Taking the Long View: Enhancing Learning On Multi-Temporal, High-Resolution, and Disparate Remote Sensing Data

Santiago Correa
UMass Amherst
USA
scorreacardo@umass.edu

Gustavo Perez
UMass Amherst
USA
gperezsarabi@umass.edu

Paulina Jaramillo
Carnegie Mellon University
USA
pjaramil@andrew.cmu.edu

Jay Taneja
UMass Amherst
USA
jtaneja@umass.edu

ABSTRACT

The progress made in computer vision and satellite technology has opened up new possibilities for observing societies and infrastructure. By analyzing vast amounts of high-resolution multi-temporal satellite data, decision-makers can gain valuable insights into population shifts, economic trends, and infrastructure performance. Nevertheless, challenges in this kind of imagery, such as varying image quality, imbalances in data collection between urban and rural areas, high costs, and the absence of image metadata, can impede the efficacy of these methods.

In this work, we develop strategies for enhancing the performance of learning methods for high-resolution multi-temporal satellite imagery. We develop custom augmentation methods and inference techniques for identifying disparate image resolutions across historical imagery. We apply our generic techniques to the problem of detecting structures in longitudinal imagery, exhibiting modest but consistent performance improvements over baseline techniques. We then develop a case study analyzing the relationship between the expansion of electricity access and the growth in human settlements over time. We discover that across 1000 communities in Kenya over a decade, settings that received electricity access grew 15% more slowly than settings that did not receive electricity access. This non-intuitive and statistically robust finding challenges conventional wisdom about infrastructure provision and rural-urban migration, with potentially broad implications for assessing the impacts of infrastructure investments on rural lives and livelihoods. All data processing and modeling scripts are available at https://github.com/santiagocorrea/DeepSatGSD.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

BuildSys '23, November 15–16, 2023, Istanbul, Turkey

ACM ISBN 979-8-4007-0230-3/23/11…$15.00
https://doi.org/10.1145/3600100.3623722

ACM Reference Format:
Santiago Correa, Gustavo Perez, Paulina Jaramillo, and Jay Taneja. 2023. Taking the Long View: Enhancing Learning On Multi-Temporal, High-Resolution, and Disparate Remote Sensing Data. In *The 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '23), November 15–16, 2023, Istanbul, Turkey*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3600100.3623722

1 INTRODUCTION

Remote sensing has revolutionized how we measure, detect, and observe phenomena around the globe. It has become an indispensable tool for scientists, researchers, and policymakers due to its non-intrusive nature, extensive spatial coverage, and ability to collect data over time. Today's growing availability of hyperspectral and high spatial resolution satellite imagery has opened the door for applications across various domains such as economics, agriculture, sociology, and public services that report weather conditions and traffic patterns. Pairing this trend with increasing computing power and recent advances in artificial intelligence and computer vision, this source of information has uncovered innumerable previously known patterns and insights into Earth and infrastructure systems. The fast-growing compendium of historical satellite information enables unprecedented multi-temporal (longitudinal) observations, enabling examination of system changes over time, efficiently and at large scale [57]. Today, deep learning techniques can map low and high-level image representations by learning filters that produce responses to visual features (such as edges and colors), ultimately enabling scene recognition [4].

Multi-temporal imagery has become especially relevant for urban planners and policymakers. By analyzing scene changes over time, planners can better understand how human settlements are growing and changing. This information can help with zoning, infrastructure planning, and natural risk assessment decisions. These capabilities are particularly valuable in low-resource settings, where historical ground data collection has typically been limited, new data collection can be prohibitively expensive, and changes in population and climate may be highly dynamic. Despite these challenges, there have been numerous demonstrations using deep learning methods with remote sensing data, including for land-use monitoring [26], electricity consumption predictions [19], and poverty assessment [21, 54, 56].

However, coupling multi-temporal imagery with data-driven methods remains challenging because longitudinal images often have highly varying quality – with different feature resolutions and color profiles over time (see Figure 1 for examples). This is because sequences of satellite observations often come from multiple sensors over time. These sequences can also have a variety of noise sources, missing data, and gaps due to meteorological phenomena (snow and clouds) during the acquisition or distortions during storage and transmission [40]. Additionally, this problem of different sensors is exacerbated by satellite data collection strategies. Typically, urban and rural areas will have substantially different revisit frequencies – ranging from days and weeks in urban settings to months and years in rural settings – with high-resolution imagery data often collected in irregular intervals due to business reasons of the commercial entities that operate the satellites. The resulting imagery datasets are substantially imbalanced over time, requiring significant care to learn from with good performance. This may include curating training samples from different periods or even developing individual models based on image age or resolution.

This work aims to alleviate some inherent challenges in employing machine learning methods on multi-temporal satellite imagery. In particular, we develop a system for collecting publicly available high-resolution satellite imagery from Google Earth Pro, construct a pipeline for identifying a crucial metadata feature from those images – ground sample distance – and develop deep learning models that leverage these metadata to improve performance specifically for image segmentation tasks with multi-temporal data. In our analysis, we leverage publicly available data – including satellite imagery, census data, and building footprint data – and cutting-edge computer vision techniques. To demonstrate the value of our approach, we implement a deep learning-based technique to monitor changes in the rural built environment to conduct a national-scale case study on the relationship between the expansion of electricity access and growth in human settlements in developing regions, ultimately uncovering a statistically-robust, policy-relevant, and non-intuitive finding. Our key contributions to the literature are:

- A collection of strategies for enhancing segmentation-based building footprint identifiers, particularly for multi-temporal imagery. We expressly do not claim novelty for enhancing the core and well-studied segmentation techniques used for building footprint identification but demonstrate particular methods for improving their performance with longitudinal imagery, including developing distinct models for images at different resolutions and employing custom data augmentation methods for multi-temporal imagery to accommodate varying color profiles and imbalanced training sets.
- A learning model to infer ground sample distance metadata from satellite imagery. This model was motivated by the circumstance arising from the freely available satellite imagery source that we employ (Google Earth Pro), which does not provide metadata information about the acquisition sensor, and they can vary between multi-temporal images of the same location. While calibration objects can help estimate ground sample distance, their presence is not always guaranteed, and their size can also vary. However, a learning model can detect and calculate this feature from pixels and estimate it at scale.
- An unexpected correlation between lack of electricity access and faster growth in human settlements. While it is typically thought that expansion of electricity services will attract rural migration, our statistically significant findings across 1000 communities throughout the country of Kenya over a decade show that settings that received electricity access grew 15% slower than those that did not receive electricity. This calls for further research and can have implications for strategies for rural economic development via infrastructure investments.

2 RELATED WORK

In the last decade, remote sensing data has become essential for places with difficult access, limited resources to collect data on the ground, and studies that require broad spatial coverage. This section highlights three trends intersecting our work: applications that leverage remote sensing in developing regions, state-of-the-art semantic segmentation techniques using deep learning, and building detection and change monitoring using satellite imagery.

