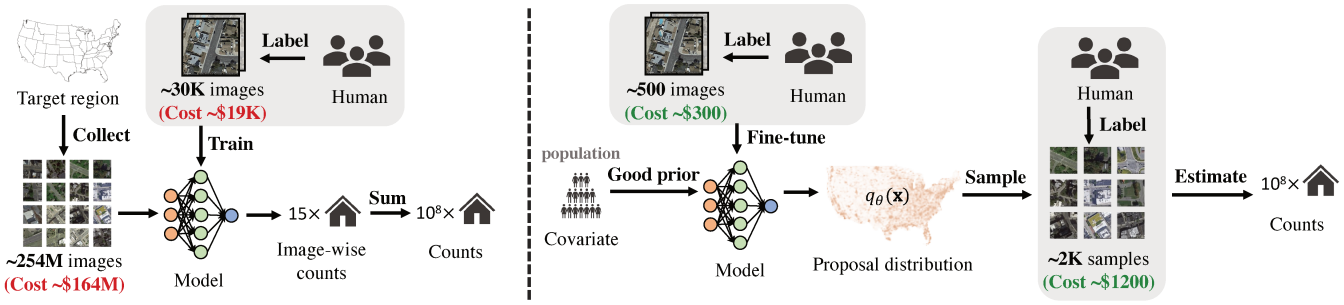
Proposal Distributions with Task-specific Tuning

Given the base covariate distributions, we can fine-tune the proposal distribution using a small number of labeled satellite images to design a task-specific proposal distribution. Isotonic regression provides an approach to learning a non-decreasing transformation, allowing fine-tuning the proposal distribution based on the input covariate distribution. More specially, let $h(\mathbf{x}) \in \mathbb{R}$ be the covariate pixel value at geolocation $\mathbf{x}$, and $f(\mathbf{x}, l)$ be the object count (see Figure 2), we learn a non-decreasing map $g_\theta(\cdot, l) : \mathbb{R} \to \mathbb{R}$ which maps $h(\mathbf{x})$ to be close to its corresponding object count $f(\mathbf{x}, l)$. The objective function is defined as

$$\theta = \arg\min_{\theta} \sum_{i=1}^{n} w_i(g_\theta(h(\mathbf{x}_i), l)) - f(\mathbf{x}_i, l))^2, \tag{6}$$

subject to $g_\theta(a, l) \leq g_\theta(b, l)$ whenever $a \leq b$. In Equation (6), w_i are positive weights and $\{\mathbf{x}_i\}_{i=1}^n$ are coordinates sampled from the target region R.

Although it might seem natural to use a general regression model, which takes multiple input covariates, to predict the



(a) Object counting with an exhaustive approach

(b) Object counting with IS-Count

Figure 3: Object counting frameworks comparison (example on the US.). Figure (a): An exhaustive approach downloads all image tiles covering the target region, maps the objects in each image using a trained model, and takes the summation of counts in all the images to produce a total count. However, purchasing satellite imagery for a large target region can be expensive. Figure (b): In contrast, IS-Count selects a small number of informative areas for object counting by sampling from a learnable proposal distribution which captures the representative areas. IS-Count largely reduces the number of satellite images and human annotations while achieving a high accuracy.

corresponding object count, we observe that such approach requires a much larger number of training data to learn a robust model. As the training size is deducted from the sampling budget, a general regression does not have a strong estimation performance due to the reduction in sample size. We observe that isotonic regression, being more restrictive, has stronger empirical performance than general regressions when the training size is limited, potentially due to the strong inductive bias of being monotonic so that a higher covariate value would be mapped to a higher object density.

Given the learned model $g_\theta(\cdot, l)$, the proposal distribution $q_\theta(\mathbf{x})$ can be derived as

$$q_\theta(\mathbf{x}, l) \triangleq \frac{g_\theta(h(\mathbf{x}), l)}{\int_{\mathbf{x}} g_\theta(h(\mathbf{x}), l) d\mathbf{x}}, \tag{7}$$

where $\int_{\mathbf{x}} g_\theta(h(\mathbf{x}), l) d\mathbf{x}$ is computed by taking the summation of all covariate elements weighted by their volumes in the region R, which can be achieved because there are only finitely many elements in the covariate dataset. As a special case, when $g_\theta(h(\mathbf{x}))$ is proportional to $h(\mathbf{x})$ for all $\mathbf{x} \in R$, sampling from $q_\theta(\mathbf{x}, l)$ is equivalent to sampling proportionally to the covariate distribution. We provide a visualization of a learned proposal distribution in Figure 4.

