Postprint

Towards Explaining Satellite Based Poverty Predictions with Convolutional Neural Networks

Hamid Sarmadi*, Thorsteinn Rögnvaldsson*,Nils Roger Carlsson*, Mattias Ohlsson*‡, Ibrahim Wahab†, Ola Hall†

*Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Sweden

{hamid.sarmadi, thorsteinn.rognvaldsson, nils_roger.carlsson, mattias.ohlsson}@hh.se

†Department of Human Geography, Lund University, Sweden

{ibrahim.wahab, ola.hall}@keg.lu.se

‡Centre for Environmental and Climate Science,

Lund University, Sweden

Abstract—Deep convolutional neural networks (CNNs) have been shown to predict poverty and development indicators from satellite images with surprising accuracy. This paper presents a first attempt at analyzing the CNNs responses in detail and explaining the basis for the predictions. The CNN model, while trained on relatively low resolution day- and night-time satellite images, is able to outperform human subjects who look at high-resolution images in ranking the Wealth Index categories. Multiple explainability experiments performed on the model indicate the importance of the sizes of the objects, pixel colors in the image, and provide a visualization of the importance of different structures in input images. A visualization is also provided of type images that maximize the network prediction of Wealth Index, which provides clues on what the CNN prediction is based on.

Index Terms—Poverty prediction, Deep Convolutional Neural Networks, Satellite Images, Explainable AI

I. INTRODUCTION

Measuring poverty with remote sensing data has been a lively research field since the introduction of nighttime lights in the late 1990's [1]. In a seminal study, Jean et al. [2] [3] extended this work and showed how to fine-tune a deep Convolutional Neural Network (CNN), pre-trained on the ImageNet data, to predict poverty levels from high-resolution daytime satellite images in sub-Saharan Africa. This paper has been followed by many studies, see [4] for a review. However, despite the successes of this approach, the results have yet to be employed for policy decisions. A recent review [5] showed that little effort has been devoted to explaining why CNNs make such good predictions of poverty levels, which may be a reason why there is hesitation to use these predictions in applications.

We show that high accuracy wealth estimates can be achieved with lower resolution satellite images, with an accuracy higher than if human experts are asked to rate the wealth level from high-resolution satellite images. In addition, several approaches are explored to explain what features or details in the images that the CNNs base their predictions on.

II. PREVIOUS WORK

Since 2016, there have been several applications of machine learning to poverty prediction. Many have used the transfer learning approach on high-resolution images and sometimes there have been attempts at interpreting the model, i.e. exploring what parts of the image that the model reacts to. Jean et al. [2] [3] show that the CNN filters react to structures in the landscape, with more reactions in urban areas than in rural areas. An analysis of if and how these filters contribute to the Wealth Index predicted value is not provided. Liu et al. [6] do gradient visualization with guided back-propagation in order to see what features give high night-time light intensity prediction. They conclude that for cities, these are contours of roads and areas of buildings are important. Tan et al. [7] use low-resolution satellite images and multitask learning for the transfer learning, i.e. more tasks than night-time lights. They show some examples of CNN filter activations for the fine-tuned network, but no analysis of how these filter activations relate to the wealth prediction. A general comment on the transfer learning features are that the illustrated filter responses may have been pretty much the same with a CNN trained on the ImageNet data, before fine tuning and without any connection to poverty.

Pandey et al. [8] train a CNN directly from high-resolution satellite images using multitask learning with labels from the 2011 Census of India. These tasks are, e.g., roof cladding material, source of drinking water, and source of lighting (a total of 24 such tasks). They illustrate, with some examples, that the CNN filters react to edges and seem to segment the images into farmland or human settlements, but there is no analysis of how these features relate to the Wealth Index prediction.

In short, published peer-reviewed work on explaining or interpreting CNN models' responses have been anecdotal, illustrating with a few filter responses but never analyzing how the filter responses relate to predicted Wealth Index values.

A different approach to using machine learning and high-resolution satellite images for wealth prediction have used features, or objects, extracted from the satellite images and

then algorithms like random forest used to model the Wealth Index from the extracted features. Ayush et al. [9] fine-tune a YOLOv3 model [10] to detect objects in the satellite images. They model the Wealth Index with gradient boosting trees and apply SHAP analysis [11] to the final model. Their conclusion is that trucks and other vehicles are the most important features; the number of trucks in the image being by far the most important feature. Huang et al. [12] train a roof detector and claim that changes in wealth can be detected through changes in roof material. Zhao et al. [13] feed several features, computed from satellite images and other sources, into a random forest model and conclude that distance to closest urban area and night-time light radiance are the two most important features for wealth prediction. Thus, the conclusions from the feature based wealth prediction vary and depend on the researchers' decision on which features to feed into the model. The feature based studies have not provided consistent clues on what the direct CNN-based Wealth Index prediction is based on.

Tang et al. [14] use the low-resolution normalized difference vegetation index (NDVI) as input. The CNN model is trained using the same transfer learning trick as used by Jean et al. [3]. They then use random forest for estimating the Wealth Index from the CNN features, concluding that NDVI-based predictions are as good as the predictions achieved with daytime satellite images. This shows that color is important but not what structure or other features that are important.

Finally, the current state-of-the-art on predicting wealth from satellite images appears to be the recent work by Lee and Braithwaite [15], who use both high-resolution daytime satellite images as well as OpenStreetMap data. Their approach is to iterate between a feature based prediction and the satellite image based prediction and thus achieve improved predictions. We have not used their model as the starting point for our study since their code is not available and there are many details in their approach that are hard to replicate. For example, they compute their own International Wealth Index, they manually correct survey site locations, and they compute several complicated features from OpenStreetMap data.

III. DATA

Our study is based on three data sources: satellite data, both day-time imagery and night-time data, and survey data on household status. All data are for countries in sub-Saharan Africa.