2.1 Remote Sensing Applications

A key area of application of remote sensing-based techniques has been agriculture, with substantial work developing deep learning models to monitor crops and agricultural systems from space [20, 33, 39, 43, 45, 51, 52] as well as map cropland using semantic segmentation [52]. These techniques mainly leverage multi-spectral imagery at a moderate resolution ($30m$) from MODIS and Landsat-8 satellites. These satellites provide multi-temporal observations at regular fixed intervals, presenting different challenges as compared to applications like ours that use high-resolution imagery, which arise from a variety of sensors at irregular temporal resolutions.

Another prominent problem space is monitoring and estimating poverty and well-being [22, 23, 31, 55]. Given the geographic distribution of people under the poverty line, poverty maps are essential inputs for poverty alleviation, political accountability, and impact evaluation. For instance, in [31], the authors combine high-resolution satellite imagery with nighttime maps, arguing that nighttime data are a rough proxy for determining economic wealth.

Satellite imagery and convolutional neural networks (CNNs) have also been used in rural and developing regions for monitoring road quality in Kenya [6], mapping infrastructure quality in Africa using survey data as ground truth [36], and evaluating electrification using globally-available and temporally-rich nighttime satellite data [13]. Like our work, these approaches of poverty measurement and infrastructure monitoring typically reveal reduced performance in rural regions but often do not consider longitudinal analysis, limiting their findings to individual points in time (with the exception of [13])..

A number of other applications employ CNNs and high-resolution longitudinal satellite imagery to monitor infrastructure, including the detection of rooftop solar panels [7, 28] and solar power plants [27, 30]; the estimation of generation capacity of these systems using weather forecasting data, solar irradiation [29, 41]; localization of wind turbines [61]; monitoring of grid infrastructure such as power lines [44]; and modeling of building energy consumption [17]. However, these applications are typically restricted to urban areas and high-income regions with substantially more

			Desired Features			
Dataset	**Resolution(m)**	**Data Source**	**VHR(0.5m/px)**	**AoI**	**Annotations**	**Multi-temporal**
Landcover.ai [5]	0.25-0.5	Aerial	√	×	√	×
UC Merced [53]	0.3	USGS National Map	√	×	*	×
PatternNet [62]	0.06-5	Google Earth	√	×	*	×
LEVIR-CD+ [46]	0.5	Google Earth	√	*	√	*
SpaceNet 2 [50]	0.3-1.24	WorldView-3	√	×	√	×
SpaceNet 7 [18]	4	Planet	×	*	√	√
RESISC45 [11]	0.2-30	Google Earth	√	×	*	×
Google Open Buildings [47]	–	Google Earth	×	√	√	×

Table 1: Some state-of-the-art datasets and approaches that aim to measure urban change using remote sensing techniques and artificial neural networks. The desired features are shown in the last four columns to estimate longitudinal structure change in our Area of Interest (AoI). Check marks indicate that the listed approach complies with the feature, an asterisk indicates partial compliance, and the x mark indicates that the feature is absent.

voluminous and homogeneous high-resolution imagery and greater ground truth data availability.

2.2 Semantic Segmentation

Image semantic segmentation divides an image into multiple regions by assigning a class to each pixel within the image. Some applications include improving the performance of object detection and recognition algorithms, scene understanding, or video indexing [59]. CNNs are the most effective approach to performing semantic segmentation [12] and currently, U-Net-based methods [42], as well as spatial pyramid pooling-based architectures such as PSPNet [58], PAN [34] and DeepLabViewV3+ [10], are commonly used.

U-Net [42] uses skip connections to help prevent information loss as data flows through the network, leading to faster training speed and improved accuracy. PSPNets [58] use a pyramid scheme to enable a more effective learning process from data distributed unevenly across spatial dimensions. Recent segmentation methods such as Segment Anything [32] leverage Vision Transformers [16] motivated by its scalability and representational power.

In our work, we apply these methods and explore ways to enhance their performance specifically for multi-temporal imagery via custom augmentation techniques to replicate the idiosyncrasies of varying satellite sensors and to accommodate training data imbalances. We also develop a method to infer ground sample distance from satellite imagery with unknown metadata in order to enhance model training for specific vintages and resolutions of imagery.

2.3 Change and Building Detection Using Satellite Imagery

Given its significance for infrastructure assessment, urban planning, and development, building extraction from remotely sensed data and change detection have been intensively explored in the literature [9, 25, 35, 37, 60]. Most building detection approaches rely on encoder-decoder architectures and semantic segmentation techniques as described in 2.2.

Multiple building detection competitions have enabled the proliferation of models and publicly available imagery [15, 18, 50]. However, they tend to have limited geographical and temporal scope, with few examples focused on developing regions. To efficiently perform longitudinal building detection in developing regions, the input datasets for training a supervised learning model must include imagery from the area of interest (context matter), a very high spatial resolution (VHR) to detect small structures, annotations, and multi-temporal composites to assess changes over time.

Table 1 summarizes state-of-the-art datasets and approaches that aim to measure settlement change using remote sensing techniques and artificial neural networks. As we can observe, none of the existing approaches comply completely with the desired features to accurately estimate structure change in developing settings. For instance, recently released building footprint estimates for the entire African continent were recently released but without the input imagery or temporal metadata [47]. [18] provides multi-temporal composites and annotations but at a much lower resolution and only for select urban areas. The characteristics that make this task most difficult are the multi-temporal nature and imagery of rural places in underdeveloped settings. With these settings experiencing rapid urbanization and swift changes due to climate change (as in much of the African continent), it is essential to track changes over time, particularly for applications related to infrastructure provision, economic growth, and livelihood development.

3 METHODS

Figure 2 illustrates our change and building detection framework. First, we develop a tool to collect raw imagery in our area of interest at VHR. Second, we perform significant image pre-processing to mitigate the presence of artifacts. Third, we train a classifier to improve the performance of the building detector. Our classifier leverages the relationships between images captured at different scales with a building segmentation model specialized for ground sample distance (GSD). Finally, we perform building matching on subsequent imagery to assess structure change and growth.

3.1 Data Collection

We developed an automatic data collection tool using Google Earth Pro Desktop (GEP), a computer freeware that creates 2D and 3D models of the Earth, mostly using satellite images. Users can navigate globally by entering addresses and coordinates using a user interface. By controlling the zoom levels, users can access imagery resolution ranging from $15m$ to $15cm$ per pixel. Additionally, GEP

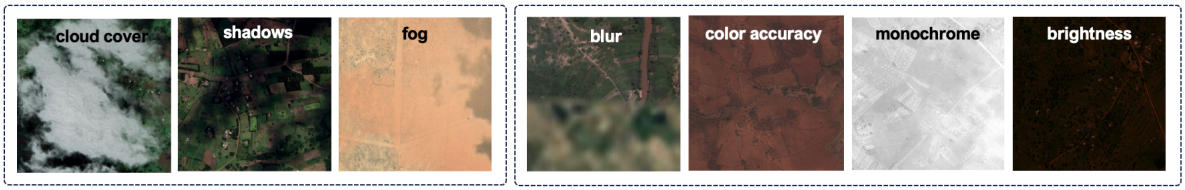

Figure 1: (a) Natural artifacts during the image capture: undesired and inherent image phenomena were identified during the data collection step such as cloud cover, shadows produced by clouds, and fog due to sand in desert areas. (b) Examples of artificial artifacts encountered in the imagery collected using our data collection tool.

provides historical images that can be accessed and downloaded using a time slider icon. Large cities and urban areas typically have the best spatial and temporal resolution. However, selecting and downloading multi-temporal imagery from a given location requires user manual intervention, which is tedious and not scalable. On average, downloading multi-temporal imagery for a single tile covering an area of $\approx 2km^2$ can take up to four minutes.

