International Journal of Geo-Information

MDPI

Article

Toward Model-Generated Household Listing in Low- and Middle-Income Countries Using Deep Learning

Robert Chew *, Kasey Jones, Jennifer Unangst, James Cajka, Justine Allpress, Safaa Amer and Karol Krotki

RTI International, Research Triangle Park, NC 27709, USA; krjones@rti.org (K.J.); junangst@rti.org (J.U.); jcajka@rti.org (J.C.); jla@rti.org (J.A.); samer@rti.org (S.A.); kkrotki@rti.org (K.K.)

* Correspondence: rchew@rti.org; Tel.: +1-919-541-5823

Received: 19 September 2018; Accepted: 14 November 2018; Published: 16 November 2018

Abstract: While governments, researchers, and NGOs are exploring ways to leverage big data sources for sustainable development, household surveys are still a critical source of information for dozens of the 232 indicators for the Sustainable Development Goals (SDGs) in low- and middle-income countries (LMICs). Though some countries' statistical agencies maintain databases of persons or households for sampling, conducting household surveys in LMICs is complicated due to incomplete, outdated, or inaccurate sampling frames. As a means to develop or update household listings in LMICs, this paper explores the use of machine learning models to detect and enumerate building structures directly from satellite imagery in the Kaduna state of Nigeria. Specifically, an object detection model was used to identify and locate buildings in satellite images. In the test set, the model attained a mean average precision (mAP) of 0.48 for detecting structures, with relatively higher values in areas with lower building density (mAP = 0.65). Furthermore, when model predictions were compared against recent household listings from fieldwork in Nigeria, the predictions showed high correlation with household coverage (Pearson = 0.70; Spearman = 0.81). With the need to produce comparable, scalable SDG indicators, this case study explores the feasibility and challenges of using object detection models to help develop timely enumerated household lists in LMICs.

Keywords: sustainable development goals (SDGs); machine learning; object detection; household enumeration; survey statistics; remote sensing

1. Introduction

In September 2015, the United Nations (UN) General Assembly created the Sustainable Development Goals (SDGs), a list of 17 shared objectives to eradicate poverty, protect the planet, and ensure global prosperity for all [1]. To monitor progress toward the SDGs, the Interagency and Expert Group on SDG Indicators (IAEG-SDGs) developed a global indicator framework consisting of 232 specific statistical measures that member states could adopt and extend [2]. While comprehensive, the SDGs have been criticized for their sprawling scope and the high expected cost of implementing and monitoring the various indicators [3,4]. As of June 2017, even wealthy countries belonging to the Organization for Economic Cooperation and Development (OECD) only have the capacity to evaluate 57% of all the SDGs targets [5]. Without support, it will be particularly difficult for LMICs to monitor these indicators and measure progress toward achieving the SDGs.

While governments, researchers, and NGOs are exploring ways that big data sources may reduce the burden of developing and monitoring SDG indicators, household surveys are still a critical source of information. Roughly one-third of the SDG indicators can currently be derived from existing household surveys, and up to two-thirds could be covered with further enhancements to these programs [6]. Furthermore, surveys are one of the only data sources capable of systematically collecting the desired

level of information that can be "disaggregated by income, gender, age, race, ethnicity, migratory status, disability, geographic location and other characteristics relevant in national contexts (SDG Target 17.18)" [2]. Though bias may be introduced into survey data at many points throughout the design, data collection, and data processing phases of a study [7], survey researchers have designed methods to recognize and address risks of various error components [8,9].

While household surveys provide high-quality estimates to support the analysis of SDG indicators, the scope and frequency of these programs is limited by their cost [10]. This is, in part, due to the extensiveness of the operations required to select a probability-based sample of households. In probability-based sampling, a sampling frame must exist so that each member of the population has a known probability of being selected. The quality of this list is crucial, as it determines the degree to which the observed sample represents the intended population. Some countries' statistical agencies are able to maintain comprehensive and up-to-date databases of persons or households for sampling, but in many cases, including low- and middle-income countries (LMICs), it is necessary to create a complete listing at the time of the survey. For example, the Demographic and Health Survey (DHS) is a large-scale study that currently captures data to support up to 30 SDG indicators across nearly 90 developing countries [11]. While impressive in its scope, the standard approach of constructing a household listing for the DHS is cumbersome. While existing census data for each country typically provide a list of logistically manageable geographic areas for a first stage of sampling (e.g., counties or districts), field staff are typically required to visit sampled areas on foot to roster households. The practice of enumerating households is not only expensive, but also potential dangerous; researchers have noted concerns of robbery or violence when sending field staff into high-risk areas [12,13], especially for listings that require field staff to spend the majority of time surveying from the streets instead of inside respondents' dwellings [14].

In this paper, we focus on a method to reduce data collection costs and timeline for household surveys in LMICs: constructing the household listings required for probability-based samples. We explore a streamlined approach to obtaining household listings using machine learning models to detect and enumerate settlement units directly from satellite imagery. Specifically, we use the Kaduna state in Nigeria as a case study for applying an object detection model to identify and locate buildings from satellite images. These methods may reduce the level of effort required for probability-based household surveys, thus mitigating a barrier to more frequent measurements of SDG metrics.

In Section 1, we introduce the issues of developing and maintain household listings in LMICs. We also provide related work in the literature that complements this study. In Section 2, we describe the data used to train and evaluate our building detection model, as well as describing the model and associated evaluation metrics. In Section 3, we summarize the results of the study and conclude with a discussion of the findings in Section 4.

Related Work

In the past few years, the application of deep learning to satellite images has become a dominant trend in remote sensing [15,16]. Largely using variants of convolutional neural networks (CNNs), this class of models have shown state-of-the-art performance on tasks as diverse as scene classification [16–22], which assigns an aerial image into one of several distinct land-use or land-cover categories (residential, industrial, agriculture, etc.), to aerial image retrieval [23,24], which returns aerial images with similar visual content to a reference aerial image. The line of research most closely related to methods used in this study are the use of deep learning models for building detection [25,26] and building footprint segmentation [27–30]. Building detection tasks aim to determine both the presence and location of any buildings within an image, providing rectangular bounding box demarcations around the extent of any predicted buildings. Building footprint segmentation, while also concerned with the presence and location of buildings, approaches the problem by predicting the class label of every pixel in an image, allowing for a more flexible building boundary.