A. Household survey data

The Demographic and Health Surveys (DHS) program is a key source of data for researchers and policymakers interested in understanding health, population, and social dynamics in low- and middle-income countries. DHS surveys collect nationally representative data on a wide range of health and social indicators, including assets wealth. DHS data are publicly available and can be accessed through the DHS program website upon request.

The surveys are cross-sectional, they cannot establish causality or track individual-level changes over time. The methodology is however standardized, ensuring that data is collected using consistent and rigorous procedures across all countries, making it possible to compare data across countries and over time, and thus facilitating cross-country analyses and comparisons. Surveys are conducted 3-5 years apart, sometimes with much longer intervals. Survey waves are uncoordinated between countries resulting in temporally heterogeneous data. To protect the privacy and confidentiality of survey participants, DHS surveys scramble the GPS coordinates of the households and clusters by a small, random amount. This process, known as "geographic displacement", ensures that the precise location of the households and clusters are anonymized.

The larger dataset contains a total of 64,864 clusters across 34 Sub-Saharan Africa countries, spanning the period between 1990 and 2019. A subset of this dataset was drawn from the 2015-16 Tanzania DHS dataset which covers 12,563 households, aggregated into 608 clusters across the 30 administrative regions of the country. Our human evaluators were tested on this subset data.

As Wealth Index value, we use the mean of the Wealth Index factor score combined, with five decimals This is in line with previous work, e.g. Head et al. [16].

B. Satellite images

Three sets of satellite data were used: Sentinel-2 daytime images for the CNN network, Visible Infrared Imaging Radiometer Suite (VIIRS) night-time images to get night-time light intensities, and high-resolution Google Maps images for humans to view.

Sentinel-2 and night-time light images were downloaded from Google Earth Engine in GEOTIFF format. Night-time light images came from the VIIRS Stray Light Corrected Nighttime Day/Night Band Composites Version 1, with a resolution of 750 meters per pixel, and daytime images came from Sentinel-2 MSI: MultiSpectral Instrument, Level-1C, bands B2, B3, and B4, with a resolution of 10 meters per pixel. Both daytime and night-time images were downloaded for the year 2016, daytime images were retrieved with cloud filtering percentile p45.

Sentinel-2 daytime images were split up into sub-images of 224×224 pixels, to fit as input to the MobileNet-V2 network. This corresponds to areas of 2240×2240 square meters. The VIIRS night-time light images were split up into matching sub-images of 3×3 pixels, corresponding to 2250×2250 square meters. The summed light intensity over the 3×3 pixel sub-image was used as the night-time light value for the corresponding daytime satellite image. If the light intensity for a VIIRS pixel was below 0.5 then it was set to zero, based on an estimate of the noise level. All images with night-time light values ≥ 1 and a random selection of images with night-time light values < 1 were used for the CNN training.

Three datasets of Sentinel-2 and VIIRS data were collected, with increasing sizes. The smallest, denoted "TZA", consists

of images from Tanzania only. This data set has 325,785 images; of which 323,966 (99.4%) have night-time light values < 1. The medium sized one, denoted "TZA + 11 cities", consists of images from Tanzania plus images from 11 cities in neighboring countries. This data set has 333,221 images; of which 330,198 (99.1%) have night-time light values < 1. The largest data set, denoted "TZA + surroundings", consists of images from Tanzania plus images from neighboring countries. This data set has 2,030,657 images; of which 1,848,356 (91.0%) have night-time light values < 1. The group of images with night-time light < 1 were heavily downsampled for all CNN trainings. Typically, for a CNN training session, 58% of the images would have night-time light < 1.

The Google Maps Platform hosts a set of APIs through which developers can retrieve data from the platform. Images are from different sensors and combined into a mosaic of images taken over multiple periods. Some of the sensors used are Landsat 8, Pleiades 1A, Quickbird and WorldView 4. For the study with human evaluators, 608 images were downloaded at zoom level 18, which corresponds to a resolution of about 0.6 meters per pixel. These images covered 608 Tanzania DHS sites, but with the location corrected to adjust for the random displacement in the DHS data. The correction was based on a guess of the correct location.

C. Human evaluation data

The main inclusion criteria for our human experts is to be development researchers, broadly defined, with some field-work experience in any region of SSA. The human evaluators were shown a batch of high-resolution satellite imagery corresponding to each DHS cluster (village or city) location in a sequential manner and asked to rate the level of welfare on a five-grade scale ranging from *Poorest*, *Poorer*, *Middle*, *Richer*, to *Richest*; akin to the DHS wealth quintiles. The median rating for each cluster image was then taken as representing the relative poverty score for the cluster. Raters were allowed to rate additional batches of 30 images if they so wished. Overall, we received 2,174 ratings from 102 experts. As can be seen from Table 1, a significant majority of our human evaluators were males, with more than half of the sample being between 36 and 55 years old and a further 18 per cent being even older. In terms of educational attainment, half of the human evaluators have a Ph.D. as the highest education level attained. With regards to the main region of fieldwork experience in SSA, they are fairly well-distributed - 47 per cent for West Africa, 36 per cent for East Africa, including Tanzania, and 17 per cent for the Central and Southern Africa region. For the length of experience, the sample has substantial field experience in these regions, with exactly half of the sample having more than 10 years of fieldwork experience. A summary description of our human experts is presented in Table I.

Variable Description		Proportion(%)
Sex	Females	22.5
	Males	75.5
	Non-binary	2.0
Age Distribution	Up to 35 years old	30.4
	Between 36 and 55 years old	52.0
	56 years and older	17.6
Highest Educational Attainment	Bachelor's Degree	16.7
	Master's Degree	33.3
	Ph.D.	50.0
Region of Experience	East Africa, incl. Tanzania	36.3
	Central and Southern Africa	16.7
	West Africa	47.1
Length of Experience	Between 1 and 10 years	50
	Between 11 and 20 years	30.4
	21 years and above	19.6

TABLE I: Summary Description of Human Experts (N=102)

IV. METHODS

A. Modeling the Wealth Index

MobileNet-V2 was taken as starting point for the model. It is a 53 layer deep CNN pre-trained on the ImageNet database. It was trained to model night-time lights in a two-stage process, first adjusting only the weights in the final layer to the output, and then fine-tuned by also adjusting weights earlier in the network. The first step was done for 20 epochs, and the fine-tune training was done for 20 epochs. The learning rate was set to 0.01 in the first part, using Adagrad training, and then the learning rate was reduced to 0.001 during fine-tuning, using L2 regularisation with parameter 0.1. The batch size was set to 100. The MobileNet-V2 takes the 224×224 daylight images as input and is trained to output the corresponding $\log(1 + $ night-time light).