Knowing these constraints, we automate this process by building an application that mimics the actions on the screen required to find a region of interest and download each temporal snapshot. We use `pyautogui` [3], a Python library that controls the mouse and keyboard to emulate interactions with a Graphical User Interface (GUI). Our data collection tool works as follows: we first construct a geospatial grid of the area of interest. Each grid cell has the size of the GEP tiles that provide the desired spatial resolution ($\approx 2km^2$ at $1,600m$ above the ground). Using the centroid's coordinates for each grid cell, our tool inputs the location to GEP and downloads the capture. Then, our tool navigates the time slider iteratively to collect tiles across time until all the grid cells are covered. All tiles are $3840x2160$ pixels at a sub-meter resolution.

To improve the scalability of our data collection process, we deployed multiple instances of our tool and covered numerous areas of interest. Our tool increases the data collection speed 100x over manual effort. So far, we have more than 5,124 multi-temporal tiles, covering an area of $\approx 10,248km^2$ in Kenya.

3.2 Image Pre-processing

The quality of satellite imagery can be affected by natural and unnatural artifacts; the most common artifacts are caused by the atmosphere. The atmosphere scatters sunlight in all directions, producing images with different distortions. In addition, cloud cover and environmental conditions can occlude land observations, making detecting structures challenging. Figure 1(a) illustrates different kinds of natural artifacts that arise in imagery. These natural artifacts include shadows produced by clouds, fog presumably occurring during sand storms in desert areas, and the occlusion caused by clouds. Another set of artifacts occurs due to satellite glitches or electronic noise. Figure 1(b) shows different artificial artifacts, such as partially blurred and gray-scale images and distortions affecting color accuracy and brightness.

While these artifacts are often challenging to eliminate without sacrificing some image quality, there are ways to reduce their negative impact on the training step. For example, image filtering can flag and remove samples affected by the abovementioned artifacts.

Artifact	Frequency (Proportion)
Gray-scale	1395 (3.6%)
Cloud cover	3785 (9.8%)
Blur	5578 (14.5%)

Table 2: Number of images (and proportion) with extreme artifacts in the initial training set (38,304 patches). The most common artifact detected is blurred images, mainly due to low-quality sensors in the early 2000s. The final training set contains 27,546 patches.

For this work, we use the variation of the Laplacian filter as a mechanism to detect artifacts associated with blur [38]. A Laplacian is a differential operator defined in equation 1 where f represents an image. An in-focus image tends to have discontinuities observed with a Laplacian filter. A blur-free image has high spatial frequency content, which causes the edges of objects in the image to become sharp and clear. In contrast, an out-of-focus image will have low spatial frequency content, which will cause the edges of objects to become blurry and indistinct. By calculating the variance of the Laplacian filter, we can estimate how spread the high-frequency content is and determine if the image is blurry. We estimate a variance threshold (0.0003) based on the imagery collected to discard samples that suffer from low variability and are likely to be blurry.

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{1}$$

Images with cloud cover are detected by measuring the number of white pixels. The pixel is likely white if high pixel values are encountered in all the color channels (RGB). Then, we calculate the proportion of white pixels in the entire image and filter out examples with more than 40% white pixels. Gray-scale images have the same pixel values in all three RGB channels, so we apply a logical AND operator between the channels to identify such cases. We use data augmentation techniques explained in the following section to mitigate the impact of brightness and color accuracy differences. Table 2 shows the number of extreme artifacts encountered in the training dataset where the main category was blurred images. In the pre-processing step, we aim to remove extreme artifacts that are not useful in the building segmentation process and add noise to the training data.

3.3 Ground Sample Distance

Ground Sample Distance (GSD) refers to the geometric separation of pixels on the ground, which is specific to the instrument and

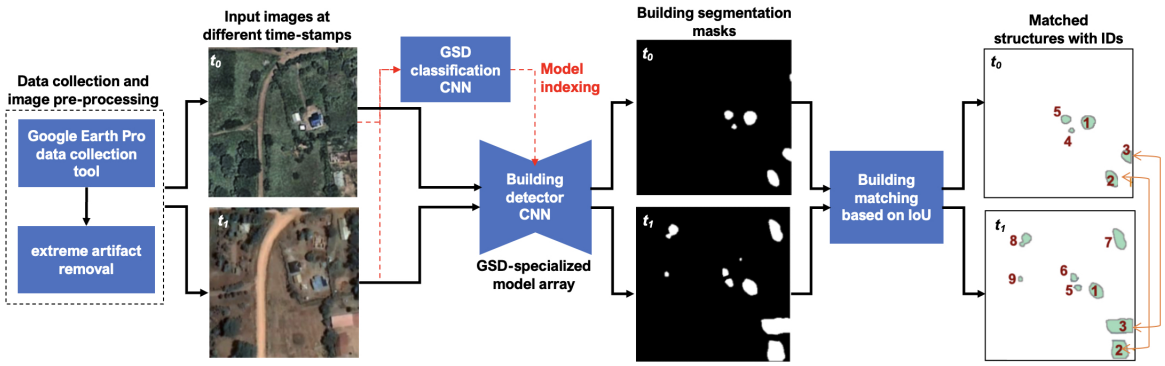

Figure 2: Approach to estimate longitudinal structure change. The top row illustrates the initial timestamp, and the bottom row a subsequent period. (a) Data collection is performed using an automatic data collection tool and extreme artifacts are removed in the pre-processing step. Each input image is fed to a Ground Sample Distance (GSD) classification CNN to identify its GSD, which is used to select the appropriate GSD-specialized building detector. (b) At each period t the corresponding RGB imagery is fed into the segmentation model to produce a binary building prediction mask. (c) A unique identifier is assigned to each building. Overlapping structures in subsequent periods receive the same identifier.

technology of the satellite used to obtain an image. Multiple sensors with different GSDs may have captured the exact location in multi-temporal applications. Therefore, due to the limited availability of labels at various periods, longitudinal building segmentation models pre-trained on a single GSD may overlook this feature. This can affect performance during inference if the model does not learn from images of different scales.