Method advances in this area have also helped drive novel applications, including studies utilizing satellite imagery to inform important SDG areas. Using a combination of publicly available daytime satellite images and nighttime lighting images, Jean et al. [31] applied a transfer learning approach on pre-trained CNNs to improve predictions of the average household expenditures or household wealth in five African countries. Gebru et al. [32] used deep learning object detection models to identify the make, model, and year of cars observable from satellite images in the United States. When aggregated, these counts were found to be highly predictive of neighborhood sociodemographic characteristics, such as household income, level of education, and race, as reported in the American Community Survey. Oshri et al. 2018 [33] used CNNs on Landsat 8 and Sentinel 1 satellite imagery to predict coverage of important infrastructure, such as sewage systems, piped water, and electricity services, across 36 African countries.

Compared to prior work, our key contribution is training and assessing deep learning building detection models to support household listing development in LMICs. While survey researchers have proposed the use of aerial imagery to help with household enumeration [14,34,35], to the authors' knowledge, this is the first work to assess the feasibility of these recommendations through use of machine learning models. Additionally, our study utilizes survey data from the Alive and Thrive initiative as an on-the-ground verification of model predictions. Because "ground truth" data are typically rare for aerial object detection tasks, we provide valuable insight into implementation considerations.

2. Data and Methods

2.1. Data

2.1.1. Building Detection Data

To develop the building detection model, we used 200 satellite images covering 11 local government areas (LGAs) in the Nigerian state of Kaduna. The Kaduna state occupies a land area of roughly 46,000 km^2 and despite rapid urbanization in past years, has vast areas of undeveloped savanna vegetation and farmland [36]. The selected areas ranged from isolated, rural areas, to the densely populated urban centers of Kaduna and Zaria. Due to the sample design described in Section 2.1.2, the settlement patterns in images from the selected LGAs are generally similar to areas found in the other 12 LGAs that were not included in the sample.

These images represent grid cells that are 100 × 100 m in size and were accessed using a layer package in ArcGIS (ESRI, Redlands, CA, USA) that contained the Bing Maps aerial imagery web mapping service. The Bing imagery data used in this study are sourced from organizations such as Digital Globe and are the most recently available images from the service for any given location. The use of the imagery for the purposes of this study complies with the terms of use agreement for the service offering. The spatial resolution of the images ranges from 0.31 to 0.5 m, though the spatial resolution of any particular image is difficult to verify because the image service provides a mosaic of imagery taken on different dates within a small range of spatial resolution values. The date of the imagery used for this analysis ranges from 2010–2016, with images in remote areas tending to be older (with dates as old as 2010), and urban images exhibiting more recent updates (with dates ranging from 2014–2016). The size and location of images were chosen to align with the application of the geo-sampling methodology [37] used in the Nigeria Alive and Thrive baseline survey fielded from 2016–2017. The Alive and Thrive survey and geo-sampling methodology are explained in further detail in Section 2.1.2.

These 200 images were allocated into training and test sets using a 64/36 split. We then created bounding box annotations around all building structures visible in the images, resulting in 2711 labeled buildings in the training set and 1844 labeled buildings in the test set. For the purpose of this study, "buildings" were defined as roofed structures visible from overhead imagery. This high-level category is expected to contain detached, semi-detached, and attached housing structures, as well as potentially

commercial or agricultural buildings. Of the 200 satellite images, 192 contained at least one building structure. Table 1 shows a summary of the data used for modeling. Figure 1 depicts an example of an annotated image after bounding boxes were applied around buildings.

Table 1. Summary of training and test data sets.

Summary Statistic	Training	Test
Total images	128	72
Total annotated buildings	2711	1844
Annotated buildings per image		
Mean	22.5	26
Min	1	1
Median	18	18
Max	107	76

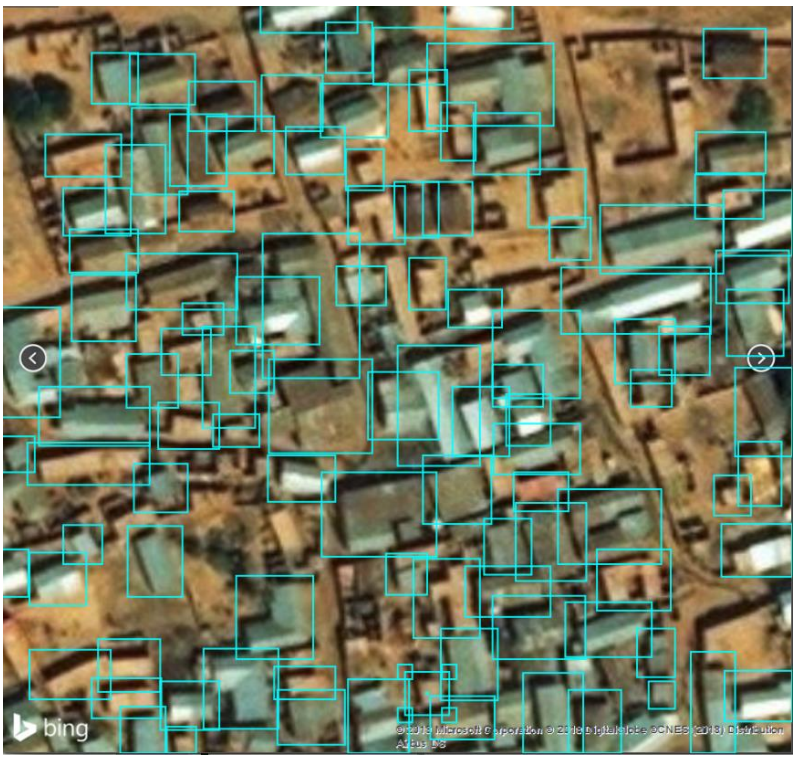


Figure 1. Example of an 100 × 100 m image annotated with human-labeled bounding boxes.

