

Learning to Score Economic Development from Satellite Imagery

Sungwon Han

lion4151@kaist.ac.kr

School of Computing, KAIST

Jeasurk Yang

e0508688@u.nus.edu

Department of Geography, National University of Singapore

Hyunjoo Yang

hyang@sogang.ac.kr

School of Economics, Sogang University

Donghyun Ahn

segaukwa@kaist.ac.kr

School of Computing, KAIST

Susang Lee

susang88@kaist.ac.kr

School of Business and Technology Management, KAIST

Sangyoon Park

sangyoon@hku.hk

Faculty of Business and Economics, University of Hong Kong

Sungwon Park

psw0416@kaist.ac.kr

School of Computing, KAIST

Jihee Kim

jiheekim@kaist.ac.kr

School of Business and Technology Management, KAIST

Meeyoung Cha

mcha@ibs.re.kr

Institute for Basic Science & School of Computing, KAIST

ABSTRACT

Reliable and timely measurements of economic activities are fundamental for understanding economic development and designing government policies. However, many developing countries still lack reliable data. In this paper, we introduce a novel approach for measuring economic development from high-resolution satellite images in the absence of ground truth statistics. Our method consists of three steps. First, we run a clustering algorithm on satellite images that distinguishes artifacts from nature (siCluster). Second, we generate a partial order graph of the identified clusters based on the level of economic development, either by human guidance or by low-resolution statistics (siPog). Third, we use a CNN-based sorter that assigns differentiable scores to each satellite grid based on the relative ranks of clusters (siScore). The novelty of our method is that we break down a computationally hard problem into sub-tasks, which involves a human-in-the-loop solution. With the combination of unsupervised learning and the partial orders of dozens of urban vs. rural clusters, our method can estimate the economic development scores of over 10,000 satellite grids consistently with other baseline development proxies (Spearman correlation of 0.851). This efficient method is interpretable and robust; we demonstrate how to apply our method to both developed (e.g., South Korea) and developing economies (e.g., Vietnam and Malawi).

CCS CONCEPTS

- Applied computing → Economics;
- Computing methodologies → Unsupervised learning.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike International 4.0 License.
KDD '20, August 23–27, 2020, Virtual Event, CA, USA
 © 2020 Copyright held by the owner/author(s).
 ACM ISBN 978-1-4503-7998-4/20/08.
<https://doi.org/10.1145/3394486.3403347>

Donghyun Ahn

segaukwa@kaist.ac.kr

School of Computing, KAIST

Susang Lee

susang88@kaist.ac.kr

School of Business and Technology Management, KAIST

Sangyoon Park

sangyoon@hku.hk

Faculty of Business and Economics, University of Hong Kong

KEYWORDS

Satellite imagery; Economic development; Human-in-the-loop; Clustering; Partial order graph

ACM Reference Format:

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 *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3394486.3403347>

1 INTRODUCTION

Collecting data on socioeconomic activities is crucial for designing sound government policies aimed at promoting and sustaining economic development. Not surprisingly, most developed countries and international organizations, such as the World Bank, deploy significant amounts of financial and human resources to conduct censuses and surveys. For instance, there are more than 130 ongoing surveys in the U.S. on demographic and economic activities currently conducted by the Census Bureau.

Unlike rich countries, however, developing countries often lack resources for conducting expensive surveys and suffer from data reliability issues. To solve these limitations and improve the quality of economic measures, researchers from various fields have attempted to use alternative data sources. Remote sensing data can be an attractive source for economic measurements because of its extensive geographic coverage, timeliness, and high granularity. For example, high-resolution satellite images allow social scientists to analyze the local impact of air pollution, changes in land use or crop choice, and fluctuations in retail demand predicted by counting cars [8].

Despite the rapid development of deep learning techniques in analyzing satellite imagery to construct economic development proxies, there remain two considerable limitations to use them in practice. First, the existing classification or regression models cannot be applied to regions without ground truth data, where these techniques are most needed, because the models require a massive amount of labels for training. Second, deep learning remains as a

black-box function. Its lack of interpretability hinders practitioners from applying such techniques to real-world problems.

This paper proposes a novel approach to overcome these limitations. We introduce a model that learns from high-resolution satellite images to *rank relative scores* of economic development. Our deep neural network model clusters images based on visual features and defines paired sets of clusters, i.e., a partial order graph (POG). The POG is an essential element in our approach addressing the limitations of the existing methods: it is an interpretable input to the scoring model that estimates the economic development, which can be generated either by readily available data or by light human annotation.

Our model operates in three stages. The first stage (*siCluster*) uses an entire collection of satellite images of a target country and clusters them by a deep learning-based unsupervised learning and transfer learning. *siCluster* uses labels for the general land cover types, such as rural, urban, and uninhabited. The second stage (*siPog*) builds a POG of the clusters from *siCluster*. The order in a POG captures the relative level of economic development, for which we use urbanization as a comprehensive proxy, following the economics literature [33]. Two different methods to generate a POG are suggested: the clusters are ordered either by humans (*human-guided*) or by data such as population density or nightlight intensity (*data-guided*). Lastly, the final stage (*siScore*) uses the POG from *siPog* to assign a differentiable score, via a CNN-based model. The proposed computational framework to measure sub-district level economic development from satellite imagery without the guide of any partial ground-truth data is novel and shows remarkable performance gain over existing baselines. Codes and implementation details are made available at the project repository.¹

The main contributions of this paper are as follows:

- We propose a novel model to predict economic development in the absence of ground-truth data. The model only requires the generation of a POG. As we demonstrate in this paper, a POG can be constructed easily using either public online resources or light human annotation.
- POGs provide an interpretable explanation of which human activity patterns in satellite imagery depict a more advanced level of economic development.
- Our approach provides new insights on how to decompose a challenging problem into sub-tasks and combine human intelligence with machine intelligence.
- Extensive ablation studies not only show that all components in our model are critical but also present several possible applications of our model.

2 RELATED WORK

Satellite imagery has been frequently used in economic and demographic research. In the past, the nightlight intensity of satellite images was used as a proxy for socioeconomic factors such as power consumption [25], income [6], well-being [13], urban growth and population [3].

Recent studies have begun to exploit daytime images to estimate economic statistics better. For example, Jean et al. [22] present a deep neural network model that combines both daytime and

nighttime images to determine the poverty in sub-Saharan African countries. This model has shown improvement in R^2 value by 0.1 or more, compared to models using only nightlight luminosity for various regions examined. Head et al. [19] confirmed the same model could be “transferred” to broader geographic areas, yet this required significant effort in hyper-parameter tuning. The performance also varied by attribute types. For instance, predicting education level and electricity access reached R^2 scores of 0.47 and 0.69, respectively. However, predicting water access and children’s weight-for-height index remained low at 0.26 and 0.06.