Our framework includes a GSD-aware model to address the challenge of identifying the sensor resolution of an image. This model indexes a specialized building segmentation model tailored to the specific GSD. We trained this model using supervised CNN techniques and natural color satellite images from DigitalGlobe in Kenya. DigitalGlobe offers a range of imagery products with varying processing levels and geolocational accuracy. Each product comes with support files containing metadata about the respective GSD. For example, WorldView-2 and GeoEye-1 have a 50cm GSD, QuickBird-2 has a 65cm GSD, and WorldView-3 has a 124cm GSD. We used this metadata as ground truth to train a CNN multi-classifier with a RestNet18 encoder. Ensemble learning, which combines models for better performance, was not used due to computational cost.

To increase the classifier's robustness, we expanded the ground truth classes by synthetically generating various GSD classes from the original imagery using bilinear sampling. We trained a classifier of 10 different GSDs (50cm, 65cm, 80cm, 100cm, 124cm, 150cm, 175cm, 200cm, 250cm, and 300cm), which covers most of the GSD range available in GEP. We trained our classifier using more than 200,000 patches of 256x256 pixels each distributed across different classes. 70% of the patches were used for training and 30% for testing. In Section 4 we discuss the performance of this model.

3.4 Building Detection and Tracking

Our longitudinal monitoring of structure growth model boils down to identifying the presence of structures in time $t = 1$ and evaluating the changes at $t = n$. We build a set of GSD-specific supervised learning models that aim to perform semantic segmentation of buildings at each t, assign structure identifiers, and propagate them across time by assigning the same identifier to each building that overlaps with an intersection-over-union (IoU) ratio greater than 0.5. Finally, we quantify the changes in the number of identifiers at each timestamp to estimate growth.

We used our data collection tool to gather imagery to develop accurate segmentation models, which we then paired with building footprint annotations from Google's publicly available dataset. This dataset covers 64% of the African continent and over 514 million structures. While Microsoft Open Buildings and Open Street Maps also offer building footprints for some areas in Kenya, we found that the annotations were lower quality and fewer than the Google dataset. It is possible that the annotations we used were produced using a model trained on the same satellite imagery we gathered, which could explain any differences in accuracy. Additionally, we matched building annotations with images taken within six months of the release of the building footprints since the annotations were not multi-temporal.

Our segmentation model was developed using over 27,000 patches, each with 256x256 pixels, distributed across various GSDs. For training purposes, 70% of the patches were utilized, while the remaining 30% were used for testing. Furthermore, we initiated the model with pre-trained weights from ImageNet.

Our approach to estimating longitudinal structure changes in rural areas is summarized in Figure 2. We use a similar method to the baseline algorithm presented in the SpaceNet7 challenge [18]. Our GSD classifier receives imagery at different timestamps t and sends the input image to the corresponding building segmentation model, which is based on the GSD. The segmentation model then produces binary building prediction masks, which are converted to building instances with unique identifiers. To ensure that multiple instances of the same building are associated over time, structures in the exact location are assigned the same identifier.

4 EVALUATION

GSD classifier. We use a pre-trained deep residual learning model, ResNet18 [24], to train a 10-class supervised model. We modify

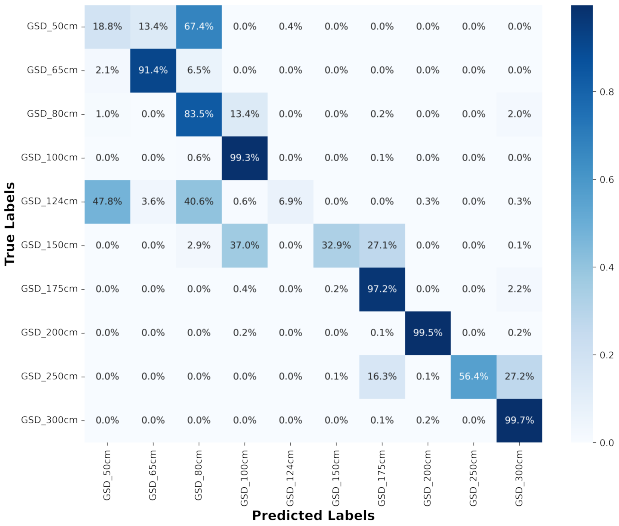

Figure 3: Confusion matrix for the GSD classifier's inference on images with GSD jitter. The overall accuracy for this sample is 71%, whereas for images without jitter (with the same GSD as the labels), the accuracy is 95%.

the last fully connected layer to match the number of classes and fine-tune the model by training it with a learning rate of $1e-4$ and a batch size of 16. We train and evaluate the GSD classifier on GSD-augmented images using bilinear resampling. Finally, We assess the model's performance on a balanced dataset, achieving a 95.6% classification accuracy.

To test the robustness of the classifier for images with different GSDs from the 10 pre-defined, we created augmented images with different GSDs from DigitalGlobe products for evaluation only by adding a uniformly distributed jitter of +/- 5cm for each pre-defined GSD. Since GSD is a continuous variable, we expected the classifier to assign the closest GSD class to each image. The results are shown in a confusion matrix in Figure 3. This experiment shows that the model correctly classified images most of the time, even with GSD jitter. This resulted in an accuracy of 71.3%, with 95% of predictions falling within one class of the correct label. It is possible that the model performs poorly in the $124cm$ category due to the type of area being imaged. We noticed that most of the images in this category have dense forests in very rural areas, which makes it difficult for the model to learn the spatial features of the GSD. On the other hand, the $64cm$ category performs best, as it contains more desert areas, highlighting structures and making estimating the GSD easier. Nonetheless, our current performance provides sufficient confidence in our GSD predictor.

Building detector. We trained GSD-specialized segmentation models using all the training images collected through GEP and paired with Google Building Footprint masks. The GSD classifier indexes the GSD-specialized building detectors. We perform ablation on different segmentation models and evaluate using Intersection-Over-Union (IoU) for the predicted building segmentation mask and the Mean Squared Error (MSE) and Mean Percentage Error (MPE) for the predicted number of buildings (See Table 3). We vary different hyperparameters such as learning rates, batch size, and patch size

Model	Encoder	Loss	BS	IoU	MSE	MPE
UNet	ResNet34	Jaccard	8	0.453	10.284	-17.794
UNet	ResNet34	Dice	8	0.468	13.524	-17.648
UNet	ResNet34	Dice	16	0.470	8.60	-16.203
UNet	ResNet34	Jaccard	16	**0.474**	**6.567**	**-7.185**
DeepLabV3+	ResNet34	Jaccard	16	0.368	9.651	-18.942
PSPNet	ResNet34	Jaccard	16	0.243	19.296	-41.533
PAN	ResNet34	Jaccard	16	0.282	14.89	-40.008
UNet	ResNet50	Jaccard	8	0.469	11.993	-22.889
UNet	ResNet101	Jaccard	32	0.472	10.489	-17.826
UNet	EfficientNet-b5	Jaccard	16	0.355	15.48	-27.15

Table 3: Summary of quantitative results for different combinations of architectures, encoders, and hyperparameters. Intersection-over-Union (IoU), Mean Squared Error (MSE), Mean Percentage Error (MPE), and Percentage Area were used as evaluation metrics. The best results are in bold.

and evaluate different loss functions and encoders (ResNets and efficientNet-b5 [49]).