To use as model inputs, RGB (red, green, blue) values were extracted for each pixel in the sampled satellite images. Specifically, we created a tensor of RGB values for each image representing a 100 × 100 m gridded area. The RGB values for each pixel in each image were extracted using the Python Imaging Library (PIL) and were resized from 720 × 720 px to 750 × 750 px to match the pre-processing steps outlined in the papers for the model architectures used.

2.1.2. Alive and Thrive Survey Data

While building detection represents a step toward the automated construction of household rosters, it merely serves as an approximation to household detection. This is because not all building structures represent occupied households (e.g., buildings may be commercial or vacant). To understand the relationship between predicted buildings and households, we use survey data from the Alive and Thrive program to review the number of buildings and households within sampled areas.

Alive and Thrive (A&T) is a multinational program designed to increase the use of best nutritional practices for infants and young children. An impact evaluation is currently underway in two states in

Nigeria, Lagos and Kaduna, where a baseline and endline household survey are being used to assess program effectiveness. The sample for the baseline survey was selected using a gridded population method called geo-sampling. Geo-sampling applies layers of GIS grids to define manageable area units for sampling. The first GIS layer creates a set of units called *primary grid cells*, which are 1 km^2 in size. These grid cells are then further partitioned into smaller areas, called *secondary grid cells (SGCs)*, which depending on urbanicity, can be of sizes 50 × 50 m (urban), 100 × 100 m (semi-urban), or 150 × 150 m (rural). To simplify the selection of households, all households within the sampled SGCs are sampled for the survey, providing a complete representation of the number of households within each sampled SGC. Because all households within sampled SGCs were captured regardless of whether they met eligibility criteria for the survey, these data provide an opportunity to explore the relationship between household and building detection. We specifically compare the household counts from sampled 100 × 100 m grids from the Alive and Thrive baseline survey with building counts generated by our model. This grid size was chosen because it captures the broadest diversity of landscapes and development patterns of the three SGC grid sizes. While comparing geocoordinates of sampled households would provide a stronger benchmark for building predictions, these data were not used due to known measurement error in the household geolocations.

2.2. Methods

2.2.1. Building Detection Models

To determine the existence and location of buildings from satellite images, we developed a building detection model using the Tensorflow Object Detection API [38]. The model uses the Inception-Resnet-v2 [39] architecture for classification and Faster R-CNN [40] for localization. Inception-Resnet-v2, a model exhibiting state-of-the-art classification accuracy on the ImageNet benchmark dataset when it was published in 2016, combines the computational efficiency of Inception units from the popular Inception V3 architecture [41] with the residential connections introduced in the ResNet architecture [42]. Faster R-CNN is a meta-architecture that uses features from an object recognition model (in our case, Inception-Resnet-v2) to predict class-agnostic box proposal regions within an image where it believes there is a high probability of an object to exist. Then, focusing on just these few proposal regions, a final classification prediction is made to guess the object type (in our case, buildings). While relatively slower than other object detection model configurations, we selected this combination because it tends to be more thorough when scanning for objects within an image than models using only single stage detection (e.g., the Single Shot Detector (SSD) meta-architecture [43]). Since our use case of training a building detection model for use in statistical agency home offices does not require detection of buildings in real-time or on resource constrained devices, such as mobile phones, we were willing to trade off speed for improved model performance.

Training deep learning models from scratch often requires large quantities of labelled data to effectively generalize to new cases. To help our model learn a new task efficiently with less data, we used a transfer learning approach [44] to initialize our model weights. The intuition behind transfer learning is that information learned from performing a task in one domain can be built upon to more easily learn the same (or a similar) task in a different domain. For our source domain, we use a pre-trained model built on the Common Objects in Context (COCO) dataset [45]. The COCO dataset is an object detection dataset containing a total of 2.5 million labeled objects from 328 thousand images and spanning over 91 object types. While the COCO dataset does not contain labeled buildings, researchers have nonetheless used transfer learning models pre-trained on COCO to effectively detect new object classes not present in COCO, such as prohibited items in carry-on luggage [46] and traffic signs [47]. In our case, we are using the underlying patterns learned to effectively detect objects in the COCO dataset to initialize our model, thereby requiring our model to need less labeled data than if we had started with naïve weight parameters of all zeros.

2.2.2. Model Evaluation Metrics

To determine how well our model detects buildings in the test set, we use the mean average precision (mAP) evaluation metric [48]. For a given object type, the mAP summarizes the balance between model precision (the number of true positives out of all observations that are predicted positive) and recall (the proportion of true positives detected) by calculating the mean precision at a set of eleven equally-spaced recall thresholds [0, 0.1, ..., 1]:

$$mAP = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} p_{interp}(r) \quad (1)$$

where $p_{interp}(r)$ is the interpolated precision at each recall level, determined by taking the maximum precision seen for predicted objects with recall exceeding r. If a model predicts multiple object classes, the mAP additionally averages over all the class types. Note that Equation (1) assumes that the predicted objects mentioned are ranked in order of their predicted probabilities of detection. Intuitively, the mAP can be thought of as a summarization of the precision/recall curve, providing an average precision metric across various thresholds for detected labeled buildings. A mAP value of 0 means that no objects were correctly detected, whereas a value of 1 means that all objects were detected without any false positives.

In this study, a building is considered "detected" if its predicted building footprint has an intersection over union (IoU) of 0.5 or greater with a labeled building footprint. This is a common threshold used in the object detection literature [48]. The intersection over union, also known as the Jaccard index, is defined as:

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \quad (2)$$

where A is the set of pixels in a predicted building footprint and B is a set of pixels in a human labelled building footprint.