In mapping grid-shaped satellite images to economic data in administrative units, many approaches use sampling (e.g., a random selection of images within an administrative district) to correlate the land cover features with ground truth data. New approaches like [16] propose a lightweight embedding method that utilizes descriptive statistics from the entire images within an area. This approach yields the state-of-the-art prediction for population density with R^2 up to 0.96.

The common assumption behind the existing approaches is that a rich set of traditional socioeconomic data is available for training. However, the assumption may not be met for much of the developing world and even for developed countries for a specific temporal resolution. Some research has examined the potential use of supplemental data to handle the absence of socioeconomic labels. One such example is using mobile phone call detail records (CDR) along with remote sensing data. CDR data contains geo-tagged user behaviors based on either calls or SMS and can even track personal-level social networks. Steele et al. [30] used a mixture model combining both CDR and satellite imagery for poverty mapping. Sheehan et al. [29] examined the frequency of each location mentioned on Wikipedia (e.g., infrastructure, company location) and applied natural language processing (NLP) for economic predictions. This work suggested a multimodal model that combines the feature vector extracted by the Wikipedia article embedding and the vector derived by nightlight imagery.

Inspired by this line of research, we also consider the case when supplemental data is available for training and provide a comparison of outcomes between different methods. Specifically, our model exploits two additional datasets: (1) district-level demographics, which is readily available from governments, and (2) nightlight intensity, which is publicly available with low-resolution quality. More importantly, our model can be run and perform better, even without any socioeconomic data labels for training.

3 MODEL

3.1 Overview

Problem definition: Let $\mathcal{I} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be the set of satellite images for a given area. The main goal of the proposed model f is to compute a score \hat{y}_i for each image \mathbf{x}_i (i.e., $\hat{y}_i = f(\mathbf{x}_i)$) that well represents the economic development level y_i . We assume ground truth values of y_i are unknown at the training phase. Note that this prevents the use of existing methods [16, 22, 30] to predict y_i .

As a solution, we propose a weakly-supervised method to estimate relative scores that highly correlate with the target variable y_i , rather than predicting its absolute values directly. The method consists of three steps: (1) identifying satellite grid clusters of similar

¹https://github.com/dscig/urban_score

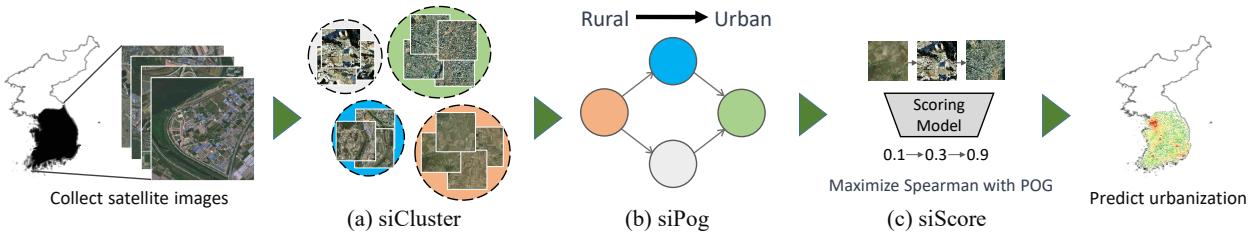


Figure 1: The overall architecture of the proposed model, composed of (a) siCluster for clustering satellite images, (b) siPog for generating partial order graph (POG), and (c) siScore for training the scoring model with POG. The example in (a) shows that the red and the green clusters are rural and urban areas, respectively, and the other two are mixed. The partial order graph would place the most rural clusters on the left and the most urban clusters on the right, as in (b). The step in (c) will assign urbanization scores to individual satellite grids while complying with the POG structure.

urbanization traits via a deep learning architecture – siCluster, (2) generating the POG of the found clusters – siPog, and (3) computing differentiable scores from the learned POG structure – siScore. The intermediate POG structure helps solve the complex problem efficiently and also makes the result interpretable. These steps are visualized in Fig. 1.

3.2 Clustering satellite imagery with siCluster

To generate scores that represent the urbanization level, one needs to know what kinds of human activities capture such values. Satellite grids contain various kinds of geo-spatial information, including vegetation, surface elevation, and human activities. Therefore, satellite images are an excellent source of examining unique patterns in agricultural and industrial zones, such as paddy fields and factories [2]. To distinguish various human activities, we apply the deep learning-based clustering that can efficiently handle the curse of dimensionality problem via hierarchical architectures [14, 24]. This work adopts DeepCluster [5], which is known for its novel end-to-end training process in the computer vision domain.

DeepCluster iteratively updates the CNN-based encoder and its clustering result every epoch. It starts by computing k-means clustering over the encoded vectors of input images. Each cluster is assigned a pseudo-label. The encoder is then updated from the classification process. Weights of the classifier g_θ and encoder h_W are jointly trained by optimizing the following loss (Eq. 1),

$$L_c = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} NLL(g_\theta(h_W(\mathbf{x}_i)), \bar{y}_i), \quad (1)$$

where \bar{y}_i represents the pseudo-labels from k-means clustering, \mathcal{B} denotes the input batch of satellite grids, and NLL is the negative log-softmax function for maximizing log-likelihood.

DeepCluster has two limitations. One is the initial randomness; the model is affected by the initial weights that can propagate through the training process. Another is the lack of consistency in the class assignment; the model relies on the pseudo-labels generated from its k-means clustering, which is subject to noise in data. As a result, DeepCluster is not directly applicable to our problem, and the satellite grids are grouped according to trivial traits such as RGB patterns [5, 23]. Our clustering algorithm, siCluster, builds upon DeepCluster with two new improvements.

Improvement #1 from transfer learning: Most of the land cover on Earth is uninhabited [9, 12] and including such grids will add noise to clustering. Hence, we first aggregated satellite grids into three categories: urban, rural, and uninhabited. This task was done by a pre-trained classifier and transfer learning. We constructed a labeled dataset for transfer learning that includes one thousand satellite images with three labels: urban, rural, and uninhabited.² We then adopted a semi-supervised learning technique to train the classifier over the small set of labels and massive amounts of unlabeled data. The Mean Teacher [31] model, which penalizes the inconsistent predictions between the teacher and student, is used.

For example, in the context of South Korea, one thousand labeled grids and 22,577 unlabeled grid images were combined to train the classification model. The fine-tuned encoder part is re-used as the initial data points of the siCluster model. Repeating the actual clustering process for urban and rural categories produced an equal number of n_c clusters. We did not further cluster the uninhabited category. The siCluster algorithm produced a total of $n_t = 2 \times n_c + 1$ clusters.