Once the MobileNet-V2 network is trained, the values from the last layer before output (1280 units) are used as features for a linear regression model of the Wealth Index. Ridge Regression is used to fit the linear regression, using a cross-validation approach to set the ridge parameter. The Ridge Regression is done several times with different training and validation data to gauge the stability of the solution. All predictions of Wealth Index are done on images that are not similar to images in the training and validation sets, and the predictions are repeated so that a prediction is done for each of the DHS survey sites.

The accuracies of the Wealth Index estimations are evaluated using three measures: the coefficient of determination R^2, the Spearman's rank correlation, and the Matthew's correlation coefficient when grouping into quintiles.

The R^2 coefficient is standard in studies of wealth estimation and allows the results to be compared to earlier published results. Spearman's rank correlation is the correlation between two ranked estimates of the Wealth Index; it varies between -1 and $+1$, with $+1$ indicating perfect agreement between the two rankings. The quintile grouping is used specifically to compare against human estimations of wealth level, since it is probably easier and more reliable for a human to group places into wealth groups rather than to provide a full ranking of all the sites. The correctness of the groupings are measured

using the Matthew's correlation coefficient (MCC) for each dichotomy of one group versus the others.

B. XAI methods

Several methods were employed to explore what features the CNN based its prediction on. None of the XAI methods required any new training of the CNN network, but the ridge regression step was always repeated for modified images.

1) Grid Shuffling: The input image is divided into equal tiles using a square grid pattern, and then tiles are randomly shuffled in the image, breaking up structures that are larger than the tile sizes [17], [18]. Tile sizes range from one pixel to half the width of the image, and when tiles are 1×1 pixels all structural information in the image is lost after shuffling and only color information is left. The aim is to find the grid size when the Wealth Index prediction starts to deteriorate, indicating the size of structures in the image that the prediction is based on.

2) Frequency Domain Filtering: This consists of low-pass, high-pass, and band-pass filters. Filtering is done in the image frequency space and is done individually for each color in the image. A low-pass filter removes high frequencies and keeps the lower frequency information. A high-pass filter is the complement of the low-pass filter and keeps higher-frequency information. A band-pass filter keeps information between two frequency thresholds in the image. We use Gaussian filters throughout for frequency domain filtering, with a signal domain standard deviation equivalent of σ for the low- and high-pass filters. The band-pass filter has a lower limit standard deviation of σ and a higher limit standard deviation of 1.5σ in signal space. The frequency domain filtering affects the color of the image but should be able to confirm what scales of features in the images that are important for the prediction.

3) L*a*b* Color Space Analysis: The L*a*b* color space is designed to correspond well to human perception [19]. In this color space, the lightness (L* component) is decoupled from the chromaticity of the color (a* and b* components), which allows separating objects based on material, regardless on the lightness variations. We used the L*a*b color space to cluster the image pixels based on chromaticities, and segment out different types of materials in the image. By keeping the chromaticities of one type (cluster) of material and removing those of other types, we can compare how chromaticities (e.g. different materials) affect the Wealth Index estimation.

4) Feature Attribution Methods: These are methods that aim to compute the importance that each part of the image has for the CNN output. We used Occlusion Sensitivity [20], Grad-CAM [21], Guided Backpropagation [22], and Guided Grad-CAM [21].

Occlusion Sensitivity shows the change in the output if a square patch in the image is grayed out. The change in output is calculated for locations on a grid and the result is visualized as a map that should show the parts of the image that are important for the output, and if they have a positive or a negative effect on the output.

Grad-CAM is a gradient-based method that uses the last convolutional layer to show a high level map of the importance of image regions on the output. Because the last convolutional layer is used, the resolution of the importance map is low, but it contains high level (as opposed to pixel level) information regarding the CNN output.

Guided backpropagation is also a gradient based method, but it estimates the importance of each pixel for the CNN output. It creates a high resolution map (the same size as the input image) with pixel-level information about the CNN output. Since Grad-CAM and guided backpropagation are complementary regarding the level of information they provide, they have been combined by multiplying the two maps with each other. This combination is denoted Guided Grad-CAM.

5) Feature Visualization: This is a recent method designed to show what patterns in the image different neurons in the CNN react to [23]. The method starts from an image with randomly assigned pixel values and then the image is optimized to maximize the output of a particular selected neuron in the CNN. Since the optimization problem certainly has many local maxima, the visualization will show an image that maximizes the output locally. The resulting *optimal* image can provide hints on the type of image features that specific neurons respond to, but it is not guaranteed that the method produces a meaningful and interpretable visualization.

V. RESULTS

A. Wealth Index estimation

The first study was done to check if the use of low-resolution satellite images and the transfer learning approach would yield Wealth Index predictions that were as accurate as those with high-resolution images.

Figure 1 shows the training and validation errors during the CNN training on night-time lights values. It shows clearly that the CNN learns to predict the night-time lights and that the fine-tuning step is important. Table II summarizes the Wealth Index prediction accuracies when the 1280 CNN features are used as input in the ridge regression model, for three different training set sizes for the night-time lights estimation and two ways to produce the features used in the Ridge Regression. The three different training sets are "TZA", "TZA + 11 cities", and "TZA + surroundings" and correspond to increasing number of images (see Section III). The two ways for producing the features are denoted "1×1" and "3×3", and correspond to using the features from one single input image of size 224×224, or the averaged features from nine images in a 3×3 grid centered on the DHS location (see Section IV). Note that the Wealth Index estimation itself is always done on the 608 DHS survey sites in Tanzania, regardless of the size of the training data used for the night-time light estimation.

