Efficient Poverty Mapping from High Resolution Remote Sensing Images

Kumar Ayush[*1] **Burak Uzkent**[*1] **Kumar Tanmay**[3] **Marshall Burke**[2] **David Lobell**[2] **Stefano Ermon**[1]
[1] Department of Computer Science, Stanford University
[2] Department of Earth Science, Stanford University
[3] Department of Electrical Engineering, IIT Kharagpur
kayush@cs.stanford.edu, buzkent@cs.stanford.edu, kr.tanmay147@iitkgp.ac.in, mburke@stanford.edu,
dlobell@stanford.edu, ermon@cs.stanford.edu

Abstract

The combination of high-resolution satellite imagery and machine learning have proven useful in many sustainability-related tasks, including poverty prediction, infrastructure measurement, and forest monitoring. However, the accuracy afforded by high-resolution imagery comes at a cost, as such imagery is extremely expensive to purchase at scale. This creates a substantial hurdle to the efficient scaling and widespread adoption of high-resolution-based approaches. To reduce acquisition costs while maintaining accuracy, we propose a reinforcement learning approach in which free low-resolution imagery is used to dynamically identify where to acquire costly high-resolution images, prior to performing a deep learning task on the high-resolution images. We apply this approach to the task of poverty prediction in Uganda, building on an earlier approach that used object detection to count objects and use these counts to predict poverty. Our approach exceeds previous performance benchmarks on this task while using 80% fewer high-resolution images, and could be useful in many domains that require high-resolution imagery.

Introduction

When combined with machine learning, high-resolution satellite imagery has proven broadly useful for a range of sustainability-related tasks, from poverty prediction (Jean et al. 2016; Ayush et al. 2020; Sheehan et al. 2019; Blumenstock, Cadamuro, and On 2015; Yeh et al. 2020) to infrastructure measurement (Cadamuro, Muhebwa, and Taneja 2018) to forest and water quality monitoring (Fisher et al. 2018) to the mapping of informal settlements (Mahabir et al. 2018). Compared to coarser (10-30m) publicly-available imagery (Drusch et al. 2012), high-resolution ($< 1m$) imagery has proven particularly useful for these tasks because it is often able to resolve specific objects or features that are undetectable in coarser imagery.

When combined with machine learning, high-resolution satellite imagery has proven broadly useful for object detection (Lam et al. 2018), object tracking (Uzkent, Rangnekar, and Hoffman 2018, 2017), cloud removal (Sarukkai et al. 2020), and a range of sustainability-related tasks, from poverty prediction (Jean et al. 2016; Ayush et al. 2020; Sheehan et al. 2019; Blumenstock, Cadamuro, and On 2015; Yeh et al. 2020) to infrastructure measurement (Cadamuro, Muhebwa, and Taneja 2018). Compared to coarser (10-30m) publicly-available imagery (Drusch et al. 2012), high-resolution ($< 1m$) imagery has proven particularly useful for these tasks because it is often able to resolve specific objects or features that are undetectable in coarser imagery.

For example, recent work demonstrated an approach for predicting local-level consumption expenditure using object detection on high-resolution daytime satellite imagery (Ayush et al. 2020), showing how this approach can yield interpretable predictions and also outperform previous benchmarks that rely on lower-resolution, publicly-available satellite imagery (Drusch et al. 2012). This additional information, however, typically comes at a cost, as high-resolution satellite imagery must be purchased from private providers. Additionally, processing high-resolution images is computationally more expensive than the coarser resolution ones (Uzkent et al. 2019; Zhu et al. 2016; Meng et al. 2017; Lampert, Blaschko, and Hofmann 2008; Wojek et al. 2008; Redmon and Farhadi 2017; Gao et al. 2018). Given these costs, deploying these models at scale using high-resolution imagery quickly becomes cost-prohibitive for most organizations and research teams, inhibiting the broader development and deployment of machine-learning based tools and insights based on these data.

To address this problem, we propose a reinforcement learning approach that uses coarse, freely-available public imagery to dynamically identify where to acquire costly high-resolution images, prior to conducting an object detection task. This concept leverages publicly available Sentinel-2 (Drusch et al. 2012) images (10-30m) to sample smaller amount of high-resolution images (<1m). Our framework is inspired from the recent studies in computer vision literature that perform conditional inference to reduce computational complexity of convolutional networks in test time (Uzkent and Ermon 2020; Wu et al. 2018).

We apply our approach to the domain of poverty prediction, and show how our approach can substantially reduce the cost of previous methods that used deep learning on high-resolution images to predict poverty (Ayush et al. 2020) while maintaining or even improving their accuracy. In our study country of Uganda, we show how our approach can re-

[*]Equal Contribution

duce the number of high-resolution images needed by 80%, in turn reducing the cost of making a country-wide poverty map using this approach by an estimate $2.9 million.

Poverty Mapping from Remote Sensing Imagery

Poverty is typically measured using consumption expenditure, the value of all the goods and services consumed by a household in a given period. A household or individual is said to be poverty stricken if their measured consumption expenditure falls below a defined threshold (currently $1.90 per capita per day). We focus on this consumption expenditure as our outcome of interest, using "poverty" as shorthand for "consumption expenditure" throughout the paper. While typical household surveys measure consumption expenditure at the household level, publicly available data typically only release geo-coordinate information at the "cluster" level – which is a village in rural areas and a neighborhood in urban areas. Efforts to predict poverty have thus focused on predicting at the cluster level (or more aggregated levels) (Ayush et al. 2020).

Ayush et al. (2020) demonstrated state-of-the-art results for predicting village-level poverty using high-resolution satellite imagery, and showed how such predictions could be made with an interpretable model. In particular, they trained an object detector to obtain classwise object counts (buildings, trucks, passenger vehicles, railway vehicles, etc.) in high-resolution images, and then used these counts in a regression model to predict poverty. Not only were these categorical features predictive of poverty, but their counts had clear and intuitive relationships with the outcome of interest. The cost of this accuracy and interpretability was the high-resolution imagery, which typically must be purchased for $10-20 per km^2 from private providers.