Theoretical Analysis

Given the estimation $\hat{C}_n$ from Equation (5), we can bound its estimation error using the Kullback-Leibler (KL) divergence between $q^*(\mathbf{x})$ and $q_\theta(\mathbf{x}, l)$. As l is a constant, we use $q_\theta(\mathbf{x})$ to denote $q_\theta(\mathbf{x}, l)$ for simplicity.

Proposition 1 *Suppose* $\frac{q^*(\mathbf{x})}{q_\theta(\mathbf{x})}$ *is well defined on* R*, let* $L = KL(q^*(\mathbf{x}) \| q_\theta(\mathbf{x}))$*,* i.e. *the KL divergence between* $q^*(\mathbf{x})$ *and* $q_\theta(\mathbf{x})$*. If* $n = \exp(L + t)$ *for some* $t \geq 0$*, then*

$$\mathbb{E}[|\hat{C}_n - C|] \leq C\left(e^{-\frac{t}{4}} + 2\sqrt{\mathbb{P}(\log \tfrac{q^*(\mathbf{x})}{q_\theta(\mathbf{x})} > L + \tfrac{t}{2})}\right). \tag{8}$$

Proposition 1 provides a bound for the estimation error of $\hat{C}_n$ using n samples. Intuitively, when $p(\mathbf{x})$ is close to $p^*(\mathbf{x})$, the count estimation using sampling is more precise with limited samples. When $q_\theta(\mathbf{x}) = q^*(\mathbf{x})$ almost everywhere, we have $\mathrm{KL}(q^*(\mathbf{x}) \| q_\theta(\mathbf{x})) = 0$, which implies

$$\mathbb{E}[|\hat{C}_{\exp(t)} - C|] \leq Ce^{-\frac{t}{4}}. \tag{9}$$

The proof of Proposition 1 can be derived from Chatterjee and Diaconis, and we provide details in Appendix.

We can also provide a probability lower bound for the estimation $\hat{C}_n$ using Markov inequality.

Theorem 1 *For any* $k > 0$

$$\mathbb{P}[\hat{C}_n \geq kC] \leq \frac{1}{k}. \tag{10}$$

Theorem 1 implies that the probability that the estimation $\hat{C}_n$ is k times larger than the ground truth (GT) counts C is no more than $\frac{1}{k}$, providing a probability lower bound for the estimation. Intuitively, we would want our proposal distribution to satisfy that $\frac{q_\theta(\mathbf{x})}{f(\mathbf{x}, l)}$ has a small variance, so that $q_\theta(\mathbf{x})$ will be close to $q^*(\mathbf{x})$. In the following, we provide a lower bound for the variance of $\frac{q_\theta(\mathbf{x})}{f(\mathbf{x}, l)}$ based on the KL divergence between q^* and q.

Theorem 2 *The variance of* $\frac{q_\theta(\mathbf{x})}{f(\mathbf{x}, l)}$ *is an upper bound to* $C^2(e^{KL(q^*(\mathbf{x}) \| q_\theta(\mathbf{x}))} - 1)$.

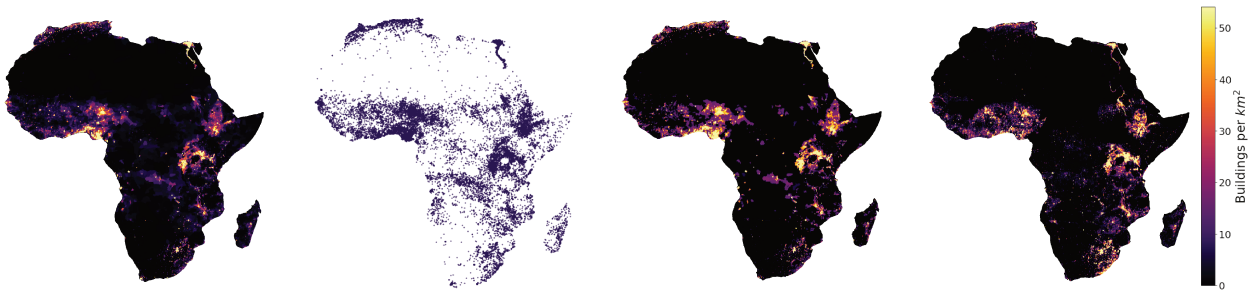
Theorem 2 implies that the variance of the estimator cannot be too small unless the KL divergence between q^* and q is also small, which encourages us to reduce the KL divergence. We provide the proof in Appendix.