For semantic segmentation tasks, Jaccard and Dice loss are commonly used loss functions. Given a vector of ground truth y_i and a vector of predicted labels $\hat{y_i}$, the Jaccard index of class c, also called Intersection-Over-Union (IoU) score is defined as in Equation 2 and its respective loss in Equation 3. Dice loss is presented in Equation 4, where i represents an example in the dataset.

$$J_c(y_i, \hat{y_i}) = \frac{|y_i = c \cap \hat{y_i} = c|}{|y_i = c \cup \hat{y_i} = c|} \tag{2}$$

$$\Delta J_c(y_i, \hat{y_i}) = 1 - J_c \tag{3}$$

$$D(y_i, \hat{y_i}) = 1 - \frac{2y_i\hat{y_i} + 1}{y_i + \hat{y_i} + 1} \tag{4}$$

The best-performing model was trained using UNet with a ResNet-34 encoder, Jaccard loss, and a batch size of 16. Our best result (0.474 IoU) is close to the IoU threshold of 0.5 defined by the SpaceNet challenges [50], even with the shortcomings related to detecting small objects in rural areas, particularly with older, less clear imagery.

GSD-aware detector. Table 4 shows the results of training different building detectors for each GSD (GSD-aware) in the dataset, indexed by our GSD classifier. We compare these results to using a single detector trained with all GSDs but without the input from the GSD classifier (GSD-unaware). We tested the performance of these models on test sets of GEP images classified at different GSDs. GSD-aware detector outperformed the generic model in 2 out of 3 resolutions, with up to 11.2% improvement.

Although the difference in IoU was negligible for a GSD of $65cm$, it's worth noting that the training data used for the GSD-aware models was only a fraction of the total training data available. Specifically, we only used data with GSD specific to each GSD-aware detector.

We also look at the performance of our model at different periods. As discussed in Section 3.2, some artificial artifacts can occur during sensing. Due to differences in sensing instruments across time, we observe that many of these artifacts occurred with older images when less sophisticated sensors were used during data acquisition.

GSD	IoU GSD-aware	IoU GSD-unaware	sample size
50 cm	**0.475 (0.20)**	0.427(0.147)	2171
65 cm	0.39(0.17)	**0.40 (0.18)**	793
80 cm	**0.366 (0.191)**	0.35(0.197)	2546

Table 4: Performance comparison of GSD-aware vs GSD-unaware segmentation models. GSD-aware outperforms the unaware model for 2 out of 3 GSD values, showing improvements of up to 11.2%. GSD-aware models use a fraction of the training data required to train the unaware model.

Figure 4: Types of data augmentation used to emulate changes in brightness, color accuracy, and blur that are present in imagery collected at different points in time.

Since our training dataset is static in time to match the release date of the building footprints, we emulate such artifacts using data augmentation techniques such as random brightness, color jitter, and blur. These emulated and mild artifacts aim to improve the model's robustness, unlike the extreme artifacts from the original dataset that were removed during pre-processing. Figure 4 illustrates the type of augmentation used to improve the performance of our model over time.

To evaluate the impact of our data augmentation and GSD awareness, we manually annotate buildings in a sequential set of tiles belonging to the same location and perform the inference step using models trained with and without augmentation. Figure 5 shows that older images (with worse sensing instruments) perform worse than recent ones. On the other hand, we can observe that data augmentation helps improve performance for all years and across different models by up to $\approx 7.8\%$. Moreover, the data augmentation on the GSD-aware model shows better performance in most of the periods even though the training set has fewer samples. This highlights the importance of performing data augmentation for our study in developing regions, where the quality of satellite imagery might not be consistent over time.

Figure 6 illustrates the qualitative results of our best model for images with different densities of structures. Places with a high density of buildings represent a challenge for individual detection since the segmentation map tends to merge buildings that are significantly close to each other. For low-density locations, the model can segment individual buildings and show robustness in the presence of other objects, such as backyards, trees, and vehicles.

5 CASE STUDY

Our tool for longitudinal structure tracking can be used for land management, urban planning, and disaster relief in developing settings. In this section, we present a case study that uses the inference of longitudinal structure estimation. We combine open-source datasets to understand correlations between changes in the built environment in rural areas and electrification.

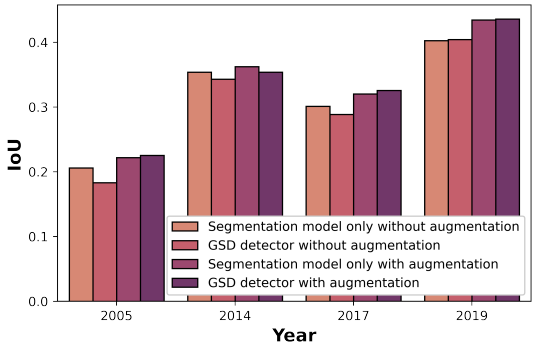

Figure 5: Performance of multiple segmentation models. The GSD-aware segmentation model performs better when data augmentation is used during training. We trained using 2019 images for consistency with building annotations release data and saw a performance improvement of up to 7.8%. Overall, data augmentation improves performance across time and model types.

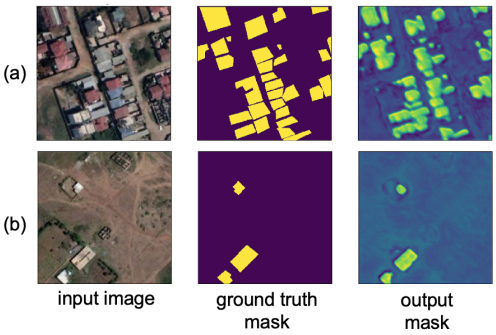

Figure 6: Qualitative results of our initial segmentation model. The figure illustrates performance for input images with (a) high and (b) low structure density.

5.1 Datasets

To perform statistical analysis, we consolidate the following datasets in a common grid with grid cells of $1km^2$ and then added the structure counts from our model:

Multi-dimensional Poverty Index(MPI) and population density. A widely used and publicly-available source of gridded population estimates is WorldPop [2], which provides annual population count estimates from 2000 to 2020 at a resolution of 3 and 30 arc-seconds (100m and 1km at the equator, respectively) globally. WorldPop estimates the population using various models that leverage census data and a stack of covariates. Specifically, we use the datasets for Kenya generated with a top-down unconstrained estimation modelling approach [48]. We use the same grid to consolidate the remaining datasets. WorldPop also provides development and health indicators with the same grid cell resolution as discussed above. We use the proportion of residents living in MPI-defined poverty for each grid cell.