Improvement #2 from consistency preserving: We added new loss terms to prevent the model from learning trivial features. Suppose a given satellite grid \mathbf{x}_i and its corresponding encoded vector \mathbf{v}_i , i.e., $\mathbf{v}_i = h_W(\mathbf{x}_i)$. We then augment \mathbf{x}_i via common techniques such as rotation, gray-scale, and flipping that do not deform the original visual context. Let us call the augmented versions $\hat{\mathbf{x}}_i$. Then, the distance between \mathbf{x}_i and its augmentations $\hat{\mathbf{x}}_i$ in the embedding space should be close enough, compared with the distance between \mathbf{x}_i to other data points. We define the *consistency preserving loss* to represent this invariant feature characteristic against data augmentation in the embedding space. siCluster is trained by jointly optimizing the negative log-likelihood loss (Eq. 1) and reducing the Euclidean distance between the input and its augmentations on embedding space (Eq. 2).

$$L_{ecp} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \| h_W(\mathbf{x}_i) - h_W(\hat{\mathbf{x}}_i) \|_2 \quad (2)$$

²One may hire human annotators or use free GIS data sources for labeling. We describe the steps to get coarse-grained labels in the supplementary material.

The final loss to train siCluster is as follows (Eq. 3), and weights of the encoder is updated via back-propagation.

$$L_{clustering} = L_c + L_{ecp} \quad (3)$$

3.3 Generating partial order graph with siPog

Images within each cluster share similar visual contexts that likely represent a similar level of economic development. The second step of the algorithm aims at ordering these identified clusters. The partial order graph (POG) is an efficient representation showing the order across different clusters while ignoring any within-cluster difference. We generated a POG in the order of economic development. Here, development refers to an economic transition from agriculture to manufacturing and service industries, which tend to cluster in more urbanized areas [33]. When two clusters showed a similar level of development, they were placed at the same level without any strict ordering between them, as illustrated in Fig. 1(b). Below are two strategies of siPog.

Human-guided.

We first considered the human-in-the-loop design and asked human annotators to sort clusters manually. Both experts (i.e., professionals in economics or geography holding a master's degree or higher) and laypeople participated in this ordering task. Annotators compared clusters and identified relative orders of clusters by examining the provided grid images. Clusters were ordered and connected as a graph by their presumed economic development level. Cluster pairs whose development levels were judged to be indifferent were placed at the same level within the POG. The strength of this method is its lower cost than the full comparison of images. In this work, a total of ten annotators who had some domain knowledge on the target country created the POG. Most annotators completed the task within 60 minutes.

Data-guided.

POG can also be generated without human guidance. While grid-level demographics are costly to obtain, there are ample resources that can be used as a proxy, such as Internet search results. Proxy data are aggregated at high-level (e.g., city or province) or are not accurate. Nonetheless, they can be used for generating POG. Below we demonstrate two examples.

- *Census data-guided:* One may generate POG based on the widely available census data. Census data for most countries, whether collected by local governments or by third party entities, contain demographic information aggregated over districts or other administrative units that are far larger than the areas covered by each satellite grid. The administrative units may be in an arbitrary polygon shape. We suppose that the census data over each district comprises the aggregate metrics in the corresponding satellite grids. Since similar human activities contribute to the economic development, in the same way, we assigned all the grids that belong to the same cluster the same census records.

Assume that each cluster C_k , composed of a collection of satellite grids, needs to be assigned a score for target metric d_k . Each satellite grid belongs to a single cluster. If the number of grids for district i with membership in cluster C_k is c_k^i and if the census data is \tilde{y}_i

over district i , we can form the following equation:

$$\tilde{y}_i = \sum_{k=1}^{n_t} c_k^i \times d_k. \quad (4)$$

Denoting \mathcal{D} as a set of districts in the target country, we can form a system of linear equations by collecting the equation from each district in \mathcal{D} . To predict the ground-truth scores of district \tilde{y}_i , we can fit a linear regression model solving the optimization problem (Eq. 5). The weight of the regressor becomes the approximation of representative scores of clusters, which is used to define relative ranks.

$$\min_{d_1, \dots, d_{n_t}} \sum_{i \in \mathcal{D}} \left\| \tilde{y}_i - \sum_{k=1}^{n_t} (c_k^i \times d_k) \right\|_2^2 \quad (5)$$

To reduce overfitting, we randomly select 50% of the equations for training the regressor and collect the weights from 100 trials. Two sample t-test at a significance level of 0.05 was used to determine significant differences in weights. If two clusters are sufficiently separable by the t-test, an edge was created to connect them in the partial order graph.

- *Nightlight data-guided:* Another proxy we use is the luminosity intensity measured in nighttime satellite imagery. This publicly free data is only available at the low-resolution quality. We first extrapolate the nighttime images to match the size of the daytime satellite grids. Once we identify all the extrapolated nighttime grids that correspond to each cluster, then nightlight intensity was averaged for each cluster. We perform a two-sample t-test with a threshold 0.01 to detect any significant intensity difference between every two clusters and consequently create an edge between them if two clusters are sufficiently comparable.

3.4 Computing scores with siScore

Now given the relative orders of clusters in the POG, the next task is to assign a score between 0 and 1 to every cluster using the CNN-based scoring model f . This model automatically detects which features from satellite imagery (belonging to clusters) determine the urbanization score via supervised learning. This scoring is identical to the metric learning framework, as the model determines the semantic scores that can measure urbanization. During training, we limit the range of values of the scoring model from 0 to 1 by clamping smaller or larger values.

Among many existing ranking methods, three main themes include: (1) a point-wise method that directly prepares the model from training dataset [32], (2) a pair-wise method that learns from the comparison of given pairs [4, 21, 26], and (3) a list-wise method that optimizes the loss function on a list of predictions and actual values [1, 17]. The first two methods either require labeled data for training or easily suffer from clustering noise leading to flipped labels. For these reasons, we adopt the list-wise method with our unique structures for the scoring model, siScore.