Lastly, to compare the actual household counts to predicted building counts within a 100 × 100 m gridded area, we use the Pearson correlation coefficient and Spearman's rank correlation coefficient. The Pearson correlation coefficient measures the linear relationship between two variables whereas the Spearman's rank correlation coefficient describes the extent two variables can be expressed using a monotonic function.

3. Results

3.1. Building Detection

On a test set consisting of 1844 annotated buildings across 72 images, our trained model detected 1081 buildings with a mAP of 0.48. Table 2 shows the precision values at all recall thresholds.

Table 2. Precision at each recall cutoff—all images.

Cut-Off	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	**mAP**
Precision	1	0.86	0.86	0.86	0.85	0.85	0	0	0	0	0	0.48

To penalize models from indiscriminately canvasing images with predicted bounding boxes (thereby maximizing the chance of finding buildings at the expensive of the false positive rate), the mAP requires that model precision is reported at different recall thresholds. The decomposition of the mAP in Table 2 suggests that our model tended to trade off precision for recall, producing high-precision predictions for areas where it had high confidence, while struggling to detect several buildings. Because the images in our test set had anywhere from 1 to 76 labeled building structures, we hypothesized that there may be a relationship between the number of labeled building in an image and the number of building detected by the model. To better understand the situations in which

our model had low recall and did not detect all buildings, we assessed the sensitivity of the model performance to building density.

Table 3 shows the precision values at all recall thresholds for images that contain less than 30 buildings. Using this cut-off, 408 out of a total 547 annotated buildings were detected across the 44 eligible images. These results indicate that, among areas with relatively fewer built structures, the model was able to retain high precision while also locating a high proportion of buildings.

Table 3. Precision at each recall cutoff among images with less than 30 buildings.

Cut-Off	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	mAP
Precision	1	0.89	0.89	0.89	0.88	0.87	0.85	0.9	0	0	0	0.65

Figures 2 and 3 illustrate the difference in predictions for an area with low building density (Figure 2), and an area with high building density (Figure 3). In the image representing low building density, the model achieved both high recall and high precision, whereas in the image with high building density, precision remained high but recall falls. For additional analyses comparing model performance at different levels of building density, please see Appendix A.

Figure 2. Predicted boxes: low building density.

Figure 3. Predicted boxes: high building density.

3.2. Correlation between A&T Households and Predicted Building Counts

As previously mentioned, an ideal detection model would perfectly identify households from satellite images to efficiently construct household listings. Unfortunately, determining whether buildings and households should be treated as synonymous is difficult without on-the-ground validation. For example, there are many types of buildings, such as commercial or agricultural structures, where people do not reside. Similarly, it may be common for a single structure, such as an apartment building, to contain multiple households. To better understand this critical relationship, we used data from the Alive & Thrive baseline survey to compare the number of households to the number of predicted buildings within 100 × 100 m grids. More specifically, we calculated the Pearson and Spearman's correlation to assess the linear and ranked associations between the two measures.

Table 4 shows the Pearson and Spearman's correlation between the number of households and predicted number of buildings within a sampled 100 × 100 m gridded area. The Pearson correlation of 0.702 suggests a strong linear relationship between the number of households and the number of predicted buildings within a grid. Likewise, the Spearman's rank correlation, which is more tolerant of outliers by only assessing the relative ranking between two variables, also showed a strong relationship with a value of 0.806.

Table 4. Correlation between A&T households and predicted building counts.

Type	Correlation
Pearson	0.702
Spearman	0.806

To further assess the relationship between the number of households and predicted buildings, we present a scatter plot of the two variables in Figure 4. Each point on the plot corresponds to a different SCG grid area sampled in the Alive and Thrive baseline survey. From Figure 4, we see that although the number of predicted buildings tended to increase as the number of households increased, the variability in the differences between the household and predicted building counts also tended to increase.

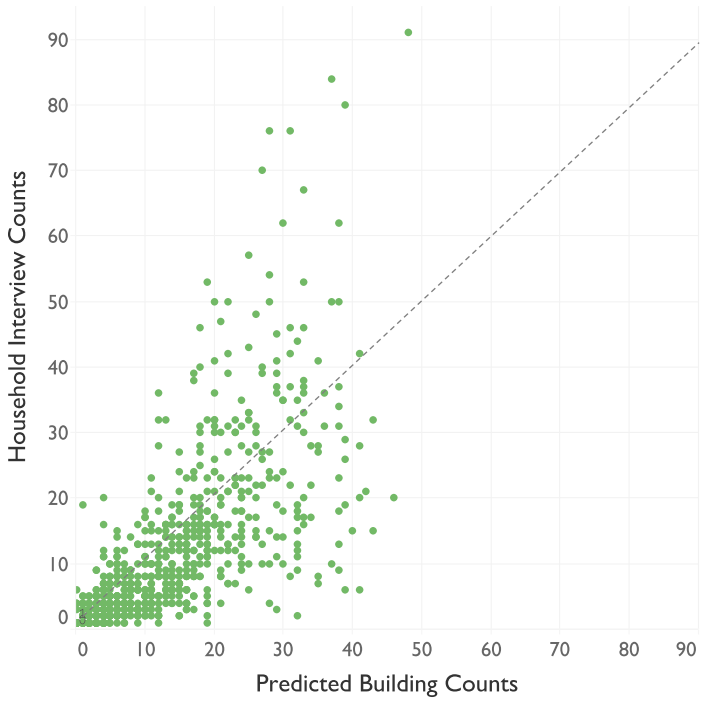
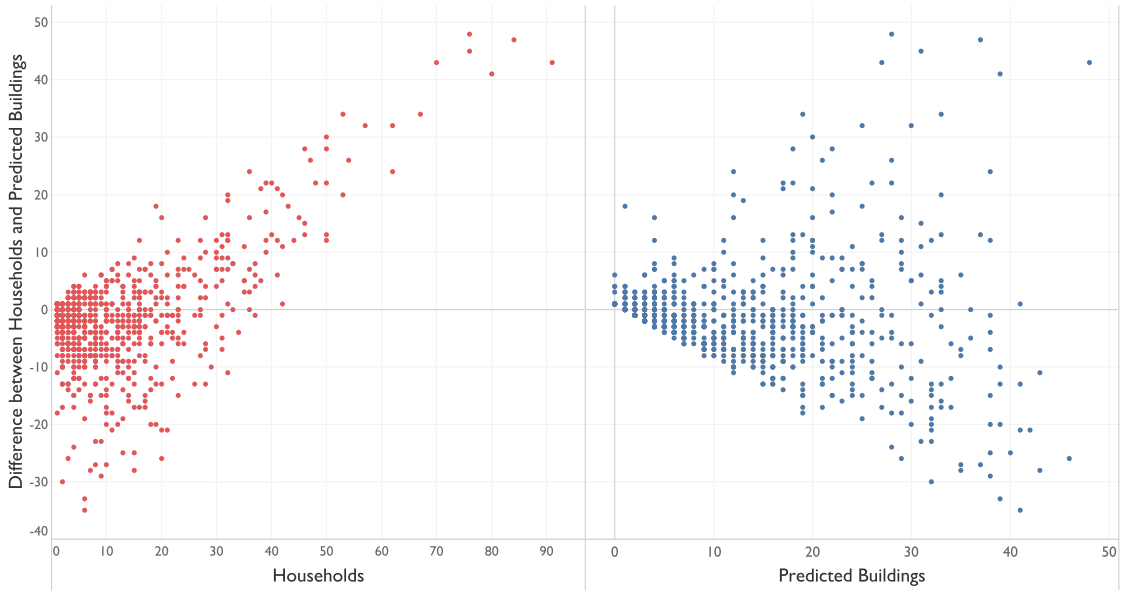