Table II shows that increasing data set sizes lead to improved estimates of the Wealth Index, although more so for the R^2 coefficient than for the rank correlation. Also, combining more images yields better estimates than using a single image. All this is in agreement with previously published findings.
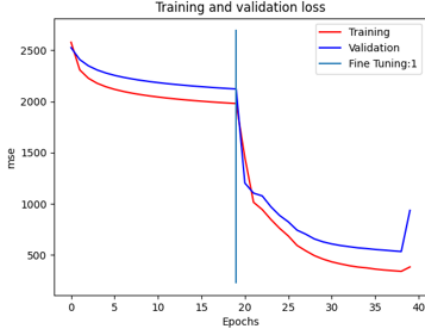
Fig. 1: Training and validation loss during the training of the CNN for night-time light prediction. The vertical bar indicates the start of the fine-tuning.

TABLE II: TZA Wealth Index Estimation From Low Resolution Images On the 608 DHS Sites

	Training data		
	TZA	TZA + 11 cities	TZA + surroundings
1×1 (R^2)	0.57	0.60	0.61
3×3 (R^2)	0.65	0.66	0.67
1×1 (rank corr.)	0.70	0.72	0.73
3×3 (rank corr.)	0.76	0.76	0.77

Jean et al. report an R^2 value of 0.57 when using the CNN transfer method to model the Wealth Index for Tanzania (2010 DHS survey data) [3]. In 2017, Head et al. replicated some of their experiments, albeit not for Tanzania, and concluded that they got pretty much the same values [16], with variations in the second decimal. Thus, the results in Table II are on par with results using higher resolution images.

The second study was done to gauge whether it mattered if survey data came from the same period as the satellite image data, or from times before or after. Here, the CNN network was always the same and the only thing that changed was the DHS data used in the ridge regression. The large DHS data set was split up into three parts: Tanzania survey results with Wealth Index for DHS survey phases 4 and 5, which meant years 2003 to 2008 (809 DHS sites), Tanzania survey results with Wealth Index for DHS survey phase 6, which meant years 2009 to 2012 (1,028 DHS sites), and Tanzania survey results with Wealth Index values for DHS survey phase 7, which meant years 2015 to 2017 (1,044 DHS sites). The three data subsets were used either as test or training sets (or a k-fold cross-validation if the time periods were the same). Table III shows the results of this experiment, always using 3×3 images, with R^2 values in the upper half and rank correlations in the lower half. The results are better for the later period, which agrees with the time for the satellite images, but the differences are small. This indicates that the features used in the prediction do not change very much over a time period of 5-20 years, or that the Wealth Index values do not change much relative to each other over this time period.

The third study compared the prediction accuracy of the

TABLE III: TZA Wealth Index Estimation Across Time

	DHS Test data		
DHS Training data	2003-2008	2009-2012	2015-2017
2003-2008 (R^2)	0.69	0.69	0.69
2009-2012 (R^2)	0.70	0.70	0.71
2015-2017 (R^2)	0.67	0.70	0.73
2003-2008 (rank corr.)	0.74	0.75	0.78
2009-2012 (rank corr.)	0.75	0.76	0.78
2015-2017 (rank corr.)	0.75	0.77	0.79

low-resolution CNN model, combining 3×3 low-resolution images, to that of human evaluators that viewed high-resolution images. This study was done on corrected DHS data, meaning that the location of the DHS survey data had been adjusted to the position believed to be the correct DHS survey position for the cluster. The CNN Wealth Index predictions where grouped into quintiles, from poorest to wealthiest. The human subjects were similarly asked to group the site they saw into one of five categories, from poorest to wealthiest.

The two confusion matrices below show how the Wealth Index predictions from the CNN and from the human evaluators compare to the quintile grouping based on the actual DHS survey result. Group 1 corresponds to the 20% poorest, and group 5 corresponds to the 20% wealthiest. An observation is that the human subjects label many more sites to be average, and fewer sites as very poor or very wealthy. Furthermore, in the average group 3, the human evaluators placed sites from all five wealth levels, almost with equal probabilities. This indicates that it is difficult for a human to estimate wealth level from a satellite image.

		Predicted group (CNN)				
		1	2	3	4	5
True group	1	63	41	11	7	0
	2	25	44	29	13	0
	3	18	27	42	29	6
	4	5	8	28	51	29
	5	1	1	12	21	87

		Predicted group (hum.)				
		1	2	3	4	5
True group	1	12	64	36	9	1
	2	13	54	47	5	2
	3	10	54	40	17	1
	4	6	37	53	25	0
	5	4	24	49	40	5

The correctness of the CNN and human groupings against the true DHS values were compared using four dichotomies in increasing wealth level: (1) whether a site was grouped into

TABLE IV: Human and CNN Quintile Groupings of TZA Wealth Index

MCC vals.	1 vs. 2-5	1-2 vs. 3-5	1-3 vs. 4-5	1-4 vs. 5
CNN	0.17	0.31	0.38	0.36
Human	0.03	0.09	0.13	0.16

group 1, instead of any of groups 2-5, (2) whether a site was grouped into group 1 or 2, instead of any of groups 3-5, (3) whether a site was grouped into any of groups 1-3, instead of groups 4-5, or (4) whether a site was grouped into groups 1-4, instead of group 5. For each dichotomy, the Matthew's correlation coefficient (MCC) was computed.

Table IV shows the MCC values for the two approaches. The CNN predicted wealth levels are consistently better than the human estimates from high-resolution images. Thus, the CNN Wealth Index predictor is producing something quite impressive from the low-resolution images and there should be a significant value in trying to understand the predictions.

Notably, the CNN model had less accuracy on the wealth indices when using corrected DHS positions instead of the original DHS positions, which puts into question if the new positions were more correct than the original positions.