Experimental Setup

To show the effectiveness, generality, and scalability of IS-Count, we conduct experiments on multiple real-world applications.

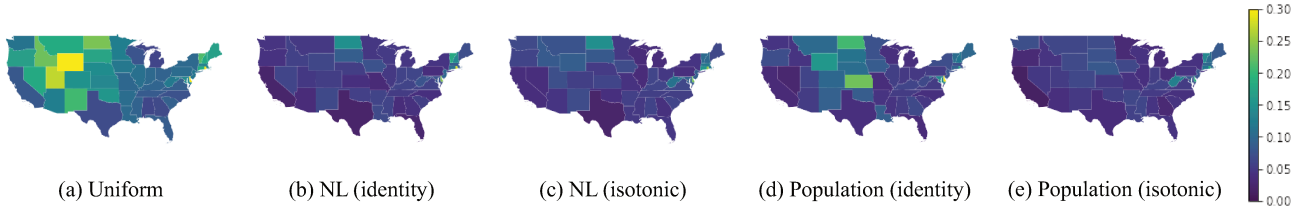
Tasks and Datasets

We consider the following four tasks across 45 countries: 1) counting buildings in the US and 43 African countries; 2) counting cars in Kenya; 3) counting brick kilns in



(a) Population density (b) Samples from (a) (c) Learned proposal distribution (d) Building density

Figure 4: Visualization of distributions in Africa. Figure (a) shows the population density in Africa; Figure (b) shows samples drawn from the population density shown in (a); Figure (c) shows the proposal distribution learned by isotonic regression with (a) as inputs; Figure (d) shows the building density derived from the Google Open Buildings dataset (Sirko et al. 2021).



(a) Uniform (b) NL (identity) (c) NL (isotonic) (d) Population (identity) (e) Population (isotonic)

Figure 5: Estimation error (see Equation (11)) on the building counts in the contiguous United States (the darker the better). We treat the count from the MS Building Footprints as the ground truth to compare with. All results are averaged over 20 runs and using 0.2% images, meaning that the ratio between the total area covered by the used images in the budget and the area of the target region is 0.2%.

Bangladesh; and 4) counting swimming pools in the US. We emphasize that our approach is generalizable to other counting tasks, and the tasks and the regions of interest are chosen such that the regions are reasonably large and validation data are either available or readily obtained from human labeling of images. We provide more details in Appendix.

Task 1: Buildings Building count provides insights into urban development and assists humanitarian efforts (Sirko et al. 2021). We evaluate IS-Count on counting buildings in the US. and 43 African countries. We use the Microsoft Building Footprints[4] and the Google Open Buildings (Sirko et al. 2021) as the ground truth building distribution for the US. and Africa respectively. The image-wise building counts are directly extracted from the corresponding datasets.

Task 2: Cars The number of cars correlates with the economic development of a region, especially in low-income countries (Litman and Laube 2002; Li et al. 2020). We focus on counting cars in a region roughly covering Nairobi, the capital of Kenya. We collect all satellite images covering the entire region, and hand-label the count of cars in all images, the sum of which is treated as the ground truth.

Task 3: Brick kilns Brick manufacturing is a major source of pollution in South Asia but is dominated by small-scale producers who are hard to monitor (Lee et al. 2021). Understanding the distribution of brick kilns is thus of importance for policymakers. In this section, we perform experiments on counting brick kilns in Bangladesh. We use the dataset from (Lee et al. 2021) and treat their results as the ground truth count and collect image-wise counts as for buildings.