Problem statement. Let $\{(\mathcal{H}_i, \mathcal{L}_i, y_i, c_i)\}_{i=1}^{N}$ be a set of N villages surveyed, where $c_i = (c_i^{lat}, c_i^{lon})$ is the latitude and longitude coordinates for cluster i, and $y_i \in \mathbb{R}$is the corresponding average poverty index for a particular year. For each cluster i, we can acquire both high-resolution (at a cost) and low-resolution (free of charge) satellite imagery corresponding to the survey year, $\mathcal{H}_i \in \mathbb{R}^{W \times H \times B}$, a $W \times H$ image with B channels, and $\mathcal{L}_i \in \mathbb{R}^{W/D \times H/D \times B}$, a $W/D \times H/D$ image with B channels. Here D represents a scalar to show the resolution difference between low-resolution and high-resolution images. Our goal is to learn (1) a regressor f_r to predict the poverty index y_i using $\mathcal{L}_i$ and parts of $\mathcal{H}_i$ (the informative regions) selected by (2) an adaptive data acquisition scheme based on $\mathcal{L}_i$. This adaptive data acquisition scheme is optimized to minimize cost (which depends on the number of selected regions) while maximizing the accuracy of f_r.

Dataset

Socio-economic Data. Our ground truth dataset consists of data on consumption expenditure (poverty) from Living Standards Measurement Study (LSMS) survey conducted in Uganda by the Uganda Bureau of Statistics between 2011 and 2012 (UBOS 2012). The survey consists of data from 2,716 households in Uganda, grouped into unique locations called clusters. The latitude and longitude, $c_i = (c_i^{lat}, c_i^{long})$, of a cluster $i = \{1, 2, \ldots, N\}$ is given, with noise of up to 5 km added in each direction by the surveyers to protect respondent privacy. Individual household locations in each cluster i are also withheld to preserve anonymity. We have N=320 clusters in the survey which we use to test the method performance in terms of predicting the average poverty index, y_i, for a group i. For each c_i, the survey measures the poverty level by the per capital daily consumption in dollars which we refer to as the "LSMS poverty score" for simplicity like (Ayush et al. 2020). Fig. 1 (bottom left corner) visualizes the surveyed locations on the map along with their corresponding LSMS poverty scores, revealing that a high percentage of surveyed locations have relatively low consumption expenditure values.

Satellite Imagery. We acquire both high-resolution and low-resolution satellite imagery for Uganda. The high-resolution satellite imagery, $\mathcal{H}_i$, corresponding to cluster c_i (roughly, a village or neighborhood) is represented by T=34×34=1156 images of 1000×1000 pixels each with 3 channels, arranged in a 34×34 square grid. This corresponds to a 10km×10km spatial neighborhood centered at c_i. A large neighborhood is considered to deal with up-to 5km of random noise in the cluster coordinates that has been added by the survey organization to protect respondent privacy. These high-resolution images come from DigitalGlobe satellites with 3 bands (RGB) and 30cm pixel resolution. Formally, we represent all the high-resolution images corresponding to c_i as a sequence of T tiles as $\mathcal{H}_i = \{H_i^j\}_{j=1}^{T}$. We acquire all the high-resolution tiles representing a cluster for comparison with (Ayush et al. 2020). However, in real-word scenario our method requires only a small fraction of HR tiles in test time unlike (Ayush et al. 2020) that acquires HR tiles exhaustively.

We also acquire low-resolution satellite imagery, $\mathcal{L}_i$, corresponding to cluster c_i and represented by a single image of 1014×1014 pixels with 3 channels. These images come from Sentinel-2 with 3 bands (RGB) and 10m pixel resolution and are freely available to the public. Each image corresponds to the same 10km×10km spatial neighborhood centered at c_i, however the resolution is much lower – each Sentinel-2 pixel corresponds to roughly 1000 pixels from the high-resolution imagery. Because of this low-resolution, it is not possible to perform fine-grained object detection just using these images. Fig. 1 illustrates an example cluster from Uganda.

Fine-grained Object Detection on High-Resolution Satellite Imagery

Similar to (Ayush et al. 2020), we use an intermediate object detection phase to obtain categorical features (classwise object counts) from high-resolution tiles of a cluster. Due to lack of object annotations for satellite images from Uganda, we use the same transfer learning strategy as in (Ayush et al. 2020) by training an object detector (YOLOv3 (Redmon and Farhadi 2018)) on xView (Lam et al. 2018), one of the largest and most diverse publicly available overhead

