IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 6, 2021, accepted July 1, 2021, date of publication July 9, 2021, date of current version July 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3096034

Methods for Measuring Geodiversity in Large Overhead Imagery Datasets

AARON M. WESLEY [1,2], (Member, IEEE), AND TIMOTHY C. MATISZIW [1,3,4], (Senior Member, IEEE)

[1] Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA
[2] National Geospatial-Intelligence Agency, Saint Louis, MO 63118, USA
[3] Department of Civil and Environmental Engineering, University of Missouri, Columbia, MO 65211, USA
[4] Department of Geography, University of Missouri, Columbia, MO 65211, USA

Corresponding author: Aaron M. Wesley (amwvg5@umsystem.edu)

ABSTRACT Geographic variation in the appearance of objects on Earth is readily observable in remotely sensed imagery (RSI) and somewhat intuitive to understand for most people – many classes of objects (houses, vehicles, crop fields etc.) simply look different depending on their location. This variation has recently been shown to have important implications when training machine learning models on geotagged image datasets for specific object detection and classification tasks. For example, models trained on datasets with ethnocentric biases in image content have been shown to misclassify objects in under-sampled regions, particularly in least-developed countries. The need to evaluate the growing corpus of RSI datasets for representativeness, heterogeneity and *geodiversity* is therefore high; yet scalable methods for measuring these concepts are absent in the remote sensing domain. This paper introduces the first dataset analysis methods for detecting and assessing geodiversity problems in large RSI datasets, based on geospatial adaptations of the Fréchet Inception Distance and Inception Score in the deep learning framework. Geospatial Fréchet Distance is proposed as a dissimilarity measure for image features of an object class across geographic regions – useful for comparing differences in object class appearance in different locations and/or spatial scales. A complementary Geospatial Inception Score is proposed to quantify heterogeneity of geographic context present in dataset labels within particular regions/locations, taking into account the labels themselves as well as their immediate surroundings. Rigorous tests of these methods on simulated RSI datasets demonstrate their stability, sensitivity, and the broad range of dataset analyses to which they can be applied.

INDEX TERMS Big geodata, deep convolutional neural networks, explainable ai, geographic domain shift, GIS-remote sensing fusion, imagery interpretation, spatial data analysis.

I. INTRODUCTION

The appearance of common types of geographic phenomena (natural or of human origin) can exhibit a tremendous amount of variation not only from place to place but also over time. For example, a building or set of buildings categorized as a 'secondary school' located in an urban community in the USA can appear quite different than one located in a rural community in Liberia (see Fig. 1). The ubiquity of spatiotemporal variation in appearance within an object class (i.e., a distinct category of geographic objects) is in fact thought to exert some influence on human cognitive development [1], perception and processing of visual stimulus [2], conceptualization of distinct locations and regions [3], [4], organization of social systems within space [5], wayfinding capability [6], sense of place [7], and, of course, object recognition and ontological classification [8], [9]. The first law of geography – Waldo Tobler's assertion that 'everything is related to everything else, but near things are more related than distant things' [10] offers an implicit theoretical explanation for the often complex relationship between geographic regions and the appearance of objects found therein [11].

An overhead view of visual variation in landscape features over geographically expansive areas has evolved considerably over the last 50 years given advances in the field of remote sensing (RS). Increased spatial, spectral, radiometric

The associate editor coordinating the review of this manuscript and approving it for publication was Weimin Huang.

and temporal resolution of air and space-borne RS systems have made fine scale physical characteristics of increasingly smaller objects observable and available for analysis by the scientific community as well as the general public [9]. Almost concurrently, methods for image processing and computer vision have matured to exploit pixel, object and region level features of images [12], [13] in increasingly sophisticated classification algorithms, contributing to the widespread usage of deep learning (DL) classifiers based on convolutional neural networks (CNNs) [14]. Recently, extensive efforts at geospatial data labeling and indexing, channeled through open-source annotation platforms such as OpenStreetMap [15], fee-based labelling services such as Amazon Mechanical Turk [16], as well as other annotation campaigns in industry [17] and government [18], have resulted in a large number of public and proprietary remotely sensed imagery (RSI) datasets. Many of these datasets contain thousands or millions of examples of object classes as varied as bridges [19], coconut trees [20], sea lions [21] and damaged buildings [22], to name a few.