Task 4: Swimming pools The number of swimming pools informs policy makers about urban planning and assists larger-scale water quality control (Hlavsa et al. 2021). In this task, we estimate the count of swimming pools in the US. at country-level. It is estimated that there are 10,709,000 swimming pools in the United States (Taylor 2020), which we treat as the ground truth. As we are not aware of existing datasets on swimming pools, we sample a small amount of images and collect the counts from human annotators (Amazon MTurk) for estimating the total count.

Base Distributions from Covariates

For constructing base proposal distributions, we focus on two covariates, population density and nightlight intensity, and leave the exploration of other covariates for future work.

Population The population density raster is a single-channel image, with each pixel a positive float value denoting the (estimated) population for a 1000m$\times$1000m area on the ground (see Figure 4). We can derive a density function $q(\mathbf{x})$ to be proportional to the population density by dividing each pixel value with the normalization constant—the summation of all pixel values in the population raster. We treat

[4]https://github.com/microsoft/USBuildingFootprints

	US (16.77%)		Africa (21.02%)	
Methods	NL	Population	NL	Population
Identity	8.46%	9.71%	15.58%	18.27%
Isotonic	8.80%	8.08%	15.79%	17.43%
Isotonic*	**8.09%**	**7.48%**	**14.86%**	**16.85%**

Table 1: Averaged error over all states in the US. and 43 African countries. All results are averaged over 20 runs. The total area of used satellite images covers 0.1% of each target state (US) and country (Africa). The result of uniform sampling is provided in the parenthesis next to the region name. We use 20% of the budget for training isotonic regression.

the density within each pixel as a uniform distribution. The derived density $q(\mathbf{x})$ can be used as the proposal distribution for IS-Count. More details can be found in Appendix.

Nightlight We use the global nightlight raster, which is also a single channel image, with each pixel a positive float value denoting the nightlight intensity for a 750m×750m area (see Appendix **??**). Similarly, we can derive a density function $q(\mathbf{x})$ to be proportional to the nightlight intensity and treat the density within each pixel to be uniform.

Experiment Settings

We consider the following three settings for constructing the proposal distribution for IS-Count.

Identity We directly use the base covariate distribution as the proposal distribution without learning.

Isotonic We fine-tune the base proposal distribution with isotonic regression using a small number of labeled samples (*e.g.*, 100 samples). We deduct the size of the training data from the total budget, meaning that the larger the training set, the fewer satellite images we can sample for count estimation.

Isotonic* Depending on task, there could already exist a certain amount of observed labeled data, which could potentially be sampled from an unknown distribution. Although these data might not be used for count estimation as they are not sampled from the proposal distribution, they can still be used to fine-tune the proposal distribution. This observation motivates us to have the second isotonic setting where the size of the training data is not deducted from the total budget.

Evaluation Metrics

We evaluate the performance using percent error defined as

$$\text{Error} = \frac{|\hat{C}_n - C|}{C} \times 100\%, \tag{11}$$

where C is the "ground truth" (GT) object count obtained from existing datasets or human annotators, and $\hat{C}_n$ is the estimation using n samples (see Equation (5)).

Results

In this section, we evaluate the performance of IS-Count on the tasks introduced in the previous section. We show with empirical results that IS-Count drastically reduces the variance of sampling in most settings, leading to huge savings of up to $163millions for purchasing high-resolution satellite images and 1 million hours for image labeling with human annotators compared to an exhaustive approach. We provide extra details in Appendix.

IS-Count with Base Proposal Distributions

To evaluate the performance of importance sampling with covariates as the proposal distributions, we compare IS-Count (identity) with uniform sampling in this section. In Table 2, we show the errors of object count estimation in different tasks, where IS-Count consistently outperforms uniform sampling by a large margin on counting buildings and cars. Moreover, as sample size increases, the estimates based on IS-Count converge quickly while the estimates based on uniform sampling show no obvious trend of convergence (see Figure 6(b)), whereas covariate-based estimates have reduced variance.

It is interesting to note that all methods give a high error rate on the count of swimming pools, and all estimates converge to approximately the same value (see the last column of Table 2). One plausible reason is that a significant number of swimming pools are indoors, and therefore not visible in satellite images. However we are not aware of data sources on the count of outdoor swimming pools to perform additional evaluation. Given that all approaches in Table 2 converge to approximately the same counts and our method has strong performance on the other tasks, we believe IS-Count should provide a reasonably accurate estimate of outdoor swimming pool counts.