List-wise metric learning: The only knowledge from POG is the orders of clusters, but not the orders of individual images. The third step, siScore, use the POG structure in learning scores of every satellite grid in the following way. We first extract every ordered paths \mathcal{P} from the POG. After choosing one of them P_j (i.e., $P_j = \{C_1, C_2, \dots, C_m\} \in \mathcal{P}$ where C_k is a k -th cluster on a chosen

path P_j , and m is a length of P_j), an equal batch size of n_s images are sampled from each cluster C_k along the selected path. siScore produces the scores to these selected images and the final scores are concatenated as s_i in Eq. 6, where $\mathbf{x}_{C_k}^i$ indicates the i -th image in cluster C_k .

$$s_i = \text{concat}([\mathbf{f}(\mathbf{x}_{C_1}^i), \mathbf{f}(\mathbf{x}_{C_2}^i), \dots, \mathbf{f}(\mathbf{x}_{C_m}^i)]) \quad (6)$$

The rank $\text{rank}(s_i)$ is calculated by sorting the score vector s_i . Since the selected path is already ordered, the actual rank r_j of clusters in P_j should be $\mathbf{r}_j = [0, 1, \dots, m]$. The scoring model f trains to match the following two rank vectors: $\text{rank}(s_i)$ and \mathbf{r}_j . The Spearman correlation evaluates how well given two variables are related as a monotonic function in terms of measuring rank correlation. Accordingly, we calculate and maximize the Spearman correlation of the estimated ranks against the actual rank to train the model as follows:

$$\max_f \left(1 - \frac{6 \|\text{rank}(s_i) - \mathbf{r}_j\|_2^2}{m(m^2 - 1)} \right), \quad (7)$$

where the rank function returns the sorted order of a list. If we assume that the length of the selected path is always the same (i.e., m), maximizing Eq. 7 is equivalent to minimizing the following loss function:

$$L_s = \frac{1}{n_s} \sum_{i=1}^{n_s} \|\text{rank}(s_i) - \mathbf{r}_j\|_2^2, \quad (8)$$

where n_s represents the batch size for training. siScore repeats this step for all paths P_j in \mathcal{P} .

Differentiable sorter: Besides determining the correct rank order of the selected image batches in path P_j , it is desirable to find scores that can “differentiate” one cluster from another. The Spearman correlation, in contrast, is unsuitable for such training due to the back-propagation nature of deep learning. Based on recent advances in computing ranking losses [10], we use an approximation method to mimic the sorting algorithm and make the algorithm differentiable to use the Spearman correlation as a loss directly.

Suppose a score vector $\hat{s} = \{e_1, e_2, \dots, e_n\}$ with n elements. Given all records of pair-wise comparisons among the elements, the algorithm sorts elements by counting wins for each element. We define the sigmoid function with hyper-parameter λ to make records of binary comparison (Eq. 9).

$$\sigma_\lambda(e_i, e_j) = \frac{1}{1 + e^{-\lambda(e_j - e_i)}}, \text{ where } \lambda > 0 \quad (9)$$

This sigmoid function takes two variables e_i, e_j and produces close to 1 if e_j is larger than e_i with sufficient margin, while producing close to 0 if e_i is larger than e_j . By calculating all records of comparisons with the sigmoid function, rank is generated by summing up the values of records for each element (Eq. 10). The hyper-parameter λ in sigmoid function affects the size of gradients during the backpropagation and the precision of estimated ranks by sorter h . The rank of i -th element in the list can be calculated as:

$$h(\hat{s}, i) = \sum_{k=1; k \neq i}^n \sigma(e_i, e_k). \quad (10)$$

To sort the images, we utilized differentiable sorter h to rank the values. Equation 8 is then re-written as follows:

$$L_s = \frac{1}{n_s} \sum_{i=1}^{n_s} \|h(s_i) - \mathbf{r}_j\|_2^2, \quad (11)$$

Variance regularization: Note that h is sensitive to the value of λ . Because of the characteristics of λ that controls the slope of comparison function σ_λ , it gives a trade-off between training efficiency and precision of the sorter. For a large λ , the precision of the sorter becomes higher, but only small differences in values between two elements result in substantial differences in ranks. This can cause every score from siScore to gather within the small range and thereby cause flipped results. Meanwhile, if we set λ small, large differences in values are required to make differences in rank, but the precision of the sorter would be low, and noise can be generated.

To leverage the high precision of sorter and avoid errors from the flipped results, we added a loss to regularize the score distribution of each cluster to satisfy small variance. With small variances in score distributions, the overlapping part between two adjacent score distributions, where the flipped results are brought, would be reduced. The average variance of score distributions of every cluster in the selected path P_j is minimized as a regularization loss:

$$L_{var} = \frac{1}{|P_j|} \sum_{C_i \in P_j} \text{Var}(f(\mathbf{X}_{C_i})), \quad (12)$$

where \mathbf{X}_{C_i} indicates the batch images in cluster C_i and Var denotes the function that calculates the variance of the given score list.

Finally, the loss function to train siScore is adjusted with the weight parameter α as in Eq. 13.

$$L_{score} = L_s + \alpha \times L_{var} \quad (13)$$

4 DATA AND SETTING

4.1 Satellite imagery

Remote sensing enables analysis over an extensive geographic coverage, near real-time accessibility, high spatial resolution, sub-district- or grid-level measurements, and change detection in land use or urban sprawls [8, 15]. ESRI, a GIS software company, provides satellite-based remote sensing data at various zoom levels. The zoom level (Z) refers to the spatial resolution, where the entire earth is fit into a single tile when $Z = 0$ and to 2^k tiles for $Z = k$. Since the number of pixels remains identical to 256-by-256 per image tile, a higher zoom level means richer resolution. We chose the $Z = 15$ images with 4.7m-resolution, which can distinguish individual buildings as well as other artifacts such as roads. Each tile contains three spectral bands (RGB), and the images are cloud-free for most of the area. We consider data from three countries: South Korea, Malawi, and Vietnam, where images are from between 2015 and 2017.

4.2 Public data sources

Census data: We utilize the total population data collected by Michael Bauer Research (MBR) GmbH, which investigates demographics data over 130 countries. The ArcGIS API allows for access to the latest data, and all data used in this paper were downloaded

in January 2020.

Nighttime satellite imagery: This data is known to contain a blurring effect of light emission that affects nearby areas with less luminosity. Hence the best resolution available for public use if at $Z = 9$, released by the NASA Earth Observing System Data and Information System. Due to the zoom level difference (15 vs. 9), we treat the data such that every 64-by-64 pixel within the $Z = 15$ grid matches with one single pixel of nightlight imagery. Hence, each $Z = 15$ grid holds a 4-by-4 pixel of nightlight intensity data.

4.3 Ground truth dataset

The grid-level ground truth labels were prepared in two forms: **Population** and **Gross Floor Area**. Facebook has contributed to making the most precise population map of the world, and they open the data that covers most of the Asian and African countries online [11]. The estimation is at the resolution of an arcsecond-by-arcsecond scale. Since this unit area is smaller than our grid size, we sum up the estimated population density of each unit located inside the grid for evaluation. Also, the Gross Floor Area, which is an amount of total construction in buildings, can refer to the degree of economic development in a given area. Such information can be computed from GIS data of building shapefiles. The local government can collect such information; we utilize the official data released by the Ministry of Land, Infrastructure, and Transport in South Korea.