The confluence of data availability, pervasive dataset creation and readily deployable models has helped usher in the era of big geospatial data [23]–[27] and the mobilization of RSI to address complex problems in areas such as humanitarian assistance and disaster response (HA/DR) [28], public health surveillance [29], precision agriculture [30], mapping social inequities [31], wildlife tracking [32], among others. An understanding of the actual content of these datasets, particularly the level of meaningful and representative variation in *how* objects appear visually as well as *where* (i.e., geospatial context) they might be situated, is essential to effectively address geospatial problems. However, a main problem within the remote sensing domain is that elements of *geodiversity* of object classes within the current corpus of RSI datasets have been largely unexamined in efforts to vet these sources for operational use [33], [34]. This oversight is not entirely surprising given the source of the data: a persistent downlink from a constellation of multimodal satellites with global coverage and daily reimaging at submeter resolution or better. Thus, the assumption that datasets derived from such sources might contain extremely high levels of object class heterogeneity and comprehensive representation of objects' geographic contexts is quite reasonable. However, there is a need for tools and techniques for empirically testing such assumptions and for characterizing and analyzing geodiversity of appearance within the object classes in a dataset.

In this paper, the importance of addressing geodiversity in RSI datasets is established through a review of similar forms of data bias identified by other disciplines as well as recent calls for action from industry and government. A conceptual framework for measuring geodiversity of specific RSI object classes is then outlined. This framework relates the often-subjective concepts of equitability, fairness, heterogeneity, representativeness, and inclusivity, and allows for project-specific data diversity requirements. Finally, deep-feature RSI dataset analysis methods are introduced to fill the most basic gap – the need to quantify diversity of dataset object class representations within and among geographic regions, taking into account heterogeneity of the object's spatial context. These methods are applied to a simulated RSI dataset to test for stability, sensitivity, and scale dependence.[1] Whereas this paper's primary focus is on testing the spatially-explicit characteristics of the proposed methods, the results demonstrate both their efficacy and the rich and relatively untapped RSI dataset analyses to which they can be applied. To our knowledge, this research provides the first training dataset geodiversity analysis methods in the remote sensing literature (see Table 4) and represents the beginning of the academic conversation regarding representational bias in large overhead imagery sources. Hopefully, the insights provided here will spur the development of a wider variety and deeper complexity of diversity-related spatial dataset analysis methods and tools to enable better decision making in the era of big geospatial data.

II. BACKGROUND

A. CURRENT CHALLENGES

RSI datasets are just one subset of a much broader phenomena of large dataset creation for deep learning applications in the scientific and commercial domains. They have be used in a variety of contexts, ranging from text-based genomic analysis [35] and natural language processing [36] to image- and video-based vision tasks for media scraped from the web [37] or captured from distributed sensors like smartphones and traffic cameras [38], [39]. Given the sheer size and complexity of such multimodal, multidisciplinary data sources, it might be expected that models trained on these datasets contain finely-discriminative, highly-diverse, and inclusive features suitable for any number of uses worldwide. However, recent research on analysis of big datasets suggests that this is not the case. Serious problems of representational bias, often related directly to undersampling of distinct geographic regions and/or demographic groups, have been uncovered in gold-standard genomic databases used throughout the biomedical sciences [40]–[42], in speech and facial/gesture recognition services [43], [44], in popular cloud-based object detection and classification services [45], [46], and in benchmark computer vision training datasets [47]–[49]. Importantly, these studies show that models trained on datasets with pronounced Amerocentric and Eurocentric bias can perform poorly for artificial intelligence applications in least-developed countries (i.e., the Global South [33]) as well as the rural and low-income communities prioritized by international security and development goals [50]–[53].

There is little literature in which the presence and/or repercussions of geographic representation bias in RSI datasets

[1] Here, we follow the tradition in computer vision, remote sensing and image processing of first internally validating a spatially-explicit model with simulated data derived from known parameters to better understand the model's likely performance on real-world data, including any limits of model reliability due to spatial effects [121]–[127].

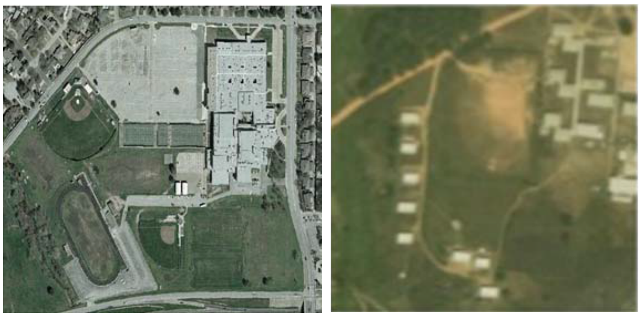

FIGURE 1. An example of visual variation in the 'secondary school' object class in Texas, USA (left) and Liberia (right) as seen in satellite imagery [102], [103].