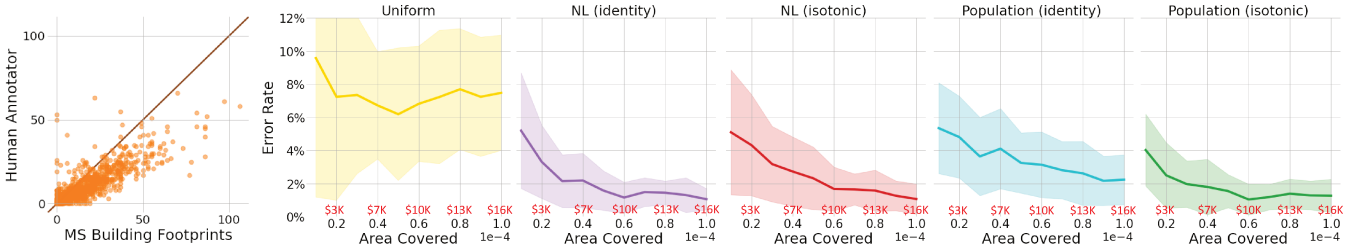
The choice of covariates for building the proposal distribution also affects the estimation. On the car counting task (see Table 2), NL (identity) method outperforms both uniform and population (identity) methods, while NL (identity) does not outperform the uniform sampling on brick kiln counting (see Table 2). We believe one potential reason is that the distribution of cars could be more correlated with NL than population, while the distribution of brick kilns could be more correlated with population than NL. Therefore, in order to generate a task-specific proposal distribution with strong performance, we propose to fine-tune the covariate distribution using isotonic regression.

IS-Count with Tuned Proposal Distributions

We observe that learning the proposal distribution with isotonic regression further improves the performance in most of the settings (Table 2). In car and brick kiln experiments, we observe that fine-tuning with isotonic methods consistently improves the performance of identity methods even with the training size deducted from the sampling budget. In addition, in Figure 6(b), we observe that the population (isotonic) proposal distribution converges fastest to the ground truth building count, compared to the population (identity) method. We believe our empirical results support the effectiveness of IS-Count.

	Buildings (US)		Cars (Nairobi region)		Brick Kilns (Bangladesh)		Pools (US)
Percentage sampled	0.001%	0.01%	2.6%	3.9%	2.0%	4.0%	0.001%
Uniform	9.58 ± 8.39	7.48 ± 3.48	9.63 ± 7.27	7.76 ± 4.60	9.83 ± 5.66	4.72 ± 4.35	47.93 ± 15.42
NL (identity)	5.20 ± 3.49	1.06 ± 0.64	5.60 ± 2.94	4.37 ± 3.34	11.49 ± 6.98	7.10 ± 5.02	64.44 ± 4.68
NL (isotonic)	5.10 ± 3.77	1.07 ± 0.91	6.28 ± 4.26	4.12 ± 2.32	7.46 ± 5.35	4.62 ± 2.78	46.80 ± 16.83
NL (isotonic*)	4.16 ± 3.77	$\mathbf{0.63 \pm 0.50}$	$\mathbf{5.02 \pm 4.11}$	$\mathbf{4.00 \pm 2.96}$	$\mathbf{6.26 \pm 3.94}$	$\mathbf{4.61 \pm 3.08}$	50.35 ± 6.46
Population (identity)	5.35 ± 2.74	2.24 ± 1.50	9.89 ± 8.18	7.76 ± 6.50	8.28 ± 5.62	6.02 ± 4.23	45.69 ± 7.10
Population (isotonic)	4.03 ± 2.17	1.27 ± 1.00	7.48 ± 5.36	6.66 ± 4.11	6.73 ± 5.50	4.75 ± 4.17	61.79 ± 2.62
Population (isotonic*)	$\mathbf{3.74 \pm 2.43}$	0.80 ± 0.60	5.27 ± 3.63	4.30 ± 2.74	7.39 ± 5.23	4.65 ± 3.04	63.70 ± 4.18
Cost saved on images ($)	163,692,910	163,678,178	22,906	22,600	2,459,183	2,408,995	163,692,910
Labeling time saved (hrs)	1,871,956	1,871,788	266	263	28,123	27,549	1,871,956

Table 2: Error rate (%) of object count estimation using different methods (averaged over 20 runs). "Percentage sampled" denotes the ratio between the total area covered by the used images and the area of the target region. For isotonic based methods, we use 20% of the total budget as the training size. We report the cost and time saved for purchasing high-resolution images and labeling with human compared to an exhaustive approach on the target region. Details are provided in Appendix.