Urban-rural catchment areas. This dataset provides different categories of grid cells located in urban centers based on their sizes and travel time when the cells are located in rural areas. The dataset is global and has 30 arc-second (1km) spatial resolution [8].

Variable mean (SD)	Treatment	Control
Population 2009	189.50(55.87)	189.51(55.81)
URCA	12.71(3.043)	12.71(3.039)
MPI	0.404(0.118)	0.404(0.117)

Table 5: Comparison of summary statistics between treatment and control groups after matching 1,000 randomly selected samples. The matching algorithm considers population, Urban-Rural Catchment Areas (URCA), and Multidimensional Poverty Index (MPI).

Transformer locations and commissioning dates. Our ground-truth data of grid electricity access comprises the geographic location of distribution transformers and minigrids in Kenya. The national power utility provided transformer locations. This dataset includes latitude and longitude, date of commissioning, and power capacity in kVA units for the more than $57k$ transformers in Kenya. Dates of commissioning span from 1966 to 2017.

5.2 Grid Electricity and Structure Growth

We are especially interested in identifying correlations between building changes and access to grid electricity in rural communities. These correlations are particularly useful when changes in the building stock can be used as an explanatory variable for causality studies. To understand this correlation, we use a difference-in-differences analysis, a quasi-experimental identification strategy for estimating effects that predate an intervention [14]. In our setting, we define the intervention as when a community gets its initial distribution transformer. Due to intrinsic biases in electrification policies, estimating causal effects requires access to randomized experiments and additional details of the communities but proving causality is out of the scope of this work. Nonetheless, quantifying the changes in the built environment in a large variety of communities in the periods immediately after grid expansion can help planners better understand the range of structure growth experienced previously across communities and perhaps prepare coordinated infrastructure provisions (including water, sanitation, health, and school services) in future settings.

Using the historical electrification data, we define a controlled study where we compare if there is a difference in how the built environment changes in places that received the intervention (treatment group). We define our treatment group as areas (grid cells of 30 arc-seconds) electrified after 2009 and the untreated group (control) as the places that did not experience the intervention. These groups were identified using the location and commissioning date of distribution transformers. We randomly sampled 1,000 grid cells with population densities between 100 and 300 inhabitants to comply with the World Bank's definition of rural areas [1]. To find the corresponding grid cells for the control group, we use a nearest-neighbor matching approach (k=1). We match each grid cell in the treatment group with the cell that does not belong to the treatment group but has the closest Euclidean distance based on population density, multidimensional poverty index, and catchment area. These two groups are described in Table 5.

Using our semantic model, we estimate the number of structures over time for each grid cell and define the difference-in-differences model as follows:

	coef	std err	t	P> |t|	[0.025	0.975]
const	40.638	1.924	21.183	0.001	36.881	44.536
treat	-7.849	2.639	-3.2	0.001	-14.201	-3.687
time_treat	44.831	1.918	22.342	0.000	42.123	50.123
did	-14.201	2.874	-4.585	0.001	-18.356	-7.667

Table 6: OLS Regression Results for the difference-in-differences analysis. The did coefficient (-14.201) shows a negative correlation, which indicates that after electrification, there are ≈ 14 (15%) fewer buildings than expected in 2019. The number of observations is 22118.

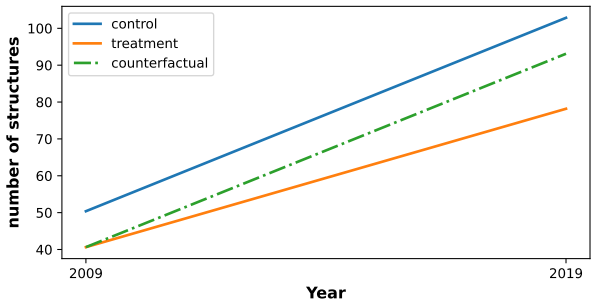

Figure 7: Average number of structures during 2009 and 2019 for places that were electrified after 2009 (treatment group) and places that did not experience the intervention (control). Electrified settlements tend to physically grow at a slower pace (≈ 15% less than expected).

$$Y_s = \alpha + \gamma X_1 + \lambda X_2 + \delta(X_1 X_2) + \varepsilon \tag{5}$$

Where Y_s represents the dependent variable referring to the number of structures, X_1 is a dummy variable representing the treatment (1) and control group (0), X_2 represents a dummy variable indicating treatment pre- (0) and post- (1) intervention (electrification). The regression coefficients α, γ, and λ represent the group means, and ε represents the random error term. The δ coefficient, a key parameter of our analysis, indicates how much the average number of structures in the treatment group has changed during the period after the treatment, compared to what would happen to the same group if the intervention did not occur.

Table 6 summarizes the results for the regression. The regression coefficients `const`(α), `treat`(γ), `time_treat`(λ), and `did`(δ) are statistically significant so it is possible to reject the null hypothesis. As we can observe, the `did`(δ) coefficient indicates the number of structures negatively correlates with electrification. On average, places that did not receive the intervention have ≈ 14 structures more than electrified places. This result is initially non-intuitive - one may theorize that electricity access would attract more residents and encourage more structure construction. Below, we discuss some ideas as to why this occurs.

Figure 7 illustrates the average differences and the counterfactual (what would have happened to Y_s if the intervention did not happen). We can observe that the slope of the treatment group is lower than the counterfactual, indicating the negative correlation observed in table 6. Even though we are not implying causality in this study, one hypothesis about the negative correlation is that electrification occurs way after the structures have been established,

so small changes in the built environment are observed. Geographical cost factors (how distant settlements are from the main grid) may have had a major role in the growth of the energy infrastructure, which advantages wealthy regions. This circumstance may demonstrate some socioeconomic biases in providing additional services to energy infrastructure (health, education, financial, and so on). Another hypothesis is that electricity grid access may not significantly affect decisions about where populations relocate. We leave the evaluation of these hypotheses to future work. This example shows the utility of our approach, particularly concerning growth in structures – seeding new hypotheses, quantifying some aspects of the efficacy of infrastructure investments, and testing the validity of long-held assumptions.

6 CONCLUSIONS

High-resolution satellite imagery can transform the monitoring of rural regions, including buildings and infrastructure, for governments, NGOs, investors, and private organizations. Current methods for monitoring development – usually involving infrequent, inaccurate, and insufficient surveys – are often unreliable, outdated, or nonexistent. Satellite imagery can provide a more accurate and up-to-date picture of development patterns, enabling tracking of changes in dynamic settings. This information is crucial for making informed decisions about allocating scarce investment resources and promoting economic and livelihood growth.