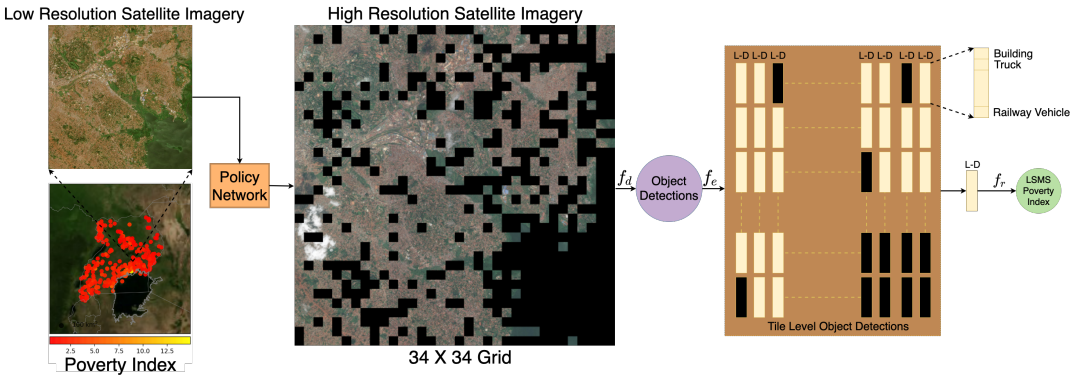


Figure 1: Schematic overview of the proposed approach. The Policy Network uses cheaply available Sentinel-2 low-resolution image representing a cluster to output a set of actions representing unique 1000×1000 px high-resolution tiles in the 34×34 grid. Then object detection is performed on the sampled HR tiles (black regions represent dropped tiles) to obtain the corresponding class-wise object counts (L-dimensional vectors). Finally, the classwise object counts vectors corresponding to the acquired HR tiles are added element-wise to get the final feature vector representing the cluster. Our reinforcement learning approach dynamically identifies where to acquire high-resolution images, conditioned on cheap, low-resolution data, before performing object detection, whereas the previous work (Ayush et al. 2020) exhaustively uses all the HR tiles representing a cluster for poverty mapping, making their method expensive and less practical.

imagery datasets for object detection with 10 parent-level and 60 child-level classes. Earlier work (Ayush et al. 2020) studied both parent-level and child-level detectors and empirically find that not only the parent-level object detection features are better for poverty regression but at the same time are more suited for interpretability due to household level descriptions. Thus, we train YOLOv3 detector using parent-level classes (see x-axis labels of Fig. 3).

As described in previous section, each $\mathcal{H}_i$ representing a cluster is a set of T high-resolution images, $\{H_i^j\}_{j=1}^{T}$. To obtain a baseline model that uses all the high-resolution imagery available, we follow the protocol in (Ayush et al. 2020) and run the trained YOLOv3 object detector on each 1000×1000px tile (*i.e.* H_i^j) to get the correspoding set of object detections . Similar to (Ayush et al. 2020), we use these object detections to generate a L-dimensional vector, $\mathbf{v}_i^j \in \mathbb{R}^L$ (where L=10 is the number of object labels/classes), by counting the number of detected objects in each class. This class-wise object counts can be used in a regression model for poverty estimation (Ayush et al. 2020).

Ayush et al. (2020) exhaustively uses all T=1156 HR tiles of a cluster for poverty estimation. In contrast, we propose to use a method that adaptively selects informative regions for high-resolution acquisition conditioned on the publicly available, low-resolution data. Thus, we reduce the dependency on HR images that are expensive to acquire thereby reducing the costs of poverty prediction models that use HR images exhaustively (Ayush et al. 2020) making their method costly and less practical. We describe our solution in the next section.

Adaptive Tile Selection

Due to the large acquisition cost of HR images, it is non-trivial and expensive to deploy models based on HR imagery at scale. For this reason, we propose an efficient tile selection framework to capture relevant fine level information such as classwise object counts for downstream tasks. We represent the HR image covering a spatial cluster i centered at $c_i = (c_i^{lat}, c_i^{lon})$ as $\mathcal{H}_i \in R^{W \times H \times B}$ where W, H and B represent height width and number of bands. Additionally, we represent the LR image of the same spatial cluster i as $\mathcal{L}_i \in R^{W/D, H/D, B}$ where D represents a scalar for the number of pixels in width and height. For example, in the case of Sentinel-2 (10 m GSD), we have $D = 30$ times smaller number of pixels than the high-resolution DigitalGlobe images (0.3m GSD). With an adaptive approach, our task is to acquire only small subset of $\mathcal{H}_i$ conditionally on $\mathcal{L}_i$ while not hurting the performance in our downstream tasks that uses object counts from the cluster i. This adaptive method is formulated as a two-step episodic Markov Decision Process (MDP), similar to (Uzkent, Yeh, and Ermon 2020). In the first step, we adaptively sample HR tiles and in the second step, we run them through a pre-trained detector.

Task Definition. The first module of our framework finds HR tiles to sample/acquire, conditioned on the low spatial resolution image covering a cluster (which is always acquired). However, a cluster is represented by 34000×34000 px HR images. Directly learning actions with reinforcement learning on such a large area can be very challenging and unstable. For this reason, we decompose our task to many independent sub-tasks where each sub-task focuses on sampling the important parts of the corresponding area with HR images. Following this, we divide a cluster-level HR image $\mathcal{H}_i = (H_i^1, H_i^2, \ldots, H_i^T)$ into equal-size non-overlapping tiles, where T is the number of tiles. Similar to $\mathcal{H}_i$, we decompose $\mathcal{L}_i$ as $\mathcal{L}_i = (l_i^1, l_i^2, \ldots, l_i^T)$ where l_i^j represents the lower spatial resolution version (from Sentinel-2) of H_i^j. In this set up, we model $\mathcal{H}_i$ as a latent variable as it is not directly observed and it is inferred from the observation $\mathcal{L}_i$. We associate each tile, H_i^j, of $\mathcal{H}_i$ with an L-dimensional

classwise object counts feature represented as $\mathbf{v}_i^j$.

In a simple scenario, we can take a single binary action for each H_i^j whether to acquire it or not conditioned on l_i^j. However, we believe that choosing multiple actions representing different disjoint subtiles of tile H_i^j can help us avoid sampling areas of tile H_i^j where there are no objects of interest. For this reason, we divide tile H_i^j into S number of disjoint subtiles as $H_i^j = (h_i^{j,1}, h_i^{j,2}, ..., h_i^{j,S})$. We then define our task as learning a policy network conditioned on l_i^j to only choose HR sub-tiles from H_i^j where there is desirable number of objects characterized by a reward function. Once we learn the policy network, in test time we run it on each l_i^j of a cluster i to sample HR images and run them through detector to find out the cluster-level object counts.