(a) Human vs. GT Count

(b) Error Rate vs. Proportion of Area Covered by Samples

Figure 6: Figure (a): Counts from existing building datasets are consistent with counts from human annotators on the same images. Due to potential tendencies of excluding objects on the boundaries, human annotator can give slightly smaller counts than counts from the building datasets. Figure (b): Estimation error of building counts at country-level for the United States averaged over 20 runs. The total number of buildings in the US. from MS Building Footprints is treated as the ground truth (GT). We plot the ratio between the area covered by the used images and the area of the US. on the x-axis and label the corresponding cost for purchasing images (in USD). We observe that IS-Count has smaller variances than uniform sampling and converges faster to the true building count than uniform sampling.

Cost Analysis

We compare the cost required for purchasing images for the exhaustive approach and IS-Count in the last two rows of Table 2. We observe that IS-Count saves as much as 99.99% of the cost for purchasing images and 99.99% hours needed for labeling with human annotators[5]. When the target region is as large as the US, IS-Count saves $163 million for purchasing images and 1 million hours for labelling images, while achieving less than 1% error on building count estimation. We provide more details on cost estimation in Appendix.

Discussion and Societal Impact

Understanding where humans build things and what they build is a central component of measuring the productivity of economies and the livelihoods of individuals worldwide. Knowledge of this physical capital is also central for a range of policy decisions, from planning infrastructure investments to delivering services to adapting to climate change. Finally, accurate measurement of the physical capital stock over time is central to answering fundamental questions about sustainable development, in particular in understanding whether we are leaving future generations as well or better off than current and past generations (Solow 1992).

We provide a new approach to accurately estimate object counts at large scale and low cost. Compared to existing brute-force approaches which require purchasing vast amounts of high-resolution images at very high cost (*e.g.*, $164 million for the entire US), our approach enables high accuracy counts of objects while eliminating 99.99% of the cost of purchasing and labeling imagery. As our IS-Count model is scalable to large-scale object counting and can be easily adapted to different tasks, it is applicable to a broad range of research and policy tasks, including contributing to near-term tracking of multiple Sustainable Development Goals. For instance, the number of cars and buildings reflects economic development in low-income countries and can be used to measure livelihoods directly (Ayush et al. 2021) (SDG 1 No Poverty), and the number of brick kilns reflects pollution from informal industries (SDG 13 Climate Action). Our approach can also make real-world counting efforts more inclusive for researchers and policymakers with more limited budgets, including those from developing countries, democratizing contributions and facilitating progress towards real-world sustainability applications.

[5]Hour estimation based on Amazon Mechanical Turk

Acknowledgements

The authors would like to thank everyone from the Stanford Sustainability and AI Lab for the constructive feedback and discussion. This work was supported by NSF awards (#1651565, #1522054), the Stanford Institute for Human-Centered AI (HAI), the Stanford King Center, the United States Agency for International Development (USAID), a Sloan Research Fellowship, and the Global Innovation Fund. WN was supported in part by ONR (N000141912145), AFOSR (FA95501910024), ARO (W911NF-21-1-0125), DOE (DE-AC02-76SF00515).

References

Ayush, K.; Uzkent, B.; Burke, M.; Lobell, D.; and Ermon, S. 2020. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612*.

Ayush, K.; Uzkent, B.; Tanmay, K.; Burke, M.; Lobell, D.; and Ermon, S. 2021. Efficient Poverty Mapping from High Resolution Remote Sensing Images. In *AAAI*.

Bazi, Y.; Al-Sharari, H.; and Melgani, F. 2009. An automatic method for counting olive trees in very high spatial remote sensing images. In *2009 IEEE International Geoscience and Remote Sensing Symposium*, volume 2, II–125. IEEE.

Blumenstock, J.; Cadamuro, G.; and On, R. 2015. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264): 1073–1076.