In this work, we have developed techniques to improve the performance of segmentation tasks for high-resolution multi-temporal imagery, which often suffers from variation due to different sensors, imbalanced datasets, and potentially unavailable metadata. We built a pipeline with an automatic imagery collection tool to obtain high-resolution imagery in Kenya and used a CNN-based model to detect ground sample distance, segment buildings and analyze structure changes. We quantify the enhanced performance from our custom data augmentation and ground sample distance metadata inference techniques, particularly for multi-temporal applications. Last, we present an application in which our tool can be used to provide policy-relevant insights at country-scale for planning new electricity infrastructure systems. In general, the ability to accurately detect buildings from satellite imagery enables tracking of changes in the density of the built environment, land use, and other structural features. We believe that this approach can lead to developing explanatory indicators for socioeconomic analysis, natural risk assessment, and policymaker tools to boost development in emerging economies, particularly in underdeveloped rural areas.

REFERENCES

[1] 2022. *How do we define cities, towns, and rural areas?* Retrieved Feb 28, 2022 from https://blogs.worldbank.org/sustainablecities/how-do-we-define-cities-towns-and-rural-areas

[2] 2022. *Open Spatial Demographic Data and Research.* Retrieved Jul 7, 2022 from https://www.worldpop.org/

[3] 2022. *PyAutoGUI documentation.* https://pyautogui.readthedocs.io/en/latest/

[4] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, Berlin, Heidelberg.

[5] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. 2020. LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery. https://doi.org/10.48550/ARXIV.2005.02264

[6] Gabriel Cadamuro, Aggrey Muhebwa, and Jay Taneja. 2018. Assigning a Grade: Accurate Measurement of Road Quality Using Satellite Imagery. In *NIPS 2018 Workshop on Machine Learning for the Developing World. Workshop paper.*

[7] Roberto Castello, Simon Roquette, Martin Esguerra, Adrian Guerra, and Jean-Louis Scartezzini. 2019. Deep learning in the built environment: automatic detection of rooftop solar panels using Convolutional Neural Networks. *Journal of Physics: Conference Series* 1343, 1 (Nov. 2019), 012034. https://doi.org/10.1088/1742-6596/1343/1/012034 Publisher: IOP Publishing.

[8] Andrea Cattaneo, Andrew Nelson, and Theresa McMenomy. 2021. Global mapping of urban–rural catchment areas reveals unequal access to services. *Proc. of the National Academy of Sciences* 118, 2 (2021), e2011990118. https://doi.org/10.1073/pnas.2011990118 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2011990118

[9] Hao Chen and Zhenwei Shi. 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing* 12, 10 (Jan. 2020), 1662. https://doi.org/10.3390/rs12101662 Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. https://doi.org/10.48550/ARXIV.1802.02611

[11] Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. of the IEEE* 105, 10 (2017), 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[13] Santiago Correa, Zeal Shah, Yuezi Wu, Simon Kohlhase, Philippe Raisin, Nabin Raj Gaihre, Vijay Modi, and Jay Taneja. 2022. PowerScour: Tracking Electrified Settlements Using Satellite Data. In *Proc. of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (Boston, Massachusetts) *(BuildSys '22).* ACM, New York, NY, USA, 139–148. https://doi.org/10.1145/3563357.3564069

[14] Scott Cunningham. 2021. *Causal Inference: The Mixtape.* 572 pages. https://mixtape.scunning.com/index.html

[15] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. 2018. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* 172–17209. https://doi.org/10.1109/CVPRW.2018.00031 arXiv:1805.06561 [cs].

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* abs/2010.11929 (2020). arXiv:2010.11929 https://arxiv.org/abs/2010.11929

[17] Thomas R. Dougherty, Tianyuan Huang, Yirong Chen, Rishee K. Jain, and Ram Rajagopal. 2021. SCHMEAR: scalable construction of holistic models for energy analysis from rooftops. In *Proc. of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '21).* ACM, New York, NY, USA, 111–120. https://doi.org/10.1145/3486611.3486666

[18] Adam Van Etten, Daniel Hogan, Jesus Martinez-Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. 2021. The Multi-Temporal Urban Development SpaceNet Dataset. In *CVPR.*

[19] Simone Fobi, Joel Mugyenyi, Nathaniel J. Williams, Vijay Modi, and Jay Taneja. 2022. Predicting Levels of Household Electricity Consumption in Low-Access Settings. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* 3902–3911.

[20] Krishna Karthik Gadiraju, Bharathkumar Ramachandra, Zexi Chen, and Ranga Raju Vatsavai. 2020. Multimodal Deep Learning Based Crop Classification Using Multispectral and Multitemporal Satellite Imagery. In *Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20).* ACM, New York, NY, USA, 3234–3242. https://doi.org/10.1145/3394486.3403375

[21] Tilottama Ghosh, Sharolyn J. Anderson, Christopher D. Elvidge, and Paul C. Sutton. 2013. Using Nighttime Satellite Imagery as a Proxy Measure of Human Well-Being. *Sustainability* 5, 12 (2013), 4988–5019. https://doi.org/10.3390/su5124988

[22] Tilottama Ghosh, Sharolyn J. Anderson, Christopher D. Elvidge, and Paul C. Sutton. 2013. Using Nighttime Satellite Imagery as a Proxy Measure of Human Well-Being. *Sustainability* 5, 12 (Dec. 2013), 4988–5019. https://doi.org/10.3390/su5124988 Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[23] Sungwon Han, Donghyun Ahn, Sungwon Park, Jeasurk Yang, Susang Lee, Jihee Kim, Hyunjoo Yang, Sangyoon Park, and Meeyoung Cha. 2020. Learning to Score Economic Development from Satellite Imagery. In *Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20).* ACM, New York, NY, USA, 2970–2979. https://doi.org/10.1145/3394486.3403347

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

[25] Michael R. Heffels and Joaquin Vanschoren. 2020. Aerial Imagery Pixel-level Segmentation. *arXiv:2012.02024 [cs]* (Dec. 2020). http://arxiv.org/abs/2012.02024

arXiv: 2012.02024 version: 1.

[26] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 7 (2019), 2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242

[27] Xin Hou, Biao Wang, Wanqi Hu, Lei Yin, and Haishan Wu. 2019. SolarNet: A Deep Learning Framework to Map Solar Power Plants In China From Satellite Imagery. https://doi.org/10.48550/arXiv.1912.03685 arXiv:1912.03685 [cs, eess].

[28] Wei Hu, Kyle Bradbury, Jordan M. Malof, Boning Li, Bohao Huang, Artem Streltsov, K. Sydny Fujita, and Ben Hoen. 2022. What you get is not always what you see: pitfalls in solar array assessment using overhead imagery. https://doi.org/10.48550/arXiv.1902.10895 arXiv:1902.10895 [cs].