1st Step of MDP. In the first step, the agent observes l_i^j and outputs a binary action array, $\mathbf{a}_i^j \in \{0,1\}^S$, where $a_i^{j,k} = 1$ represents acquisition of the HR version of the k-th subtile of H_i^j *i.e.* $h_i^{j,k}$. The subtile sampling policy, parameterized by θ_p, is formulated as $\pi(\mathbf{a}_i^j|l_i^j;\theta_p) = p(\mathbf{a}_i^j|l_i^j;\theta_p)$ where $\pi(l_i^j;\theta_p)$ is a function mapping the observed LR image to a probability distribution over subtile sampling actions $\mathbf{a}_i^j$.

2nd Step of MDP. In the second step, the agent runs the object detection on the selected HR subtiles. Conditioned on $\mathbf{a}_i^j$, it observes HR subtiles if necessary and produces $\hat{\mathbf{v}}_i^j$, a L-dimensional classwise object counts vector. We find the object counts with our adaptive framework using a pre-trained object detector f_d (parameterized by θ_d) as:

$$\hat{\mathbf{v}}_i^{j,k} = \begin{cases} f_d(h_i^{j,k}) & \text{if} \quad a_i^{j,k} = 1 \\ \mathbf{0} & \text{else} \end{cases} \tag{1}$$

Then, we compute the tile level object counts as $\hat{\mathbf{v}}_i^j = \sum_{k=1}^{S} \hat{\mathbf{v}}_i^{j,k}$. Finally, we define our overall cost function J as:

$$\max_{\theta_p} J(\theta_p, \theta_d) = \mathbb{E}_p[R(\mathbf{a}_i^j, \hat{\mathbf{v}}_i^j, \mathbf{v}_i^j)], \tag{2}$$

where the reward depends on $\mathbf{a}_i^j$, $\hat{\mathbf{v}}_i^j$, $\mathbf{v}_i^j$. Our goal is to learn the parameters θ_p given a pre-trained object detector θ_d to maximize the objective being a function of the reward function.

The Reward Function. The desired outcome from our adaptive strategy is to reduce the *image acquisition cost* drastically by sampling smaller subset of tiles. Taking this into account, we design a dual reward function that encourages dropping as many subtiles as possible while successfully approximating the classwise object counts. We define R as follows:

$$R = R_{acc}(\hat{\mathbf{v}}_i^j, \mathbf{v}_i^j) + R_{cost}(\mathbf{a}_i^j) \tag{3}$$

$$R_{acc} = -||\mathbf{v}_i^j - \hat{\mathbf{v}}_i^j||_1 \tag{4}$$

$$R_{cost} = \lambda(1 - ||\mathbf{a}_i^j||_1/S) \tag{5}$$

where R_{acc} is object counts approximation accuracy and R_{cost} represents the image acquisition cost with λ as its coefficient. The R_{acc} term encourages acquiring a subtile when the counts difference between the object counts from fixed HR subtile sampling policy and the adaptive policy is positive. We increase the reward *linearly* with the smaller number of acquired subtiles for the cost component.

Modeling and Optimization of the Policy Network

In the previous section, in high level we formulated the task of efficient HR subtile selection as a two step episodic MDP. In this section, we model how to learn the policy distribution for subtile sampling.

Modeling the Policy Network. In this study, we have $T = 1156$ number of tiles as we have a 34×34 grid of images. In this case, each grid consists of 2000×2000 pixels. As mentioned in the previous section, we divide each tile into S=4 subtiles of 1000×1000 pixels each (higher values of S led to unstable training with higher variance and less sparse selections). In this study, similar to (Uzkent, Yeh, and Ermon 2020) we model the action likelihood function of the policy network, f_p, using the product of bernoulli distributions as:

$$\pi(\mathbf{a}_i^j|l_i^j;\theta_p) = \prod_{k=1}^{S} (s_i^{j,k})^{a_i^{j,k}} (1 - s_i^{j,k})^{(1-a_i^{j,k})} \tag{6}$$

$$s_i^j = f_p(l_i^j;\theta_p) \tag{7}$$

We use a sigmoid function to transform logits to probabilistic values, $s_i^{j,k} \in [0,1]$.

Optimization of the Policy Network. The previously defined objective function as shown in Eq. 2 is not differentiable w.r.t the policy network parameters, θ_p, because acquistion actions are discrete. To overcome this, we train using Policy Gradient (Sutton and Barto 2018). Our final objective function as shown below includes the reward function as well as action likelihood distribution which can be differentiated w.r.t θ_p.

$$\nabla_{\theta_p} J = \mathbb{E}\left[R(\mathbf{a}_i^j, \hat{\mathbf{v}}_i^j, \mathbf{v}_i^j) \nabla_{\theta_p} \log \pi_{\theta_p}(\mathbf{a}_i^j|l_i^j)\right], \tag{8}$$

Our objective function relies on mini-batch Monte-Carlo sampling to approximate the expectation. Especially, in scenarios where we can not afford large mini-batches, we can have highly oscillating expectations which results in large variance. As this can de-stabilize the optimization, we use the self-critical baseline (Rennie et al. 2017), A, to reduce the variance.

$$\nabla_{\theta_p} J = \mathbb{E}\left[A \sum_{k=1}^{S} \nabla_{\theta_p} \log(s_i^{j,k}\mathbf{a}_i^{j,k} + (1 - s_i^{j,k})(1 - \mathbf{a}_i^{j,k}))\right] \tag{9}$$

$$A(\mathbf{a}_i^j, \mathbf{a}_i'^j) = R(\mathbf{a}_i^j, \hat{\mathbf{v}}_i^j, \mathbf{v}_i^j) - R(\mathbf{a}_i'^j, \hat{\mathbf{v}}_i'^j, \mathbf{v}_i^j) \tag{10}$$

where $\mathbf{a}_i'^j$ represents the baseline action vector. To get $\mathbf{a}_i'^j$, we use the most likely action vector proposed by the policy network: *i.e.*, $a_i'^{j,k} = 1$ if $s_i^{j,k} > 0.5$ and $a_i'^{j,k} = 0$ otherwise. Finally, in this study we use temperature scaling (Sutton and Barto 2018) to adjust exploration/exploitation tradeoff during optimization time as

$$s_i^{j,k} = \alpha s_i^{j,k} + (1 - \alpha)(1 - s_i^{j,k}). \tag{11}$$

Setting α to a large value results in sampling from the learned policy whereas the small values lead to sampling from random policy. See appendix for the pseudocode and implementation details.