B. Explorations of explanations

Here, we present results with the XAI methods described in Section IV-B. For practical reasons, when results are reported for a 3×3 input it means that the nine images have been concatenated into a 672×672 pixel image and fed into a fully-convolutional network created by fusing nine CNNs. This is different from previous results where predictions were done using feature values averaged over nine CNNs. The XAI work was done on the CNN trained on the TZA dataset, cf. Table II.

1) Grid Shuffling: The results from doing Grid Shuffling are shown in Figure 2. Wealth Index prediction performance is shown versus tile size (section length). The performance curve for the 3×3 input is consistently above the curve for the 1×1 input. This is expected since the results with averaged image features (i.e. combining images) are better in general. It is, however, surprising that the curves begin decreasing at different scales; 1×1 starts decreasing at 10 pixels, whereas 3×3 starts decreasing at approximately 60 pixels.

2) Frequency Domain Filtering: The effects of feeding low-pass, high-pass, or band-pass filtered images to the CNN are shown in Figure 3. The x-axis label is

$$\sigma = \frac{N}{2\pi D}$$

where D is the standard deviation of the Gaussian filter in frequency space and $N = 224$.

The low-pass filter leads to a general decrease in performance as σ increases. It is only for $\sigma \leq 2$ that the prediction performance is acceptable, which means almost no filtering. The minimum performance for low-pass filtered images is considerably lower than for the Grid Shuffling, which is surprising but could be related to how color is affected by the filtering

(the colors become more and more gray). Similarly, the high-pass filter appears to not affect the Wealth Index prediction very much before it comes to $\sigma \leq 2$, which corresponds to keeping only the highest frequencies in the image, i.e. objects 1-2 pixels in size, and discarding the rest. Together, the high-pass and low-pass results indicate that almost all scales in the image are important, but high frequency components are much more important than low frequency components.

In the band-pass filtered images details at a certain frequency (size) are kept in the image while others are grayed out. The Wealth Index prediction performance with band-pass filtered images maximizes around 2-3 pixels, and the performance is almost as good as with unfiltered images. This indicates that objects of sizes around 2-3 pixels (20-30 meters) are most important.

3) L*a*b* Color Space Analysis: The colors for each pixel for each one of the 608 DHS Sentinel-2 images were converted from RGB into L*a*b* space. Then k-means clustering was applied on the pixel distribution, with the intention to find meaningful concentrations of pixels in the L*a*b* space. The number of clusters (k) was set to 3 using the elbow method.

To measure the importance of each color cluster, the colors for pixels in each cluster were then modified in two different ways, one where the color of pixels in two clusters was removed but the lightness kept intact, and one where both color and lightness in two clusters were removed. In the first type of experiment, the chromaticities of pixels in two clusters were removed by setting their a* and b* components to zero, and the modified image was then fed to the CNN. In the second type of experiment, the pixels for two clusters were "blanked out" by setting them to an average gray color (setting the R, G, and B components to the middle of their valid range).

The results for these experiments are summarized in Figures 4 and 5. The colors of the bars are the cluster centers in L*a*b* space. One cluster is centered on dark orange/light brown color that could be related to dry areas. Another cluster is centered on a dark brown color that could be related to mountainous areas, buildings, or other types of infrastructure. The third cluster is centered on a bluish green color that could be related to green areas and bodies of water. Keeping the dark orange/light brown cluster intact but removing color for the two other clusters changes the performance very little. Graying the pixels completely always has a larger effect than just removing color, but the effect is surprisingly small.

4) Feature Attribution Methods: Figure 6 shows the results from applying Grad-CAM, Guided Backpropagation, Guided Grad-CAM, and Occlusion Sensitivity to some sample 1×1 images in the 608 DHS data set (each image is centered on a DHS survey location). Grad-CAM has higher outputs for samples with larger Wealth Index values but because of the low resolution it is not very helpful in giving detailed information on how the network is making the decision. It sometimes lights up in places where there are buildings or man-made infrastructure, but it is not consistent. Furthermore, in one case where the image is in the ocean Grad-CAM highlights several regions, which is hard to interpret (the image
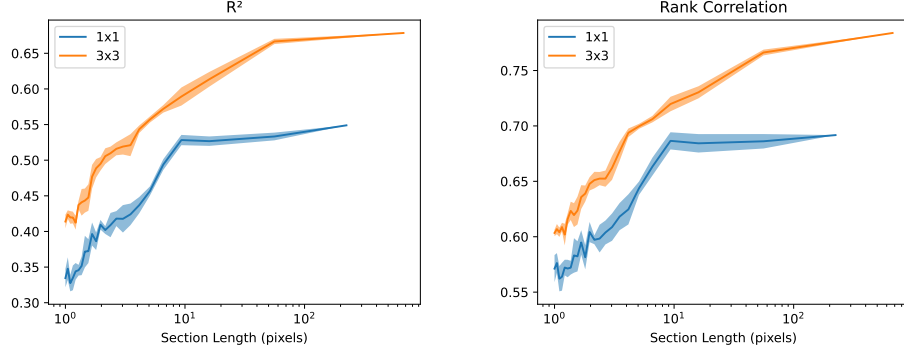
Fig. 2: R^2 and Spearman's Rank Correlation performance for Wealth Index prediction when input images are split up in tiles and tiles are shuffled. Results for 3×3 inputs are shown in orange and results for 1×1 inputs are shown in blue. Five random shuffling experiments were performed for each tile size, and the curves show the mean and standard deviation.
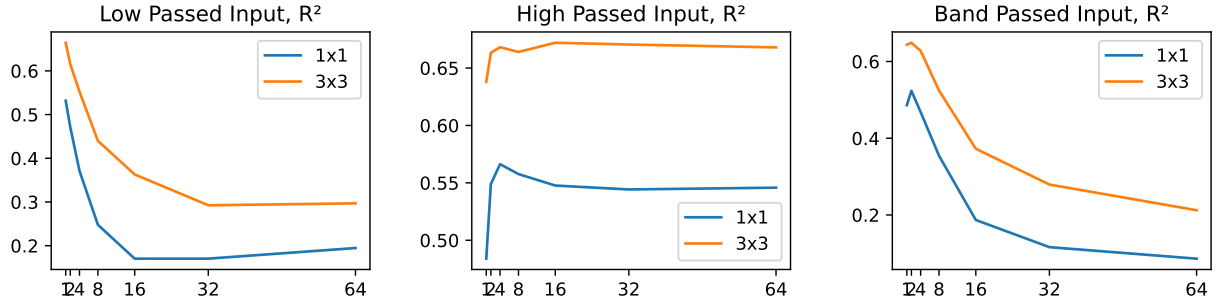