Figure 4. Scatterplot of household counts vs. predicted building counts with a 1:1 diagonal line.

For household counts, this heteroskedastic trend may be partially explained by our model under-detecting structures in highly dense areas (as reported in Section 3.1). To examine this hypothesis, the left panel of Figure 5 displays the relationship between the household count and the difference between household and predicted building counts. For grids with greater than or equal to 30 households, 89% had less predicted buildings than households (72 out of 81), suggesting a scenario where multiple households reside in a single building. This contrasted sharply with grids containing less than 30 households in which only 27% had fewer predicted buildings than households (186 out of 698), suggesting that in addition to under-detecting structures in building dense areas, households may also have been underreported.

For the plot of predicted buildings shown in the right panel of Figure 5, the shape appears to be more symmetric, making the underlying cause of discrepancies in the building and household counts less straightforward to diagnose. Due to this complication, this relationship is explored further in the Discussion section.



(**a**) (**b**)

Figure 5. Difference between Alive and Thrive (A&T) survey household counts and predicted buildings per SGU vs. (**a**) household counts per SGU and (**b**) predicted building counts per SGU.

4. Discussion

Our findings suggest that, while not without flaws, deep learning models show promise in performing building detection in LMICs settings. The model results suggest that building detection is more difficult to perform consistently in areas with high building density for the class of models examined. These results agree with similar findings in the Object-Based Image Analysis (OBIA) literature, in which building extraction tasks are reported as being more difficult to perform in residential areas than central business districts, where there is higher spectral complexity and building displacement [49]. Additionally, other studies report difficulty identifying attached building types more likely to be seen in dense urban areas (i.e., apartments) [50]. Detecting and counting objects in highly dense scenes is an active area of research in the computer vision literature, spanning from crowd counting and density estimation [51,52] to counting for cell microscopy [53,54]. Future work that explores models designed specifically for these settings could enhance building detection in urban areas and other settlement dense regions.

Furthermore, our findings suggest that predicted building counts have high correlation with the number of households within a region, which is encouraging for moving towards an automatic

listing of households. Interestingly, the trend between household counts and predicted building counts appears to be heteroskedastic in our study area, showing larger variances in differences when higher values of either variable are observed. While the grids with a notably higher number of households than predicted buildings may be partially explained by the findings in Section 3.1, in which our model detects a lower proportion of buildings in areas of high building density, the reason for having areas with significantly more predicted buildings than households is less clear. To demonstrate, Figure 6 juxtaposes two images from our test data representing urban areas of Kaduna; both of these images have a similar number of predicted buildings but vastly different number of reported households. Figure 6a has 40 predicted buildings with only 6 households, whereas Figure 6b has 48 predicted buildings with 91 households.

(**a**) (**b**)

Figure 6. (**a**) SGC grid area with 40 predicted buildings and 6 households; (**b**) SGC grid area with 48 predicted buildings and 91 households.

There are many potential reasons why this might occur. One possible explanation is that the dense areas from our study have greater diversity in the types of building structures present (e.g., some images may represent a higher concentration of commercial buildings). Though Kaduna has a history of weak land-use planning [55], which could allow for large variations in mixed-use development in its urban corridors, more research is needed to confidently test this hypothesis. Another hypothesis for the observed variation in household counts is that interviewers did not consistently adhere to the survey's household enumeration protocols within sampled SGCs. In urban neighborhoods where buildings are dense, recording all households is logistically difficult (e.g., entrances to some residences may be blocked or not visible from the road). In these cases, household counts may be underestimated. Intentional data falsification by field interviewers is a well-documented phenomenon in survey data collection [56,57] and may also contribute to some instances of low household counts in areas with many predicted buildings. We are hopeful that advances in technology will offer improvements to the "on-the-ground" meta-data typically collected during surveys. Enhancements to these data would provide more robust information for validation of building detection tasks and strengthen researchers' understanding of the relationship between buildings and households.