4.4 Implementation setting

ResNet-18 [18] is used as a backbone CNN network for both clustering and scoring model. The pre-training step of siCluster used the same hyper-parameters in the original Mean Teacher algorithm in [16]. Adam optimizer with learning rate of $1e^{-4}$ was used, where the model trained for 200 epochs. Random rotation and grey scaling with probability 0.1 are utilized for data augmentation to make our model robust against the tone of color or textures. The batch size was set to 256 for clustering and to 20 for metric learning (i.e., $n_s = 20$ in Eq. 8). The number of clusters n_c to separate satellite imagery of urban and rural each was set to 10 in South Korea (i.e., total cluster count is $n_t = 2 \times n_c + 1 = 21$), λ in Eq. 9 to 30, and α in Eq. 13 to 4.

5 EXPERIMENTS

We conducted extensive experiments to examine how each part of the model operates via component analysis and test robustness across multiple countries.

5.1 Performance evaluation

The POG was generated as follows. **Human-guided** method involved annotations from five experts and five locals, where both the average and the maximum performance are reported. For data-guided POG, we utilized the population census data collected by MBR (**Census-guided**) and grid-level nightlight intensity data (**Nightlight-guided**). Table 1 reports the correlation values between the estimated economic development and two kinds of ground truth labels: Gross Floor Area and Population. Both Spearman and Pearson correlations were calculated on the log-scaled ground truth

Table 1: Comparison of prediction models for South Korea. Results are evaluated for two grid-level statistics: Gross Floor Area and Population Count.

	Method	Gross Floor Area		Population	
		Spearman	Pearson	Spearman	Pearson
A	Human-guided (Avg)	0.825	0.787	0.764	0.766
	Human-guided (Max)	0.851	0.800	0.795	0.778
	Census-guided	0.826	0.799	0.792	0.788
	Nightlight-guided	0.846	0.801	0.794	0.789
B	Nightlight-only	0.664	0.655	0.728	0.731
	Pairwise (Human)	0.651	0.610	0.300	0.302
C	K-means	0.434	0.587	0.451	0.557
	DeepCluster	0.618	0.559	0.532	0.551
D	Triplet (POG)	0.807	0.754	0.768	0.726
	Pairwise (POG)	0.825	0.759	0.767	0.739
	w/o Score model	0.737	0.675	0.678	0.673

A : Our model, B : Baselines, C : siCluster ablation, D : siScore ablation

values. All models produced scores of solid correlation (i.e., above 0.7) with ground truth labels, even when such information was not available during training. The best performance comes from the human-guided model, reaching 0.851 and 0.795 in Spearman correlations.

Seven baselines were implemented for comparison, which used ResNet-18 as the backbone network. (1) **Nightlight-only** uses the light intensity of nighttime satellite imagery for measuring economic development. We also experimented with human-annotated labels (2) **Pairwise (Human)** that indicated the relative rank of four thousand random image pairs. Three annotators who have domain knowledge of target countries were recruited. Participants were asked to choose which image in pair showed higher economic development, and their decisions were aggregated. The pairwise loss was then used for training. This model is one kind of naive methods that directly learn from the human-annotated orderings. These baselines were not as effective as our models.

The next two baselines are ablations for siCluster. We replace this module by the conventional (3) **K-means** clustering algorithm and by the original (4) **DeepCluster** algorithm. The remaining three are ablations for siScore. The labels for (5) **Triplet (POG)** and (6) **Pairwise (POG)** were generated by an identical POG instead of human annotation. When generating labels from the POG, cluster pairs are randomly selected, and images from each cluster form pairs. The order of chosen clusters is considered as a label for these pairs. Triplet makes labels, including anchor, positive, and negative samples. The model is trained to generate a similar score between an anchor and positive, while different and order-preserving scores should be calculated between an anchor and negative. We also proposed baseline (7) **without the score model**, which gives a scalar value that preserves the orders of POG to each cluster instead of a deep learning-based score model. Census-guided POG is used for these baselines.

Ablation studies demonstrate that every tested component of the model is critical to performance. The ablation study for the cluster method shows that siCluster is an optimal cluster method for satellite images. The Pearson value of siScore is far superior to other methods in the case of scoring model ablations. Triplet and Pairwise dip into confined POG rank. On the other hand, siScore

takes the overall POG rank as a whole and leverages the hierarchical relationship between clusters efficiently.

5.2 Component analyses

Clustering review: We examine the performance of siCluster qualitatively. A total of 95,936 grids were clustered into 21 clusters for South Korea (i.e., one uninhabited, ten rural, and ten urban). Fig. 2 show example images of cluster no. 1, 6, and 17. The clustered images show remarkable visual similarity in terms of urbanization and land cover types. Cluster no. 1 consistent of highly urbanized zones with large warehouse-like structures. Cluster no. 6 also represents urban structures but composed of smaller structures that are residential areas. Cluster no. 17 show agricultural zones.

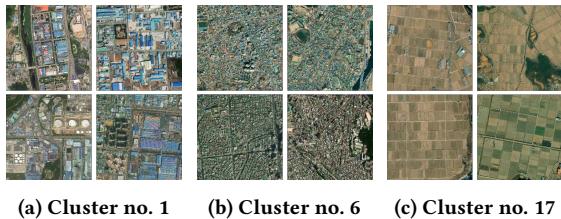


Figure 2: Example images from the same cluster. These images show similar visual characteristics of land covers.

POG design: To validate the intermediate POG designs generated by human-guided and data-guided siPog approaches, Table 2 reports that Spearman and Pearson correlation for the generated POGs and the actual orders of images according to the population data. We assigned the rank to every image following the POG, and images in the same cluster are considered to have the same rank. As a result, shown in Table 2, all three designs' correlations (Human, Census data-guided, and Nightlight data-guided POG) are near 0.7, and this high correlation infers that our POGs with three models well mimic the actual order of images.

Table 2: Measured consistency between generated POG from our proposed designs and ground-truth population dataset. Spearman and Pearson correlations are used for evaluation.

	Human-guided	Census-guided	Nightlight-guided
Spearman	0.68	0.68	0.71
Pearson	0.65	0.67	0.70

Hyper-parameters: Sensitivity of the model to parameters can be examined by enumerating the possible ranges of: (a) cluster count n_t , (b) λ in differentiable sorter (Eq. 9), and (c) α for the regularization weight (Eq. 13). Due to space limitation, we show results of the census data-guided POG in siPog in Figure 3. We make several observations.