Fig. 3: Wealth Index prediction performance of the CNN when provided with low-pass, high-pass, and band-pass filtered images as input. The x-axis indicates the standard deviation of the Gaussian filter, expressed in pixels. Only results for R^2 are shown to save space.

TABLE V: Correlation Between CNN Output and Summed Feature Attributes

Correlation	Grad-CAM	Guided BP	Guided Grad-CAM
CNN output	0.99	0.79	0.89

shows the ocean due to the random displacement applied to the true DHS survey location). Of the remaining Feature Attribution Methods, Guided Grad-CAM seems to be the most consistent in detecting human infrastructure and not reacting to blue ocean. However, none of the methods provide a very convincing basis for explaining the output value.

A test was made to see how well the features from Grad-CAM, Guided Backpropagation, and Guided Grad-CAM matched with the CNN output. The sum of the responses in the images was computed and correlated with the CNN output value. The correlations are shown in Table V. The correlation between Grad-CAM and CNN output is very high, but Grad-CAM is computed for the last convolutional layer so this should be expected. However, the correlation between the sum of Guided Grad-CAM and the CNN output is also high, higher than for Guided Backpropagation, which indicates that the features highlighted for Guided Grad-CAM explain a lot.

5) Feature Visualization: The result from applying the Feature Visualization technique on the output neuron that predicts the Wealth Index is shown in Figure 7. The three visualizations correspond to three different initial random images. This shows what type of input image that would maximize the Wealth Index output, and it is not hard to imagine that it looks like a populated place with roads, streets, houses and neighborhoods.

VI. DISCUSSION

This study was initiated to understand what it is in daytime satellite images that a CNN reacts to and enables such good predictions of poverty status, better than estimates made by a human evaluator from high-resolution satellite images. Previous analyses of, e.g., filter responses in the CNN have been anecdotal, providing a few example images but never considering how the features affect the complete set of predictions.

The studies with Grid Shuffling indicate that details 100-600 meters in size are important. The Gaussian frequency filter experiments indicate that small details, 20-30 meters in size, are important and that large details (low frequencies) are not informative. Low-pass filtering the image destroys the Wealth Index prediction. Color only is not sufficient, since
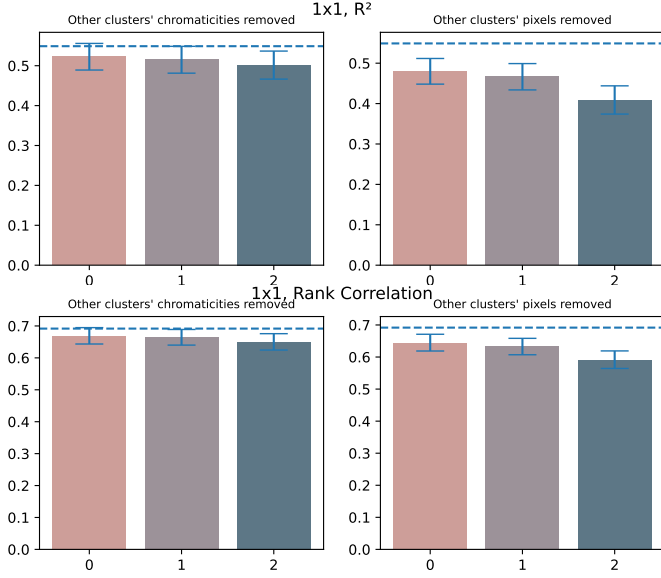
Fig. 4: Performance bar plots for experiments on the L*a*b* space pixel clusters on 1×1 images. Two types of experiments were done for each cluster: first, removing the chromaticity of other clusters' pixels, and second, graying the pixel values of other clusters' pixels. The height is the average performance and the error bars are the standard deviation, estimated using bootstrapping. The horizontal dashed line corresponds to the CNN original performance.
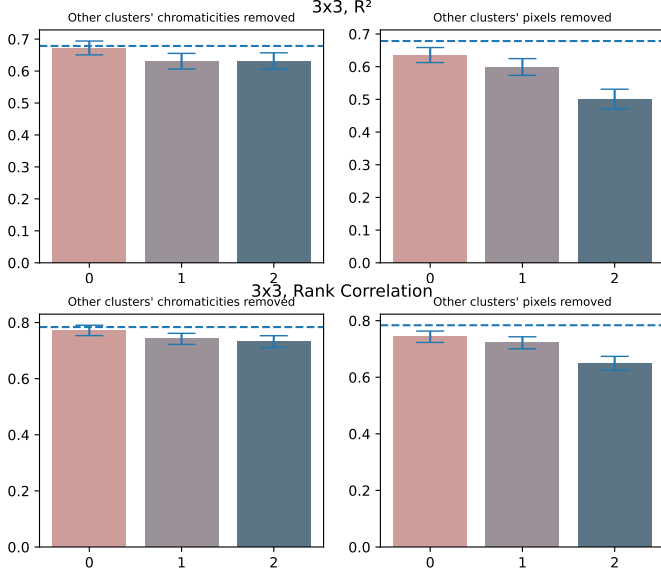

Fig. 5: Performance bar plots for experiments on the L*a*b* space pixel clusters on 3×3 images. Two types of experiments were done for each cluster: first, removing the chromaticity of other clusters' pixels, and second, graying the pixel values of other clusters' pixels. The height is the average performance and the error bars are the standard deviation, estimated using bootstrapping. The horizontal dashed line corresponds to the CNN original performance.

shuffling the individual pixels retains color but the Wealth Index prediction is quite poor.