Experiments

Training and Testing the Policy Network on xView

Our goal is to learn policies to reduce the dependency on HR images in approximating object counts in a geocluster while successfully predicting the downstream index (poverty prediction). Since our downstream dataset (Uganda) does not contain object bounding boxes, it is not possible to assess how well we approximate true object counts. To achieve this, we train our policy network on the xView dataset where our object detector is trained on. We use 2000×2000 px images and their corresponding 224×224 px LR images to train the policy network on each point. As proposed earlier, the action space has 4 units representing the top left, top right, bottom left, and bottom right part (1000×1000 px) of the full area. The detector is only run on the part chosen by the policy network. We train the policy network on 1249 points and test it on 200 points and show the results in Table 1.

Our policy network uses 42.3% HR images while approximating the fixed approach in mean Average Precision (mAP) and mean Average Recall (mAR) metrics (Redmon and Farhadi 2018). This results indicate that the policy network learns to successfully choose regions where there are objects of interest and eliminate the regions with no objects of interest. See the Appendix for more details.

	mAP	mAR	HR	Run-time
No Dropping	24.3%	42.5%	100.0%	2890 ms
RL Method	26.3%	41.1%	42.3%	1510 ms

Table 1: Results on the xView test set.

Testing the Policy Network on Poverty Prediction

Previously, we trained and tested the policy network to quantify how well we approximate the true object counts. In this section, we train and test the policy network on Uganda dataset where we have only cluster-level poverty labels.

Poverty Estimation. Previous work (Ayush et al. 2020) exhaustively performed object detection on all the HR tiles representing a cluster i to obtain T L-dimensional vectors, $\mathbf{v}_i = \{\mathbf{v}_i^j\}_{j=1}^T$, which are then aggregated into a single L-dimensional categorical feature vector, $\mathbf{m}_i$, by summing over the tiles *i.e.* $\mathbf{m}_i = \sum_{j=1}^T \mathbf{v}_i^j$. This was subsequently used in a regression model to predict poverty score for cluster i. Using our adaptive method, we obtain $\hat{\mathbf{m}}_i = \sum_{j=1}^T \hat{\mathbf{v}}_i^j$, which is an approximate classwise counts vector for cluster i. Following (Ayush et al. 2020), we consider Gradient Boosting Decision Trees as the regression model to estimate the poverty index, y_i, given the cluster level categorical feature vector (classwise object counts), $\mathbf{m}_i$ or $\hat{\mathbf{m}}_i$.

Training and Evaluation. We have N=320 clusters in the survey. We divide the dataset into a 80%-20% train-test split. We train a GBDT model using object counts features ($\mathbf{m}_i$) based on all HR tiles of the clusters in the training set. We use the clusters in the training set to train the policy network for adaptive tile selection. The trained policy network is then used to acquire informative HR tiles for each test cluster *i.e* for a test cluster i, the policy network selects HR tiles (subsequently used to obtain $\hat{\mathbf{m}}_i$) conditioned on low-resolution input representing the cluster. The obtained $\hat{\mathbf{m}}_i$ is then passed through the trained GBDT model to get the poverty score y_i. See appendix for more details. To evaluate the models, we use Pearson's r^2 to quantify the model performance. Invariance under seperate changes in scale between two variables allows Pearson's r^2 to provide insights into the ability of the model at distinguishing poverty levels. We also report mean squared error (MSE) and Explained Variance (Rosenthal and Rosenthal 2011). Explained variance measures the discrepancy between a model and actual data. Higher explained variance indicates a stronger strength of association thus meaning better predictions.

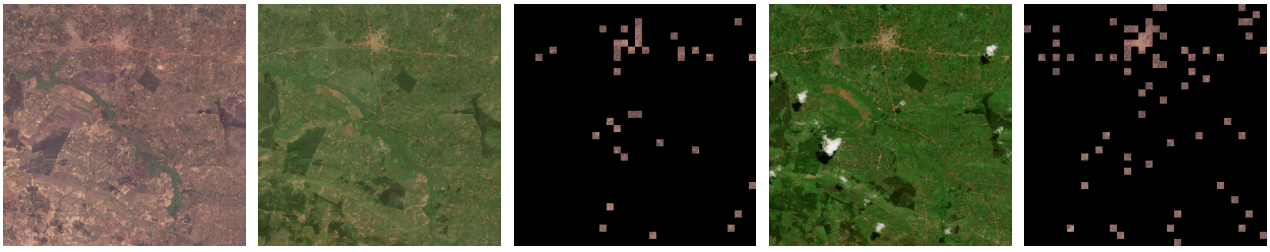
Baselines and State-of-the-Art Models. We compare our method with the following: (a) *No Patch Dropping*, where we simply use all the HR tiles in $\mathcal{H}_i$ to get the classwise object counts features (same as (Ayush et al. 2020)), (b) *Fixed Policy-X* samples $X\%$ HR tiles from the center of a cluster, (c) *Random Policy-X* samples $X\%$ HR tiles randomly from a cluster, (d) *Stochastic Policy-X* samples $X\%$ HR tiles where the survival likelihood of a tile decays w.r.t the euclidean distance from the cluster center, (f) *Green Tiles*, where we compute the average green channel value for a low-res tile and select bottom K tiles for HR acquisition with least average green channel value, where K is the number of tiles selected by the policy network for a particular cluster, (g) *Counts Prediction*, where we train a CNN (Resnet-50 backbone) to regress object counts given low-res tile as input. We find that the object counts in a tile vary from 0-500. Instead of regressing directly on raw object counts, we create 100 bins such that a tile with counts between 5i-(5i+1) has label 5i+2.5 (e.g. a tile with counts 0-5 has label 2.5, 5-10 has label 7.5 and so on). We use this network to select top K HR tiles based on predicted object counts, (h) *Settlement Layer*, where we select HR tiles based on their population density. We used the HR settlement layer maps[1] and selected top K tiles based on population density, and (e) *Nightlights*, where we use Nightlight Images (48×48 px) representing the clusters in Uganda and sample only those HR tiles which have non-zero nighttime light intensities.