[29] Zhaojian Huang, Thushini Mendis, and Shen Xu. 2019. Urban solar utilization potential mapping via deep learning technology: A case study of Wuhan, China. *Applied Energy* 250 (Sept. 2019), 283–291. https://doi.org/10.1016/j.apenergy.2019.04.113

[30] Nevrez Imamoglu, Motoki Kimura, Hiroki Miyamoto, Aito Fujita, and Ryosuke Nakamura. 2017. Solar Power Plant Detection on Multi-Spectral Satellite Imagery using Weakly-Supervised CNN with Feedback Features and m-PCNN Fusion. *Proc. of the British Machine Vision Conference 2017* (2017). https://doi.org/10.5244/c.31.183

[31] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794. https://doi.org/10.1126/science.aaf7894 arXiv:https://science.sciencemag.org/content/353/6301/790.full.pdf

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

[33] Dan M. Kluger, Sherrie Wang, and David B. Lobell. 2021. Two shifts for crop mapping: Leveraging aggregate crop statistics to improve satellite-based maps in new regions. *Remote Sensing of Environment* 262 (sep 2021), 112488. https://doi.org/10.1016/j.rse.2021.112488

[34] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. 2018. Pyramid Attention Network for Semantic Segmentation. https://doi.org/10.48550/ARXIV.1805.10180

[35] Rui Liu, Li Mi, and Zhenzhong Chen. 2021. AFNet: Adaptive Fusion Network for Remote Sensing Image Semantic Segmentation. *IEEE Trans. on Geoscience and Remote Sensing* 59, 9 (Sept. 2021), 7871–7886. https://doi.org/10.1109/TGRS.2020.3034123 Conference Name: IEEE Trans. on Geoscience and Remote Sensing.

[36] Barak Oshri, Annie Hu, Peter Adelson, Xiao Chen, Pascaline Dupas, Jeremy Weinstein, Marshall Burke, David Lobell, and Stefano Ermon. 2018. Infrastructure Quality Assessment in Africa using Satellite Imagery and Deep Learning. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. ACM, New York, NY, USA, 616–625. https://doi.org/10.1145/3219819.3219924

[37] Maria Papadomanolaki, Maria Vakalopoulou, and Konstantinos Karantzalos. 2021. A Deep Multitask Learning Framework Coupling Semantic Segmentation and Fully Convolutional LSTM Networks for Urban Change Detection. *IEEE Trans. on Geoscience and Remote Sensing* 59, 9 (Sept. 2021), 7651–7668. https://doi.org/10.1109/TGRS.2021.3055584 Conference Name: IEEE Trans. on Geoscience and Remote Sensing.

[38] J.L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia. 2000. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proc. 15th International Conference on Pattern Recognition. ICPR-2000*, Vol. 3. 314–317 vol.3. https://doi.org/10.1109/ICPR.2000.903548

[39] Reid Pryzant, Stefano Ermon, and David Lobell. 2017. Monitoring Ethiopian Wheat Fungus with Satellite Imagery and Deep Feature Learning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Workshop paper.*

[40] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature* 566, 7743 (Feb. 2019), 195–204. https://doi.org/10.1038/s41586-019-0912-1

[41] Fermín Rodríguez, Alice Fleetwood, Ainhoa Galarza, and Luis Fontán. 2018. Predicting solar energy generation through artificial neural networks using weather forecasts for microgrid control. *Renewable Energy* 126 (2018), 855–864. https://doi.org/10.1016/j.renene.2018.03.070

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. https://doi.org/10.48550/ARXIV.1505.04597

[43] Rose Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. 2019. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*

[44] Sumeet Saurav, Prashant Gidde, Sanjay Singh, and Ravi Saini. 2019. Power Line Segmentation in Aerial Images Using Convolutional Neural Networks. In *Pattern Recognition and Machine Intelligence*, Bhabesh Deka, Pradipta Maji, Sushmita Mitra, Dhruba Kumar Bhattacharyya, Prabin Kumar Bora, and Sankar Kumar Pal (Eds.). Springer International Publishing, Cham, 623–632.

[45] Joel Segarra, Maria Luisa Buchaillot, Jose Luis Araus, and Shawn C. Kefauver. 2020. Remote Sensing for Precision Agriculture: Sentinel-2 Improved Features and Applications. *Agronomy* 10, 5 (May 2020), 641. https://doi.org/10.3390/agronomy10050641 Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[46] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. 2021. S2Looking: A Satellite Side-Looking Dataset for Building Change Detection. *Remote Sensing* 13, 24 (dec 2021), 5094. https://doi.org/10.3390/rs13245094

[47] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. 2021. Continental-Scale Building Detection from High Resolution Satellite Imagery. https://doi.org/10.48550/ARXIV.2107.12283

[48] Forrest R. Stevens, Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. 2015. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE* 10, 2 (feb 2015), e0107042. https://doi.org/10.1371/journal.pone.0107042

[49] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2019). https://doi.org/10.48550/ARXIV.1905.11946

[50] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. 2018. SpaceNet: A Remote Sensing Dataset and Challenge Series. https://doi.org/10.48550/ARXIV.1807.01232

[51] Anna X. Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. 2018. Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data. In *Proc. of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '18)*. ACM, New York, NY, USA, Article 50, 5 pages. https://doi.org/10.1145/3209811.3212707

[52] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B. Lobell. 2020. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sensing* 12, 2 (2020).

[53] Yi Yang and Shawn Newsam. 2010. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In *Proc. of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (San Jose, California) *(GIS '10)*. ACM, New York, NY, USA, 270–279. https://doi.org/10.1145/1869790.1869829

[54] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* 8, 3 (2020), 1217–1229. https://doi.org/10.1038/s41467-020-16185-w

[55] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* 11, 1 (May 2020), 2583. https://doi.org/10.1038/s41467-020-16185-w Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computer science;Economics Subject_term_id: computer-science;economics.

[56] Bailang Yu, Kaifang Shi, Yingjie Hu, Chang Huang, Zuoqi Chen, and Jianping Wu. 2015. Poverty Evaluation Using NPP-VIIRS Nighttime Light Composite Data at the County Level in China. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8, 3 (2015), 1217–1229. https://doi.org/10.1109/JSTARS.2015.2399416

[57] Liangpei Zhang, Lefei Zhang, and Bo Du. 2016. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geoscience and Remote Sensing Magazine* 4, 2 (2016), 22–40. https://doi.org/10.1109/MGRS.2016.2540798

[58] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2016. Pyramid Scene Parsing Network. https://doi.org/10.48550/ARXIV.1612.01105

[59] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. 2019. Object Detection with Deep Learning: A Review. http://arxiv.org/abs/1807.05511 arXiv:1807.05511 [cs].

[60] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. 2021. Change is Everywhere: Single-Temporal Supervised Object Change Detection in Remote Sensing Imagery. *arXiv:2108.07002 [cs]* (Aug. 2021). http://arxiv.org/abs/2108.07002 arXiv: 2108.07002.

[61] Sharon Zhou, Jeremy Irvin, Zhecheng Wang, Ram Rajagopal, Andrew Ng, Eva Zhang, Will Deaderick, and Jabs Aljubran. 2019. DeepWind: Weakly Supervised Localization of Wind Turbines in Satellite Imagery. In *NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning*. https://www.climatechange.ai/papers/neurips2019/5

[62] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing* 145 (nov 2018), 197–209. https://doi.org/10.1016/j.isprsjprs.2018.01.004