The adoption of building detection methods as a streamlined approach for constructing household sampling frames ultimately depends on the minimization of prediction error. Prediction error introduces two types of challenges for a survey: (1) overcoverage, where the constructed listing includes buildings that are not of interest for the study, and (2) undercoverage, where the listing excludes buildings that belong to the intended target population. To illustrate this point, let us imagine a scenario where a list of predicted buildings is generated for an area and this list is used exclusively as

the sampling frame for a household survey measuring SDG indicators. In cases where field staff visit buildings that do not contain households (overcoverage), unnecessary costs may be incurred to the project budget from the wasted time and travel resources; however, no bias should be introduced into the survey estimates in this case, assuming that the nonresidential or vacant buildings can be identified during data collection and appropriate adjustments can be made during analysis. Alternatively, if the model fails to detect all buildings that contain households and thus excludes them from the sampling frame (undercoverage), bias becomes a substantial concern. Survey estimates derived from a sampling frame suffering from undercoverage may include error because the excluded households may be systematically different from those represented in the sample. While methods are available to compensate for housing unit undercoverage [58], they will add to the cost of the study.

Understanding these implications of prediction error allow us to better understand under what conditions the current model should be used in the field. Given the favorable household coverage enabled by the building detection model in rural areas within our study, there is support for it being an effective option for developing household lists in low building dense areas. This is especially true when considering the lack of high-quality existing frames in many LMICs. However, due to concerns of potential bias introduced by undercoverage, the presented building detection model may be better utilized as a supporting approach in urban areas where there is greater variability in differences between households and predicted buildings. As modeling approaches advance and quality annotated datasets become more widely available, we expect to see these methods become increasingly useful.

In addition to developing household lists, building detection models can provide research teams with other valuable options for conducting high-quality household surveys. Though it is not a focus of this study, predicted building counts could also be used as a measure of size for probability proportional to size (PPS) sampling [59]. PPS is a sampling technique that selects units in one sampling stage with probabilities proportional to a measure of size, followed by the sampling of a fixed number of units at the next stage. The larger the unit's size, the greater its chance of being included in the sample. The advantage of PPS is that it leads to equal overall sampling probability, while at the same time maintaining a uniform work load for each unit in the first stage. Additionally, for sampling designs in which a full enumeration of households is conducted at the final stage of sampling, calculating predicted building counts prior to data collection could provide a valuable quality check when the actual household counts deviate greatly from the predicted building counts, empowering survey managers to solicit context from field teams to better understand why differences occur.

While this study only assesses object detection models trained on imagery from one state in Nigeria, there is potential for the same or similar methods to be scaled to larger areas to develop regional or national household SDG indicators. One potential bottleneck to implementing these models in new areas is the large amount of labeled training data required to train convolutional neural networks from scratch. To address this in our study, we use transfer learning to initialize our building detection model weights prior to model training. Tiecke et al. [30] take a different approach, reducing the labeling problem to a binary classification task (labelling imagery of 30 $\times$ 30 m area grids as "containing buildings" or "not containing buildings"), which is easier to obtain labels for. They then use these labeled data to train a weakly-supervised semantic segmentation model to predict pixel-level building labels.

Besides training data, a lack of computational resources in LMICs may also hamper usage of these models in practice. While cloud computing options for even specialized Graphics Processing Unit (GPU) servers are becoming increasingly accessible and affordable, reducing the areas required for household listing can also help lessen the computational load of creating large-scale building predictions. Using traditional image processing models that do not require training data, such as conventional edge detectors, can help reduce the number of regions needed to be screened and modeled [30]. Additionally, if incorporated into a clustered sampling design, only selected enumeration areas would require household enumeration instead of requiring a comprehensive national household listing.

There are some limitations of the study. First, we do not have records that directly link buildings and households and thus were unable to build models that detect residential structures specifically. Given the existing limitations on the availability of household data in many LMICs, this information would likely need to come from existing household surveys that use mobile devices, tablets, etc., to record the location of household units during interviews. Second, there may be error in the manual labeling of building outlines for the training and test sets, as well as error introduced during data collection with respect to household coverage. Labelling for different types of residential structures (apartments, single family homes, etc.) should also help better characterize the heterogeneity in the number of households per building. This may require recruiting labelers with in-country knowledge of various building types, perhaps also using higher resolution imagery than was assessed in this study. Additionally, this methodology assumes that available satellite data provides an up-to-date portrait for the buildings and households that will be present during the survey data collection period. While it is not uncommon in the literature to have a modest temporal gap between satellite imagery dates and the survey data collection period for assessing these classes of models [30], recognizing the potential for errors can help statistical agencies be proactive in identifying emerging issues. To a certain extent, unmanned aerial vehicles (UAVs) could provide more detailed and timely imagery to help mitigate this concern. While our findings suggest that building density was correlated with the model accuracy, future work may benefit from a more exhaustive exploration into what conditions and settings are challenging for current models used in building detection. In particular, an understanding of how building size, geometry, and type might affect model performance would help survey researchers and statistical agencies better assess where there are still opportunities for improvement. Lastly, while we only provide evidence for a single state in Nigeria, we hope this case study encourages further research and resources to examine a larger-scale implementation of these methods for household enumeration in LMICs.

5. Conclusions

As countries strategize how to improve measurement of SDG indicators going forward, household surveys will likely continue to play a major role in creating reliable and comparable statistics. Given the extensive resources required to construct household listings in many LMICs, object detection, and segmentation models could help enhance household enumeration exercises by presenting survey researchers with building lists for direct sampling or quality checks. In particular, refined models of the types discussed in this paper could eventually allow survey teams to send field staff directly to sampled addresses/buildings and drastically reduce, if not eliminate, the need for on-the-ground fieldwork. When considering the cost and time required to conduct household surveys, modeling approaches to reduce the level of effort and increase data collection efficiency will be crucial for widespread adoption and effectiveness of the SDG indicator program.

Author Contributions: Conceptualization, R.C., S.A., and K.K.; Methodology, R.C. and K.J.; Formal Analysis, K.J. and R.C.; Data Curation, J.U., K.J., J.C., and J.A.; Writing-Original Draft Preparation, R.C., K.J., and J.U.; Writing-Review and Editing, J.C., J.A., S.A., and K.K.; Visualization, K.J. and R.C.; Funding Acquisition, S.A.

Funding: This research was funded by RTI International Internal R&D funds.