Additionally, since Sentinel-2 imagery is freely available, we perform a comparative analysis of the effect of season on the ability of the policy network at approximating classwise object counts. We thus acquired two sets of low-resolution imagery, one from dry-season (Dec - Feb) in Uganda and other from wet season (March-May, Sept-Nov) corresponding to the survey year. Seasonality is likely highly relevant in our rural setting, where crops are grown during the wet sea-

[1] https://research.fb.com/downloads/high-resolution-settlement-layer-hrsl/

	No Dropping	Fixed-18	Random-25	Stochastic-25	Green	Counts Pred.	Sett. Layer	Nightlights	Ours (Dry sea.)	Ours (Wet sea.)
r^2	0.53	0.43	0.34	0.26	0.33	0.49	0.45	0.45	0.51 ± 0.01	$\mathbf{0.61 \pm 0.01}$
MSE	1.86	2.20	2.67	3.13	2.56	1.91	2.16	2.17	1.89 ± 0.02	$\mathbf{1.46 \pm 0.02}$
Explained Variance	0.54	0.43	0.33	0.27	0.36	0.48	0.46	0.45	0.50 ± 0.01	$\mathbf{0.63 \pm 0.02}$
HR Acquisition.	1.0	0.18	0.25	0.25	0.19	0.19	0.19	0.12	0.19	**0.19**

Table 2: LSMS poverty score prediction results in Pearson's r^2 (and two other metrics) for various methods. *HR Acquisition* represents the fraction of HR tiles acquired. We report the mean and std of our RL model across 7 runs with different seeds.



(a) (b) (c) (d) (e)

Figure 2: (a) High-Resolution Satellite Imagery representing a cluster. (b) Sentinel-2 Imagery of the cluster from dry season. (c) Corresponding HR acquisitions when dry-season imagery is input to the Policy Network. (d) Sentinel-2 Imagery of the cluster from wet season. (e) Corresponding HR acquisitions when wet-season imagery is input to the Policy Network.

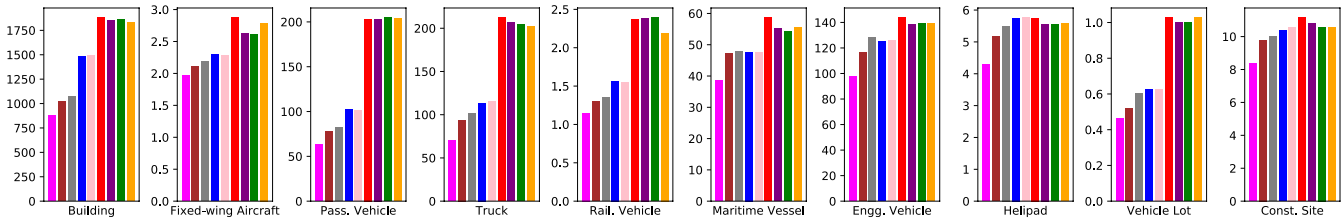
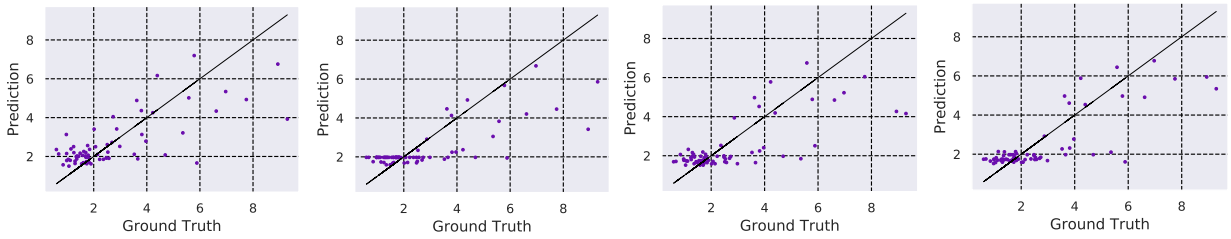


Figure 3: Number of objects missed on average across clusters for each class. Colored bars in each subplot from left-right are: Ours (wet season), Ours (dry season), Counts Pred., Nightlight, Settlement, Fixed-18, Random-25, Green Tiles, Stochastic-25.