First, the correlation is the highest at the cluster count of $n_t = 21$, and the model performance decreases for larger or fewer clusters. A possible explanation is that too many leads to flipped orders due to the overlap between distributions, whereas too few misses urbanization patterns. These results can be compared to the optimal

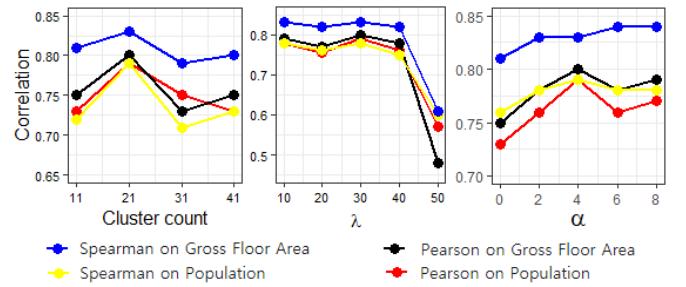


Figure 3: The effect of cluster counts and hyper-parameters (λ values in the differentiable sorter and α values in the regularization loss) on model performance.

number of clusters for Malawi and Vietnam, which were 7 and 11, respectively. This exploration shows that there exists an adequate cluster count for each target country.

Next, in terms of hyper-parameters, the prediction performance degrades for a large λ value. As explained earlier, such values can cause every score of the model to gather in a limited range and hence affects the efficiency in training due to the skewed gradients. This likely contributes to the poor performance above the λ value of 40. Also, the average accuracy reaches the highest at $\alpha = 4$ for both datasets, and no significant performance gain is observed below or beyond this point. In the case of the Spearman correlation on the Gross Floor Area, the highest performance is observed at $\alpha = 8$, although at a marginal. Altogether, we find that hyper-parameter tuning is important for the performance, and there exists some range of parameters (λ, α) that are acceptable to many models.

5.3 Robustness testing

Application to developing economies: We conducted additional experiments on two developing economies, Malawi and Vietnam, and gathered a total of 64,303 and 226,305 satellite images for them, respectively. All models were trained in the same manner as mentioned earlier, except for the cluster count n_t in siCluster. The optimal n_t was found to be 7 for Malawi and 11 for Vietnam based on grid search, respectively.

Figure 4 shows the performance comparison, based on the grid-level Facebook population data [11]. Our model repeatedly outperforms the conventional nightlight model for developing economies. In the case of Malawi (i.e., the poorer of the two), the nightlight baseline performs nearly one-third of the other best solution. As a result, the improvement is from ‘weak’ to ‘strong’ correlation. The light intensity from nighttime satellite imagery is known to be erroneous to be used for areas of extreme poverty [22]. The advantage of our model is attributed to the use of daytime imagery, which no longer suffers from the light saturation effect.

Next, Figure 5 shows the transferability of the model when it is trained on one country’s data and tested on another country’s data. As one might expect, training and testing model on the same country (i.e., the diagonal line) shows the highest performance in most cases. One exception is the census-guided POG generation for Vietnam. South Korea showed consistently above 0.6 Spearman correlation even when the model was trained on Malawi or Vietnam

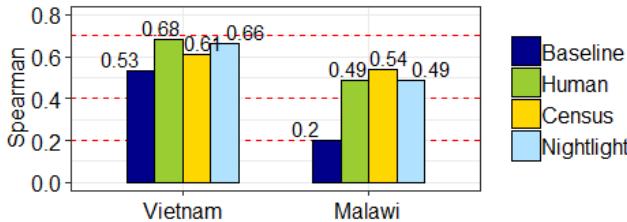


Figure 4: Results for Malawi and Vietnam. Three models (Human, Census, and Nightlight) are compared against the conventional nightlight model (Baseline). The red lines indicate the boundaries for ‘weak,’ ‘moderate,’ ‘strong’ and ‘very strong’ correlations from the bottom.

data, for all Human, Census, and Nightlight strategies. We speculate that a comparatively broad spectrum of urbanized patterns in South Korea contributes to this result. This may be linked to the fact that South Korea data required the highest number of clusters (21) compared to the other two (7 and 11), again likely because there were many more distinctive grids with the sufficient numbers of images in each cluster, showing vastly broad economic spectrum. In contrast, the poorest economy (i.e., Malawi) images may contain smaller variations across the grid images.

Human			Census			Nightlight			
	KOR	VNM	KOR	VNM	MWI	KOR	VNM	MWI	
Eval	0.649	0.771	0.795	0.720	0.682	0.792	0.632	0.778	0.794
VNM	0.592	0.738	0.609	0.657	0.616	0.518	0.578	0.665	0.551
MWI	0.573	0.524	0.482	0.540	0.224	0.328	0.493	0.483	0.428
Train									

Figure 5: Spearman correlation with ground-truth population for models that are trained in one country and evaluated in another country. Results from three designs of POG generation are visualized respectively.

Application on change detection and visual interpretation: Cities grow rapidly in many places. Monitoring changes in cities is crucial for sustaining economic development, yet this task is challenging to track. We here show that our model of estimating economic development can be used to quantify change detection. Figure 6 displays images of one district over time. Our algorithm can measure the economic development score from the grid-level images. The estimated scores increased from 0.13 in 2010 to 0.31 in 2018, quantifying the urbanizing process. Besides, the POG structure of the corresponding images as well as the visual interpretation of the model, generated by the grad-CAM algorithm [28], indicates which areas of the district contributed most to the scoring decision. The regions highlighted in the heat map were critical to the algorithmic decision. This example demonstrates how the algorithm can be used to monitor the urbanizing process over vast areas in both developed and developing economies.

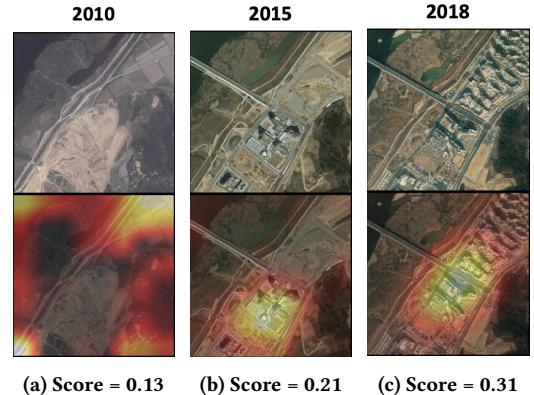


Figure 6: Satellite image and its corresponding score for urban development over multiple years for the same district. Estimated scores from our scoring model reflect the district’s evolution process (from score 0.13 to 0.31 on average). Below heat-map image extracted by grad-CAM [28] displays which image part contributes to the score increase.