Acknowledgments: We would like to thank the Alive and Thrive initiative for access to the baseline survey data.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Predicted Building Counts versus Building Density

To illustrate how building predictions are influenced by building density, we report the mean absolute error (MAE) for various maximum building count thresholds. The MAE is calculated using:

$$MAE = \frac{\sum_{i=1}^{n} \left| b_i - \hat{b}_i \right|}{n} \quad \text{(A1)}$$

where b_i is the number of human labeled buildings in grid i, $\hat{b}_i$ is the number of predicted buildings detected in grid i, and n is the total number of 100 × 100 m grid area units compared.

Table A1 compares the average labeled and predicted building counts for different building density thresholds in the test set. For example, if considering only images with fewer than 10 buildings, the test set contains 16 such images, and the MAE is 1.44.

Table A1. Errors for building counts at different maximum building count thresholds.

Maximum Buildings in Image	Test Images	Percent of Test Images	Average Building Count	Average Predicted Building Count	Mean Absolute Error
<10	16	23%	5.19	5.88	1.44
<20	37	52%	10.16	9.68	1.78
<30	44	62%	12.43	11.23	2.30
<40	50	70%	14.82	13.02	2.76
<50	62	87%	20.47	16.08	5.16
<60	64	90%	21.52	16.39	5.88
<70	69	97%	24.58	17.62	7.65
<80	72	100%	25.97	18.20	8.45

In low building density areas (images with less than 30 buildings), the model achieved an MAE of 2.30, suggesting that the predicted building count per image differed from the actual building count per image by 2.30 on average. However, as the number of buildings per image increased, so did the MAE. For images that had a high number of buildings (50+), the model was unable to locate a high number of the buildings. In particular, as we looked across all 72 images in the test set where up to 80+ buildings per image are present, the model underestimated the number of buildings per grid area by 8.45 on average.

References

1. About the Sustainable Development Goals. Available online: https://www.un.org/sustainabledevelopment/sustainable-development-goals/ (accessed on 12 May 2018).
2. United Nations Statistical Commission (UNSD). *Report of the Inter-agency and Expert Group on Sustainable Development Goal Indicators*; UNSD: New York, NY, USA, 2017.
3. The Economist. The 169 Commandments: The Proposed Sustainable Development Goals Would Be Worse than Useless. Available online: https://www.economist.com/leaders/2015/03/26/the-169-commandments (accessed on 12 May 2018).
4. Renwick, D. Sustainable Development Goals. The Council on Foreign Relations. Available online: https://www.cfr.org/backgrounder/sustainable-development-goals (accessed on 12 May 2018).
5. Organisation for Economic Co-operation and Development. *Measuring Distance to the SDG Targets: An Assessment of Where OECD Countries Stand*; OECD: Paris, France, 2017.
6. United Nations Statistical Commission. *Report of the Intersecretariat Working Group on Household Surveys*; UNSD: New York, NY, USA, 2017.
7. Groves, R.M. *Survey Errors and Survey Costs*; John Wiley and Sons: New York, NY, USA, 1989.

8. Groves, R.M.; Lyberg, L. Total survey error: Past, present, and future. *Public Opin. Q.* **2010**, *74*, 849–879. [CrossRef]
9. Biemer, P.P. Total survey error: Design, implementation, and evaluation. *Public Opin. Q.* **2010**, *74*, 817–848. [CrossRef]
10. Alkire, S.; Samman, E. *Mobilising the Household Data Required to Progress toward the SDGs*; OPHI Working Paper 72; Oxford University: Oxford, UK, 2014.
11. Demographic and Health Surveys Program. SDG Indicators in DHS Surveys. Available online: https://dhsprogram.com/Topics/upload/SDGs%20in%20DHS%2018May2017.pdf (accessed on 15 May 2018).
12. Shannon, H.S.; Hutson, R.; Kolbe, A.; Stringer, B.; Haines, T. Choosing a survey sample when data on the population are limited: A method using Global Positioning Systems and aerial and satellite photographs. *Emerg. Themes Epidemiol.* **2012**, *9*, 5. [CrossRef] [PubMed]
13. Burnham, G.; Lafta, R.; Doocy, S.; Roberts, L. Mortality after the 2003 invasion of Iraq: A cross-sectional cluster sample survey. *Lancet* **2006**, *368*, 1421–1428. [CrossRef]
14. Eckman, S.; Eyerman, J.; Temple, D. *Unmanned Aircraft Systems Can Improve Survey Data Collection*; RTI Press: Research Triangle Park, NC, USA, 2018. [CrossRef]
15. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
16. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
17. Basu, S.; Ganguly, S.; Mukhopadhyay, S.; DiBiano, R.; Karki, M.; Nemani, R. Deepsat: A learning framework for satellite imagery. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2015; p. 37.
18. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land use classification in remote sensing images by convolutional neural networks. *arXiv* **2015**, arXiv:1508.00092.
19. Han, X.; Zhong, Y.; Cao, L.; Zhang, L. Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sens.* **2017**, *9*, 848. [CrossRef]
20. Hu, F.; Xia, G.S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
21. Zhong, Y.; Fei, F.; Liu, Y.; Zhao, B.; Jiao, H.; Zhang, L. SatCNN: Satellite image dataset classification using agile convolutional neural networks. *Remote Sens. Lett.* **2017**, *8*, 136–145. [CrossRef]
22. Chew, R.F.; Amer, S.; Jones, K.; Unangst, J.; Cajka, J.; Allpress, J.; Bruhn, M. Residential scene classification for gridded population sampling in developing countries using deep convolutional neural networks on satellite imagery. *Int. J. Health Geogr.* **2018**, *17*, 12. [CrossRef] [PubMed]
23. Jiang, T.; Xia, G.S.; Lu, Q. Sketch-based aerial image retrieval. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3690–3694.
24. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sens.* **2017**, *9*, 489. [CrossRef]
25. Yang, H.L.; Lunga, D.; Yuan, J. Toward country scale building detection with convolutional neural network using aerial images. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 870–873.
26. Dickenson, M.; Gueguen, L. Rotated Rectangles for Symbolized Building Footprint Extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 225–228.
27. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. *arXiv* **2017**, arXiv:1709.05932.
28. Yuan, J. Automatic building extraction in aerial scenes using convolutional networks. *arXiv* **2015**, arXiv:1602.06564.
29. Zhang, A.; Liu, X.; Gros, A.; Tiecke, T. Building Detection from Satellite Images on a Global Scale. *arXiv* **2017**, arXiv:1707.08952.
30. Tiecke, T.G.; Liu, X.; Zhang, A.; Gros, A.; Li, N.; Yetman, G.; Kilic, T.; Murray, S.; Blankespoor, B.; Prydz, E.B. Mapping the world population one building at a time. *arXiv* **2017**, arXiv:1712.05839.
31. Jean, N.; Burke, M.; Xie, M.; Davis, W.M.; Lobell, D.B.; Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science* **2016**, *353*, 790–794. [CrossRef] [PubMed]