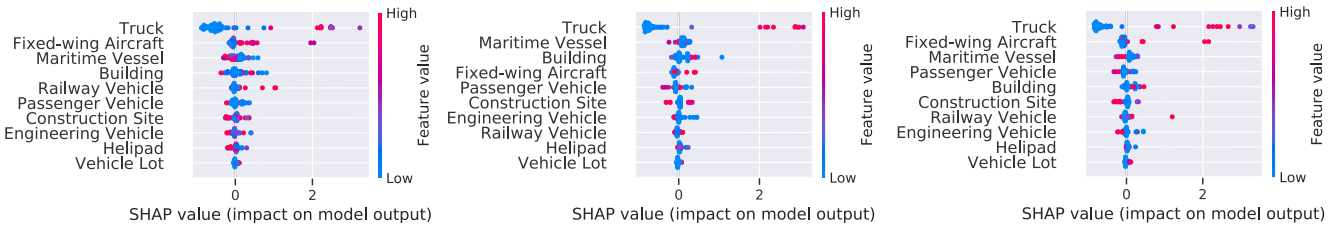
(a) No Dropping (b) Nightlights (c) Ours (Dry Season) (d) Ours (Wet Season)

Figure 4: LSMS poverty score regression results of GBDT.

son and much related market activity is highly seasonal. We hypothesize that greenery in low-resolution imagery during wet season will better indicate which patches might contain useful economic information.

Quantitative Analysis. Fig. 3 compares the ability of various methods at approximating the classwise object counts. It shows the number of objects missed on an average across clusters for each parent class, where we can see that our method (using wet season imagery) can better approximate the "true object counts" (we use object detector predictions on all the HR tiles as a proxy for true values) compared to baselines and our method (using dry season imagery). Table 2 shows the results of poverty prediction in Uganda. Our model (wet season) achieves **0.61** r^2 and substantially outperforms the published state-of-the-art results (Ayush et al. 2020) (**0.53** r^2) while using around **80%** fewer HR images.

It is interesting that we can outperform *No Dropping* method when sampling only 20% of HR tiles. Qualitatively, we observed that it is due to false positives proposed by the object detector on the tiles with no true objects of interest in it. Unfortunately, since we do not have ground truth bounding boxes for Uganda, we can not quantify it. However,



(a) No Dropping (b) Ours (Dry Season) (c) Ours (Wet Season)

Figure 5: Summary of the effects of all features using SHAP, showing the distribution of the impacts each feature has on the model output. Color represents the feature value (red high, blue low).

our experiments on xView (Table 1) show that our approach achieves higher AP than the *No Dropping* approach, suggesting our approach is able to remove false positives.

In comparison to the baselines relying on external data layers such as settlement and nightlights, our method achieves around 0.16 higher r^2. This is because such maps assume that objects are located in the tiles with large nightlight intensity or settlement index, however, some objects, i.e trucks, passenger vehicles etc., do not necessarily exist in these areas. Additionally, our approach outperforms the counts prediction model by 0.12 in r^2. This might be because the counts predictor is trained to directly regress very noisy object counts thus making it a difficult task.

Next, a scatter plot of GBDT LSMS poverty score predictions v.s. ground truth is shown in Fig. 4. It can be seen that the GBDT model can maintain explainability of a large fraction of the variance based on object counts identified from the sampled HR tiles using our method, compared to (Ayush et al. 2020) that exhaustively uses all HR tiles.

Performance/Sampling Trade-off. We analyze the trade-off between accuracy (regression performance) and HR sampling rate controlled by the hyperparameter λ in the reward Eq. 5. We intentionally change λ to quantify the effect on the policy network. As seen in Fig. 6, the policy network samples less HR tiles (a 0.09 fraction) when we increase λ to 2.0 and the r^2 goes down to 0.48. On the other hand, when we set λ to 1.0, we get optimal results in terms of r^2, while acquiring only a 0.18 fraction of HR imagery.

Cost saving. Current pricing for high-resolution (30cm) RGB imagery is 10-20$ per km^2. Given that Uganda is 240k km^2 in land area, creating a poverty map using our method would save roughly $2.9 million if imagery costs $15 per km^2. This represents a potentially large cost saving if our approach is scaled at country or continent scale.

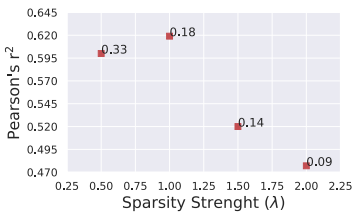


Figure 6: Trade-off between Pearson's r^2 and coefficent of image acquisition cost (λ). Text accompanying the points represents HR acquisition fraction.

Analysis based on Season. Presence of greenery during wet season allows the policy network to better identify the informative regions containing objects, compared to when trained with dry season Sentinel-2 imagery as input. Fig. 2 presents an example cluster, where it is seen that training the policy network using wet season imagery better assists the network at sampling informative tiles (see Appendix).

Impact on Interpretability. An important contribution of (Ayush et al. 2020) was to introduce model interpretability allowing successful application of such methods in many policy domains. They use Tree SHAP (Tree SHapley Additive exPlanations) (Lundberg and Lee 2017), a game theoretic approach to explain the output of tree-based models, to explain the effect of individual features on poverty predictions. Here, we show that in addition to closely approximating the classwise object counts, our method retains the same findings for interpretability as that of (Ayush et al. 2020). Fig. 5 shows the plots of SHAP values of every feature for every cluster for three different methods. The features are sorted by the sum of SHAP value magnitudes over all samples. It can be seen that our method still maintains that *#Trucks* tends to have a higher impact on the model's output. We also observe that ordering of features in terms of SHAP values is fairly similar between the *No Dropping* approach (Ayush et al. 2020) and our method.

Conclusion

In this study, we increase the efficiency of recent methods of predicting consumption expenditure using object counts from high-resolution satellite images. To achieve this, we proposed a novel reinforcement learning setup to conditionally acquire high-resolution tiles. We designed a cost-aware reward function to reflect real-world constraints – i.e. budget and GPU availability – and then trained a policy network to approximate object counts in a given location as closely as possible given these constraints. We show that our approach reduces the number of high-resolution images needed by 80% while improving downstream poverty estimation performance relative to multiple other approaches, including a method that exhaustively uses all high-resolution images from a location. Future work includes application of our adaptive method to other sustainability-related computer vision tasks using high-resolution images at large scale.