Both the Grid Shuffling and the Gaussian filtering experiments show that combining 3×3 images instead of just the single 1×1 images is always beneficial. This was not immediately obvious since one would think that objects of a certain size should be useful in wealth index estimation regardless of the input image size. However, there are at least a couple of reasons why combining images would be helpful. First, due to the (small) random displacement of the correct DHS locations, using neighboring images to the stated DHS location increases the probability of one of the images covering the correct survey site. The second relates to the statistical probability for covering important information. If the Wealth Index prediction is based on the number of small infrastructure features found in an image, then having more images increases the probability for detecting enough of these small features.

The L*a*b* color experiments show that the Wealth Index prediction can be done well even if the color and intensity of many pixels are set to an average gray, and just one color cluster is kept. This result is in line with earlier results by Tang et al. [14] who showed that it is possible to predict the Wealth Index with quite high accuracy ($R^2 = 0.60$ for Tanzania) using low-resolution satellite images with the normalized difference vegetation index (which probably is close to the complement of the orange/brown color we get in the L*a*b* clustering). However, the L*a*b* experiments do not support the supposition that roof cladding material is a very important feature. Also, the resolution of the images (10 meters per pixel) probably makes this feature difficult to detect.

At first, the Feature Attribution results seemed not very informative or consistent across samples, but the correlation between the sum of the feature values and the CNN prediction indicate that they describe a lot of the CNN decision. The Grad-CAM, which has the highest correlation with the CNN ouput, has too low resolution to be consistent with our other findings. However, the Guided Grad-CAM looks very promising as a basis for explaining the CNN response. The features it highlights match other findings about size, and it looks like it is basically the number of these features in the image that determines the CNN output.

Finally, the image Feature Visualization of the output neuron gives indications about what an image with a high Wealth Index looks like. The visualizations appear like the street structure of a city and buildings inside it. Main streets and minor streets are shown as bigger and straight, or smaller and curvy lines, respectively. Building structures can be seen as dots. It should be possible to further investigate Feature Visualization by systematically visualizing the features from layers before the output layer.

VII. CONCLUSION

The experiments indicate that a CNN that predicts Wealth Index for Tanzania based on low-resolution satellite images
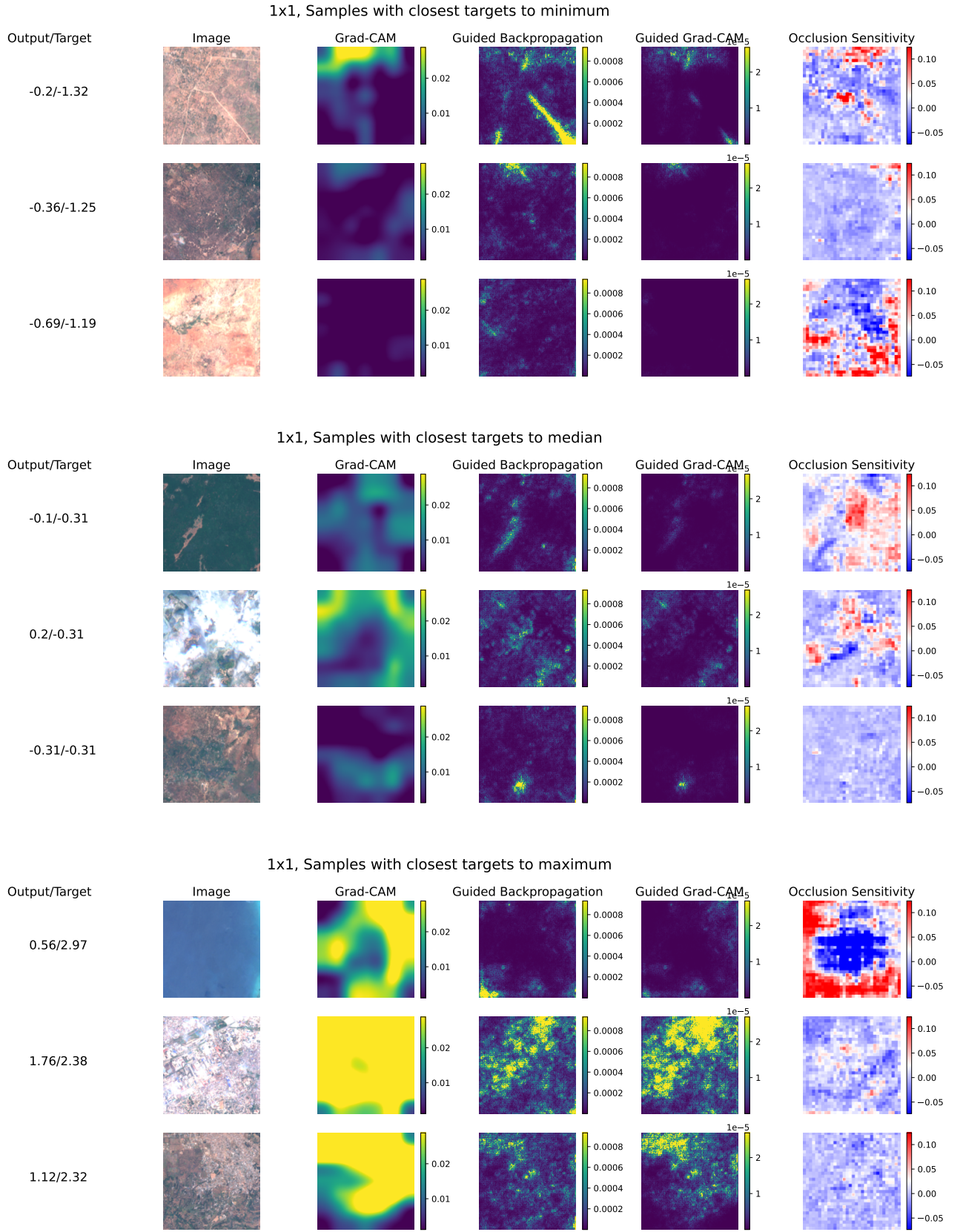
Fig. 6: Feature attribution methods for 1×1 input samples. From the testing set, three samples with the lowest target values, three samples with highest target values, and three samples with target values closes to the median are used for visualizations.
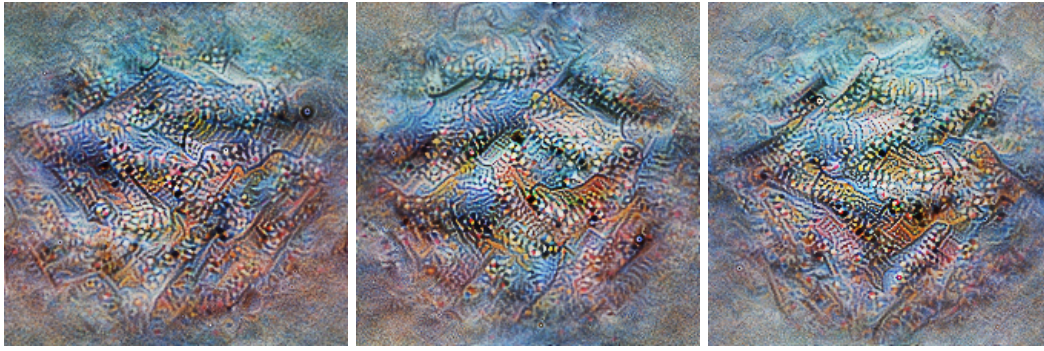
Fig. 7: Feature visualization of the most optimal input image based on the output neuron for our Wealth Index regression neural network. The three examples were created starting from three different random initial images.