Sensitivity of human annotator type: Does constructing a POG require advanced knowledge? We examined whether the expertise of annotators has any effect on the quality of POG. Our expert annotators were knowledgeable in economics and geography, whereas the local annotators had an in-depth knowledge of the target country. For different countries, we recruited different local annotators. All annotators were given the same instruction and data, and their POGs are compared in Figure 7. This small scale comparison shows there is no evidence of the quality difference in the generated POGs between experts and locals, demonstrated by the p-values of the two-sample t-tests. Furthermore, high graph consistency between the two groups, which is calculated by measuring the Spearman correlation of two partial orders, shows their POGs are similar (i.e., the value of graph consistency is near or over 0.7). These findings imply that our POG generation process is simple, and thereby anyone who wants to train the model can quickly implement by themselves.

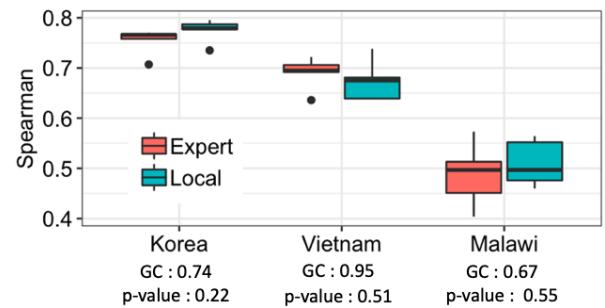


Figure 7: Spearman correlation of POGs generated by expert versus local annotators. GC refers to the graph consistency. The p-value of the t-test indicates there is no significant difference between experts and locals.

6 CONCLUSION

We presented a new method that measures economic development from high-resolution satellite imagery using deep learning. Our method first groups tens of thousands of grids into clusters of manageable sizes by visual patterns of artifacts. Then it generates a POG of the identified clusters guided by humans or low-resolution proxy data, which is lightweight compared to the overall scale of the problem. The final step assigns differentiable scores to the entire grid set based on the POG structure, which reported a high correlation to the ground truth data.

One implication of this research is the use of high representation power of deep learning in demographic analysis. The model applies to both developing economies where labeled data is limited and developed economies where grid-level census data are not gathered frequently. Generating a POG helps interpret which land cover patterns are linked to high urbanization scores. Our approach can be used in various scenarios, including change detection; repeatedly computing the economic development scores over newly collected satellite imagery can measure changes in fast-growing areas over time.

This work bares several limitations. First, since every step of our method affects the model performance, any errors in the initial clustering can propagate throughout the entire model. Combining the first two steps as an end-to-end model will be able to minimize standard errors. Second, the model is vulnerable to noise in satellite imagery. Due to its unsupervised fashion in the clustering process, noise in images, such as clouds [7], can be falsely identified as a unique pattern of human activities. Finally, while this research considered two kinds of annotators (e.g., experts and locals) as part of the human-in-the-loop design, it can be further designed to minimize the impact of variances in human performance. In the future, we would like to explore in-depth how computational intelligence can improve further by complementing it with human intelligence [20].

ACKNOWLEDGMENTS

This work was supported by the Institute for Basic Science (IBS-R029-C2). We thank Eunji Lee and Dohyeon Lee for their contributions on collecting data and visualization.

REFERENCES

- [1] Qingyao Ai, Keping Bi, Jiafeng Guo, and W Bruce Croft. 2018. Learning a deep listwise context model for ranking refinement. In *proc. of the ACM SIGIR*. 135–144.
- [2] Adrian Albert, Jasleen Kaur, and Marta C Gonzalez. 2017. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *proc. of the ACM SIGKDD*. 1357–1366.
- [3] Hasi Bagan and Yoshiki Yamagata. 2015. Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data. *GIScience & Remote Sensing* 52, 6 (2015), 765–780.
- [4] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23–581 (2010), 81.
- [5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *proc. of the ECCV*. 132–149.
- [6] Mayuri Chaturvedi, Tilottama Ghosh, and Laveesh Bhandari. 2011. Assessing Income Distribution at the District Level for India Using Nighttime Satellite Imagery. In *proc. of the APAN*.
- [7] Qing Cheng, Huanfeng Shen, Liangpei Zhang, Qiangqiang Yuan, and Chao Zeng. 2014. Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model. *ISPRS journal of photogrammetry and remote sensing* 92 (2014), 54–68.
- [8] Dave Donaldson and Adam Storeygard. 2016. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives* 30, 4 (2016), 171–198.
- [9] Erin Doxsey-Whitfield, Kytt MacManus, Susana B Adamo, Linda Pistolesi, John Squires, Olena Borkovska, and Sandra R Baptista. 2015. Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Papers in Applied Geography* 1, 3 (2015), 226–234.
- [10] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. 2019. SoDeep: a Sorting Deep net to learn ranking loss surrogates. In *proc. of the IEEE CVPR*. 10792–10801.
- [11] Facebook. 2020. *Data for Good* program. Available at <https://data.humdata.org/> organization/facebook. Date accessed 29 Jan 2020.
- [12] Jonathan A Foley, Ruth DeFries, Gregory P Asner, Carol Barford, Gordon Bonan, Stephen R Carpenter, F Stuart Chapin, Michael T Coe, Gretchen C Daily, Holly K Gibbs, et al. 2005. Global consequences of land use. *Science* 309, 5734 (2005), 570–574.
- [13] Tilottama Ghosh, Sharolyn Anderson, Christopher Elvidge, and Paul Sutton. 2013. Using nighttime satellite imagery as a proxy measure of human well-being. *Sustainability* 5, 12 (2013), 4988–5019.
- [14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.
- [15] Patrick Griffiths, Patrick Hostert, Oliver Gruebler, and Sebastian V. der Linden. 2010. Mapping megacity growth with multi-sensor data. *Remote Sensing of Environment* 114, 2 (2010), 426–439.
- [16] Sungwon Han, Donghyun Ahn, Hyunji Cha, Jeasurk Yang, Sungwon Park, and Meeyoung Cha. 2020. Lightweight and Robust Representation of Economic Scales from Satellite Imagery. In *proc. of the AAAI*.
- [17] Jing He, Xin Li, and Lejian Liao. 2017. Category-aware Next Point-of-Interest Recommendation via Listwise Bayesian Personalized Ranking. In *proc. of the IJCAI*. 1837–1843.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *proc. of the IEEE CVPR*.
- [19] Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E. Blumenstock. 2017. Can Human Development be Measured with Satellite Imagery?. In *proc. of the ICTD*.
- [20] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crișan, Camelia-M. Pintea, and Vasile Palade. 2019. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence* 49 (2019), 2401–2414.
- [21] Ziniu Hu, Yang Wang, Qu Peng, and Hang Li. 2019. Unbiased LambdaMART: An Unbiased Pairwise Learning-to-Rank Algorithm. In *proc. of the WWW*. 2830–2836.
- [22] 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.
- [23] Xu Ji, Joao F Henriques, and Andrea Vedaldi. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *proc. of the IEEE ICCV*. 9865–9874.
- [24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *proc. of the NIPS*. 1097–1105.
- [25] Husi Letu, Gegen Tana, Masanao Hara, and Fumihiro Nishio. 2011. Monitoring the electric power consumption by lighting from DMSP/OLS nighttime satellite imagery. In *proc. of the IEEE IGARSS*.
- [26] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. 2017. Computer vision uncovers predictors of physical urban change. *PNAS* 114, 29 (2017), 7571–7576.
- [27] Arthur O'Sullivan. 2019. *Urban Economics, Ninth Edition*. McGraw-Hill.
- [28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE ICCV*. 618–626.
- [29] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2019. Predicting economic development using geolocated wikipedia articles. In *proc. of the ACM SIGKDD*. 2698–2706.
- [30] Jessica E Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A Alegana, Tomas J Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M Iqbal, et al. 2017. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* 14, 127 (2017), 20160690.
- [31] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *proc. of the NIPS*.
- [32] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. 2017. Learning to rank question answer pairs with holographic dual LSTM architecture. In *proc. of the ACM SIGIR*. 695–704.
- [33] Henderson J. Vernon. 2003. Urbanization and Economic Development. *Annals of Economics and Finance* 4, 2 (2003), 275–341.