Acknowledgements

This research was supported in part by Stanford's Data for Development Initiative and NSF grants #1651565 and #1733686.

Ethics Statement

A range of organizations, from governments to non-governmental organizations to private sector enterprise, depend on accurate local-level information of economic well-being of populations for their decision-making. Such information is expensive to collect using traditional ground-based survey operations, and as a result it rarely gets collected: (Yeh et al. 2020) calculate that for most countries in Africa, at least five years pass between nationally representative household livelihood surveys, and the number of villages covered in these surveys is incredibly small relative to the population size. The result is an environment of data scarcity in which governments and NGOs have difficulty identifying those most in need of assistance or measuring the impact of the assistance they do deliver.

Our approach offers an accurate, inexpensive, and scalable method for plugging this data gap. It uses only anonmyized public data in training, and these data are thought to be a true random and unbiased sample of the population (this is the explict goal of the LSMS survey team). Thus our approach should deliver unbiased estimates of local-level well-being, and does so without having to use any individually-identifying information.

One possible concern of our approach is that better information on where the poor are could lead to them being denied certain services due to their (now-known) low income levels. We believe that this is unlikely for multiple reasons. First, the poor are already severely underserved by private markets in much of the developing world, with (e.g.) access to bank loans among the poor typically in the low single digits (Suri and Jack 2016). For these populations, better data on their location and livelihoods is expected to improve rather than diminish access to private sector services. Similarly, due to lack of data, governments have difficulty targeting existing anti-poverty programs, meaning that resources that should be going to poorer populations sometimes flow to wealthier populations; for governments that do target, ground-based survey efforts regularly cost tens of millions of dollars (Banerjee, Niehaus, and Suri 2019). More accurate data on the location and poverty levels of populations should increase rather than decrease resources flowing to vulnerable populations.

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* .

Banerjee, A.; Niehaus, P.; and Suri, T. 2019. Universal basic income in the developing world. *Annual Review of Economics* .

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

Cadamuro, G.; Muhebwa, A.; and Taneja, J. 2018. Assigning a grade: Accurate measurement of road quality using satellite imagery. *arXiv preprint arXiv:1812.01699* .

Drusch, M.; Bello, U. D.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; Meygret, A.; Spoto, F.; Sy, O.; Marchese, F.; and Bargellini, P. 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* 120: 25 – 36. ISSN 0034-4257. doi:https://doi.org/10.1016/j.rse.2011.11.026. URL http://www.sciencedirect.com/science/article/pii/S0034425712000636.

Fisher, J. R.; Acosta, E. A.; Dennedy-Frank, P. J.; Kroeger, T.; and Boucher, T. M. 2018. Impact of satellite imagery spatial resolution on land use classification accuracy and modeled water quality. *Remote Sensing in Ecology and Conservation* 4(2): 137–149.

Gao, M.; Yu, R.; Li, A.; Morariu, V. I.; and Davis, L. S. 2018. Dynamic Zoom-in Network for Fast Object Detection in Large Images. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6926–6935. Salt Lake City, UT, USA: IEEE. ISBN 978-1-5386-6420-9. doi:10.1109/CVPR.2018.00724. URL https://ieeexplore.ieee.org/document/8578822/.

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): 790–794.

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* .

Lampert, C. H.; Blaschko, M. B.; and Hofmann, T. 2008. Beyond sliding windows: Object localization by efficient subwindow search. In *2008 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mahabir, R.; Croitoru, A.; Crooks, A. T.; Agouris, P.; and Stefanidis, A. 2018. A critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums: Trends, challenges and emerging opportunities. *Urban Science* 2(1): 8.

Meng, Z.; Fan, X.; Chen, X.; Chen, M.; and Tong, Y. 2017. Detecting small signs from large images. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, 217–224. IEEE.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525. Honolulu, HI, USA: IEEE. ISBN 9781538604571. doi:10.1109/CVPR.2017.690. URL https://ieeexplore.ieee.org/abstract/document/8100173.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* .

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7008–7024.

Rosenthal, G.; and Rosenthal, J. A. 2011. *Statistics and data interpretation for social work*. Springer publishing company.

Sarukkai, V.; Jain, A.; Uzkent, B.; and Ermon, S. 2020. Cloud removal from satellite images using spatiotemporal generator networks. In *The IEEE Winter Conference on Applications of Computer Vision*, 1796–1805.

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.

Suri, T.; and Jack, W. 2016. The long-run poverty and gender impacts of mobile money. *Science* 354(6317): 1288–1292.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

UBOS, U. B. o. S. 2012. Uganda National Panel Survey 2011/2012. *Uganda* .

Uzkent, B.; and Ermon, S. 2020. Learning when and where to zoom with deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12345–12354.

Uzkent, B.; Rangnekar, A.; and Hoffman, M. 2017. Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 39–48.

Uzkent, B.; Rangnekar, A.; and Hoffman, M. J. 2018. Tracking in aerial hyperspectral videos using deep kernelized correlation filters. *IEEE Transactions on Geoscience and Remote Sensing* 57(1): 449–461.

Uzkent, B.; Sheehan, E.; Meng, C.; Tang, Z.; Burke, M.; Lobell, D. B.; and Ermon, S. 2019. Learning to Interpret Satellite Images using Wikipedia. In *IJCAI*, 3620–3626.

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

Wojek, C.; Dorkó, G.; Schulz, A.; and Schiele, B. 2008. Sliding-windows for rapid object class localization: A parallel technique. In *Joint Pattern Recognition Symposium*, 71–81. Springer.

Wu, Z.; Nagarajan, T.; Kumar, A.; Rennie, S.; Davis, L. S.; Grauman, K.; and Feris, R. 2018. Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8817–8826.

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(1): 1–11.

Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; and Hu, S. 2016. Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2110–2118.