forms its prediction from many small details in the image; roads, road networks, and buildings. How many such small details there are in the image appears important, but the exact material in buildings appears to not be important, and neither is the number of vehicles. The resolution in the images used is too low to describe individual vehicles but the Wealth Index prediction is still competitive with that using much higher resolution images, and apparently better than predictions made by human evaluators using high-resolution images.

REFERENCES

[1] C. D. Elvidge, K. E. Baugh, E. A. Kihn, H. W. Kroehl, and E. R. Davis, "Mapping City Lights With Nighttime Data From The DMSP Operational Linescan System," *Photogrammetric Engineering and Remote Sensing*, vol. 63, no. 6, Jun. 1997.

[2] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 3929—3935.

[3] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, pp. 790–794, August 2016.

[4] D. A. Castro and M. A. Álvarez, "Predicting socioeconomic indicators using transfer learning on imagery data: an application in Brazil," *GeoJournal*, 2022.

[5] O. Hall, M. Ohlsson, and T. Rögnvaldsson, "A review of explainable AI in the satellite data, deep machine learning, and human poverty domain," *Patterns*, vol. 3, October 2022.

[6] H. Liu, X. He, Y. Bai, X. Liu, Y. Wu, Y. Zhao, and H. Yang, "Nightlight as a proxy of economic indicators: Fine-grained GDP inference around Chinese mainland via attention-augmented CNN from daytime satellite imagery," *Remote Sensing*, vol. 13, p. 2067, 2021.

[7] Y. Tan, P. Wu, G. Zhou, Y. Li, and B. Bai, "Combining residual neural networks and feature pyramid networks to estimate poverty using multisource remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 553–565, 2020.

[8] S. M. Pandey, T. Agarwal, and N. C. Krishnan, "Multi-task deep learning for predicting poverty from satellite images," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[9] K. Ayush, B. Uzkent, M. Burke, D. B. Lobell, and S. Ermon, "Generating interpretable poverty maps using object detection in satellite images," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, 2020, pp. 4410–4416.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4768—-4777.

[12] L. Y. Huang, S. M. Hsiang, and M. Gonzalez-Navarro, "Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs," National Bureau of Economic Research (NBER), Tech. Rep. 29105, 2021.

[13] X. Zhao, B. Yu, Y. Liu, Z. Chen, Q. Li, C. Wang, and J. Wu, "Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh," *Remote Sensing*, vol. 11, no. 4, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/4/375

[14] B. Tang, Y. Liu, and D. S. Matteson, "Predicting poverty with vegetation index," *Applied Economic Perspectives and Policy*, vol. 44, pp. 930–945, 2022.

[15] K. Lee and J. Braithwaite, "High-resolution poverty maps in Sub-Saharan Africa," *World Development*, vol. 159, p. 106028, 2022.

[16] A. Head, M. Manguin, N. Tran, and J. E. Blumenstock, "Can human development be measured with satellite imagery?" in *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, ser. ICTD '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3136560.3136576

[17] T. Zhang and Z. Zhu, "Interpreting adversarially trained convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97. Proceedings of Machine Learning Research (PMLR), 2019, pp. 7502–7511.

[18] D. Crowder and G. Malik, "Robustness of Humans and Machines on Object Recognition with Extreme Image Transformations," May 2022, arXiv:2205.05167 [cs] version: 2. [Online]. Available: http://arxiv.org/abs/2205.05167

[19] A. R. Robertson, "The CIE 1976 Color-Difference Formulae," *Color Research & Application*, vol. 2, no. 1, pp. 7–11, 1977. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1520-6378.1977.tb00104.x

[20] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, pp. 818–833.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020. [Online]. Available: https://doi.org/10.1007/s11263-019-01228-7

[22] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," Apr. 2015, arXiv:1412.6806 [cs]. [Online]. Available: http://arxiv.org/abs/1412.6806

[23] C. Olah, A. Mordvintsev, and L. Schubert, "Feature Visualization," *Distill*, vol. 2, no. 11, p. e7, Nov. 2017. [Online]. Available: https://distill.pub/2017/feature-visualization