SUPPLEMENTARY MATERIAL

The authors note that the entire datasets used in this paper are non-proprietary online resources, except for human annotations. We share the model details and codes for the reproducibility of research outcomes at https://github.com/dscig/urban_score.

Clustering process details

The siCluster algorithm aims to capture the unique patterns of human activities in satellite imagery related to economic development. We made an improvement from the original DeepCluster by applying transfer learning. One thousand satellite images were labeled within three classes: urban, rural, and uninhabited. Four annotators contributed to generate the labels of the images; they were guided to decide the label according to the common traits on urban structure, such as compact construction, high population density, non-agricultural land use [27]. We aggregated the decisions of all annotators as average votes, which were then applied to construct a classifier. After training, the fine-tuned model without a final fully connected layer was utilized as the feature extractor during the main clustering step.

We initially split the entire dataset into three classes: urban, rural, and uninhabited. Then, the clustering algorithm runs separately for the urban and rural classes to find an equal number of n_c clusters. This makes the total cluster count to be $n_t = 2 \times n_c + 1$. Figure 8 shows example cluster images in South Korea, composed of ten urban, ten rural, and one uninhabited cluster. Each cluster shows consistent traits in image patterns. For example, cluster no. 2 consists of the urban apartment complex, cluster no. 10 and no. 14 consists of warehouses and paddy fields, respectively.

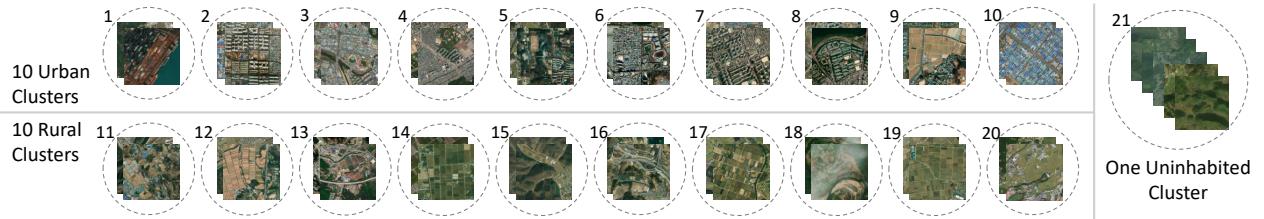


Figure 8: SiCluster results in South Korea. All satellite images are grouped into 21 clusters (i.e., $n_t = 21$).

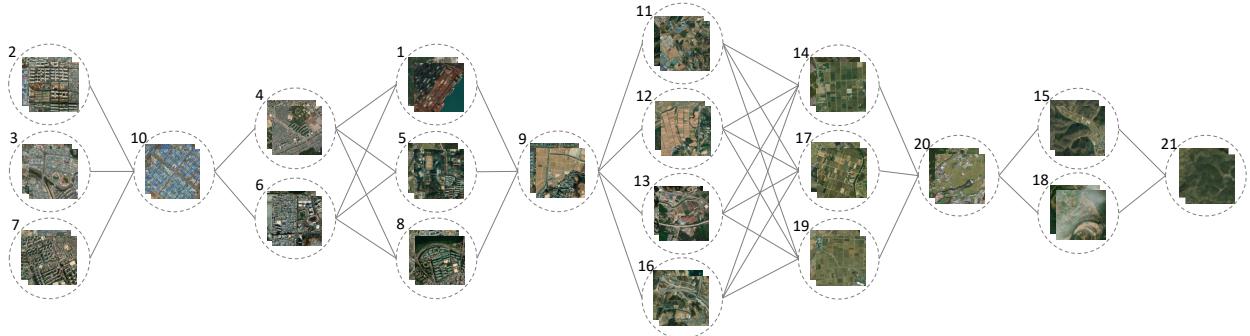


Figure 9: Example POG of South Korea from human-guided siPog. Clusters are sorted from urban to rural, from left to right.

Examples of partial order graphs

Once tens of thousands of satellite grids are grouped via siCluster, we are now able to construct a partial order graph (POG) to determine the relative order of clusters, by economic development. This step is the siPog algorithm. Figure 9 shows one instance of a POG for South Korea, generated by a human annotator. Twenty-one clusters are aligned in a graph format, from the most urban to the least. Clusters that have a similar level of development appear on the same level. This POG structure becomes the input to the final metric learning step. This step is also interpretable; later, one may use this structure to explain why certain images receive higher scores than others.

Hyper-parameter details

As suggested by Head et al. [19], parameter selection is critical when applying deep learning model on satellite images to different countries. We performed a hyper-parameter analysis in Section 5.2 and gave guidelines for appropriate ranges, which we provide more details. Our model has three hyper-parameters: (a) cluster count n_t , (b) λ in differentiable sorter (Eq. 9), and (c) α for the regularization weight (Eq. 13). In the case of cluster count n_t , optimal value mostly depends on the variations of pattern that each country has. Reflecting on the actual appearance of the target country, we adjusted the optimal n_t by grid search (South Korea: 21; Vietnam: 11; Malawi: 7). On the other hand, λ and α are the hyper-parameters comparatively less affected by changes in countries. The performance of our model is fairly robust if two parameters meet a certain condition ($\lambda < 50$ and $\alpha > 0$). Therefore, we adopted the hyper-parameters, λ and α , to Vietnam and Malawi the same as those tuned from South Korea ($\lambda = 30$ and $\alpha = 4$).