Bureau of Economic Analysis. 2003. Consumer Durable Goods in the United States, 1925-99. *Washington, DC: US Department of Commerce*.

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

Chatterjee, S.; and Diaconis, P. 2018. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2): 1099–1135.

Crowther, T. W.; Glick, H. B.; Covey, K. R.; Bettigole, C.; Maynard, D. S.; Thomas, S. M.; Smith, J. R.; Hintler, G.; Duguid, M. C.; Amatulli, G.; et al. 2015. Mapping tree density at a global scale. *Nature*, 525(7568): 201–205.

Gao, G.; Liu, Q.; and Wang, Y. 2020. Counting dense objects in remote sensing images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4137–4141. IEEE.

Hlavsa, M. C.; Aluko, S. K.; Miller, A. D.; Person, J.; Gerdes, M. E.; Lee, S.; Laco, J. P.; Hannapel, E. J.; and Hill, V. R. 2021. Outbreaks Associated with Treated Recreational Water — United States, 2015–2019.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–4.

Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M.; Bulatov, Y.; and McCord, B. 2018. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*.

Lee, J.; Brooks, N. R.; Tajwar, F.; Burke, M.; Ermon, S.; Lobell, D. B.; Biswas, D.; and Luby, S. P. 2021. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118(17).

Li, B.; Gao, S.; Liang, Y.; Kang, Y.; Prestby, T.; Gao, Y.; and Xiao, R. 2020. Estimation of regional economic development indicator from transportation network analytics. *Scientific reports*, 10(1): 1–15.

Litman, T.; and Laube, F. 2002. Automobile dependency and economic development. *Victoria Transport Policy Institute, Canada*.

Mubin, N. A.; Nadarajoo, E.; Shafri, H. Z. M.; and Hamedianfar, A. 2019. Young and mature oil palm tree detection and counting using convolutional neural network deep learning method. *International Journal of Remote Sensing*, 40(19): 7500–7515.

Owen, A. B. 2013. Monte Carlo theory, methods and examples. *Preprint*, 5–6.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271.

Salamí, E.; Gallardo, A.; Skorobogatov, G.; and Barrado, C. 2019. On-the-fly olive tree counting using a UAS and cloud services. *Remote Sensing*, 11(3): 316.

Sheehan, E.; Meng, C.; Tan, M.; Uzkent, B.; Jean, N.; Burke, M.; Lobell, D.; and Ermon, S. 2019. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2698–2706.

Sirko, W.; Kashubin, S.; Ritter, M.; Annkah, A.; Bouchareb, Y. S. E.; Dauphin, Y.; Keysers, D.; Neumann, M.; Cissé, M.; and Quinn, J. 2021. Continental-Scale Building Detection from High Resolution Satellite Imagery. *ArXiv*, abs/2107.12283.

Solow, R. M. 1992. Sustainability: An Economist's Perspective. *National Geographic Research and Exploration*, 8: 10–21.

Taylor, L. H. 2020. Fascinating Facts About Pools, Spas, Swimming and Safety. https://www.liveabout.com/facts-about-pools-spas-swimming-safety-2737127. Accessed: 2021-09-06.

Uzkent, B.; and Ermon, S. 2020. Learning When and Where to Zoom With Deep Reinforcement Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Uzkent, B.; Sheehan, E.; Meng, C.; Tang, Z.; Burke, M.; Lobell, D.; and Ermon, S. 2019. Learning to interpret satellite images using wikipedia. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Uzkent, B.; Yeh, C.; and Ermon, S. 2020. Efficient object detection in large images using deep reinforcement learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1824–1833.

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.

Yi, Z. N.; Frederick, H.; Mendoza, R. L.; Avery, R.; and Goodman, L. 2021a. AI Mapping Risks to Wildlife in Tanzania — Development Seed.

Yi, Z. N.; Zurutuza, N.; Kim, D.-H.; Mendoza, R. L.; Morrissey, M.; Daniels, C.; Ingalls, N.; Farias, J.; Tenorio, K.; Serrano, P.; and Anwar, S. 2021b. Scaling AI to map every school on the planet — Development Seed.

Yu, J.; Wang, Z.; Majumdar, A.; and Rajagopal, R. 2018. DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States. *Joule*, 2(12): 2605–2617.