32. Gebru, T.; Krause, J.; Wang, Y.; Chen, D.; Deng, J.; Aiden, E.L.; Li, F.-F. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *arXiv* **2017**, arXiv:1702.06683. [CrossRef] [PubMed]
33. Oshri, B.; Hu, A.; Adelson, P.; Chen, X.; Dupas, P.; Weinstein, J.; Burke, M.; Lobell, L.; Ermon, S. Infrastructure Quality Assessment in Africa using Satellite Imagery and Deep Learning. *arXiv* **2018**, arXiv:1806.00894.
34. Eyerman, J.; Krotki, K.; Amer, S.; Gordon, R.; Evans, J.; Snyder, K.; Zajkowski, T. Drone-Assisted Sample Design for Developing Countries. In Proceedings of the FedCASIC Workshops, Washington, DC, USA, 4–5 March 2015.
35. Haenssgen, M.J. Satellite-aided survey sampling and implementation in low- and middle-income contexts: A low-cost/low-tech alternative. *Emerg. Themes Epidemiol.* **2015**, *12*. [CrossRef] [PubMed]
36. State Development Plan: 2014–2018. Kaduna State Government, Ministry of Economic Planning. Available online: http://www.sparc-nigeria.com/RC/files/1.1.9_Kaduna_Development_Plan_2014_2018.pdf (accessed on 2 November 2018).
37. Cajka, J.; Amer, S.; Ridenhour, J.; Allpress, J. Geo-sampling in developing nations. *Int. J. Soc. Res. Methodol.* **2018**, *21*, 729–746. [CrossRef]
38. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 4.
39. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017.
40. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
41. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. *arXiv* **2015**, arXiv:1512.00567.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385.
43. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
44. Pan, S.J.; Yang, Q.A. Survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
45. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, C.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
46. Liang, K.J.; Heilmann, G.; Gregory, C.; Diallo, S.O.; Carlson, D.; Spell, G.P.; Sigman, J.B.; Roe, K.; Carin, L. Automatic threat recognition of prohibited items at aviation checkpoint with X-ray imaging: A deep learning approach. In Proceedings of the Anomaly Detection and Imaging with X-rays (ADIX) III, Orlando, FL, USA, 27 April 2018.
47. Arcos-García, Á.; Álvarez-García, J.A.; Soria-Morillo, L.M. Evaluation of Deep Neural Networks for traffic sign detection systems. *Neurocomputing* **2018**, *316*, 332–344. [CrossRef]
48. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes (VOC) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
49. Hu, X.; Weng, Q. Impervious surface area extraction from IKONOS imagery using an object-based fuzzy method. *Geocarto Int.* **2011**, *26*, 3–20. [CrossRef]
50. Belgiu, M.; Tomljenovic, I.; Lampoltshammer, T.J.; Blaschke, T.; Höfle, B. Ontology-based classification of building types detected from airborne laser scanning data. *Remote Sens.* **2014**, *6*, 1347–1366. [CrossRef]
51. Rodriguez, M.; Laptev, I.; Sivic, J.; Audibert, J.Y. Density-aware person detection and tracking in crowds. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2423–2430.
52. Sindagi, V.A.; Patel, V.M. Generating high-quality crowd density maps using contextual pyramid CNNs. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1879–1888.

53. Cohen, J.P.; Boucher, G.; Glastonbury, C.; Lo, H.Z.; Bengio, Y. Count-ception: Counting by fully convolutional redundant counting. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy, 22–29 October 2017; pp. 18–26.
54. Xie, W.; Noble, J.A.; Zisserman, A. Microscopy cell counting and detection with fully convolutional regression networks. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2018**, *6*, 283–292. [CrossRef]
55. Dyachia, Z.S.; Permana, A.S.; Ho, C.S.; Baba, A.N.; Agboola, O.P. Implications of Present Land Use Plan on Urban Growth and Environmental Sustainability in a Sub Saharan Africa City. *Int. J. Built Environ. Sustain.* **2017**, *4*. [CrossRef]
56. Bredl, S.; Storfinger, N.; Menold, N. *A Literature Review of Methods to Detect Fabricated Survey Data (No. 56)*; Discussion Paper; Zentrum für Internationale Entwicklungs-und Umweltforschung: Gießen, Genmany, 2011.
57. Murphy, J.; Baxter, R.; Eyerman, J.; Cunningham, D.; Kennet, J. A system for detecting interviewer falsification. In Proceedings of the American Association for Public Opinion Research 59th Annual Conference, Phoenix, Arizona, 20–23 May 2004; pp. 4968–4975.
58. Shook-Sa, B.; Harter, R.; McMichael, J.; Ridenhour, J.; Dever, J. *The CHUM: A Frame Supplementation Procedure for Address-Based Sampling*; RTI Press: Research Triangle Park, NC, USA, 2016. [CrossRef]
59. Yates, F.; Grundy, P.M. Selection without replacement from within strata with probability proportional to size. *J. R. Stat. Soc.* **1953**, *15*, 253–261.