is either mentioned [54]–[56] or investigated [33], [34]. Descriptions of new RSI datasets often reference the need for object class samples to be sufficiently heterogenous and representative to support generalizability of trained models and report basic object class spatial statistics such counts per region and label density/distribution as evidence of generalizability [22], [57]. However, attempts to quantify the extent and/or spatial characteristics of geographic variation in learnable image features of object classes in such datasets are absent in the remote sensing literature. As such, forceful concerns have been raised by industry about RSI dataset geodiversity problems [58] including the lack of tools to measure RSI feature diversity or predict the effect of biased data on trained model generalizability [59]. Interestingly, the U.S. Department of Defense, a main producer and consumer of RSI-based data and analyses, included a call to detect, understand and mitigate problems of dataset bias within its Artificial Intelligence Ethics Framework for the Intelligence Community [60]. Hence, image-content-based evaluation of RSI dataset bias remains a key gap in the academic literature as well as a priority for industry and government.

B. DIVERSITY AND GEODIVERSITY

Academic discussions related to diversity in datasets have used terms such as 'inclusive', 'equitable', 'heterogenous', 'fair' and 'representative', with varying agreement on the meaning of these concepts [61], [62]. Here, the overall concept of geodiversity in the context of RSI datasets is defined as 'the extent to which the observed heterogeneity of representations in sampled object class labels reflect the expected heterogeneity of the real-world object of interest.' This definition allows for an application-specific conceptual framework for evaluating dataset diversity that takes the target object(s) and project goals into account. From this starting point, it is argued that RSI training datasets achieve sufficient geodiversity for a given classification problem (and therefore equitability and representativeness) when the heterogeneity in the morphological and contextual representations of its object class(es) reflects that of the real-world object(s) in the target domain. This leaves room for datasets to be considered sufficiently geodiverse even if they consist of relatively homogenous, location-invariant object classes as long as a) there is also relatively little morphological and contextual variation in the real-world objects of interest, or b) if a detection/classification task requires very specific image features (e.g., rare object detection with few and relatively similar training examples not expected to deviate in appearance or context).

Theoretical and practical obstacles exist for assessing geodiversity of RSI datasets which relate to the uniqueness of RS phenomenology and the nature of Earth observation. First, subtle but important differences in the way humans and overhead sensors perceive and process the features of geographic objects can create equally subtle mismatches in the data models of such objects [63], [64], leading to potential differences in how humans and machines detect and interpret variation in object appearance [65]. Theoretically, this can cause humans to expect more (or less) visual variation for certain object classes in RSI data than warranted. Second, the extent of true object class variation may be unknown, and assumptions of geodiversity in image features of some object classes for certain regions may be untenable. Thus, there is a need for methods to empirically evaluate feature variation over the landscape. However, manual methods for this assessment in large RSI datasets can be impractical with thousands or millions of examples per class. As such, scalable quantitative metrics to support these types of tasks are required. Simply analyzing the coordinate locations of dataset samples (e.g., dispersion or clustering of points) would provide insufficient insight due to lack of information of the actual feature content of the images (and the untenability of any presumptions of local object diversity).

C. PRACTICAL AND SCIENTIFIC REQUIREMENTS FOR RSI DATASET GEODIVERSITY ASSESSMENT

Despite the above concerns, a growing need for RSI dataset geodiversity analysis tools exists in the scientific community as well as in industry and government. Depending on the goals and complexity of the operational use case, answers to increasingly complex questions concerning dataset geodiversity may be required, which until now have relied on manual or presumptive answers (e.g., those based on label location and point patterns). The following types of practical geodiversity-related questions may be encountered:

1. Binary – Is there detectable geographic variation in image features of object classes in a particular dataset?
2. Continuous – To what extent does image feature appearance vary over the study area?
3. Exploratory – What locations or regions have the highest (or lowest) diversity in the appearance of a particular object class or its immediate spatial context?
4. Comparative – Which locations or regions have the most similar (or dissimilar) image feature representations for an object class?
5. Explanatory – What factors contribute to the prevalence of particular image features detected in certain locations or regions?

6. Prescriptive – To what extent must more RSI data collection be done for a particular location or region to address trained model bias problems for an object class?

Answers to the above questions are not just required at the point of dataset usage; continuous triage of the dataset *creation phase* is needed, particularly for large-scale and complex imagery annotation campaigns (e.g., online volunteer annotation). Project managers have an unaddressed need to know what exactly is being collected, and if client or end user expectations for the dataset are being met.

Additionally, other basic scientific questions related to the geographic nature and statistical properties of RSI dataset geodiversity have largely gone unexplored. These include:

1. Train-test set parity/complementarity – What geographic differences in image feature geodiversity exist for the training and test split for a particular dataset?
2. Geographic bias – What specific image feature biases might a model trained on a certain dataset contain? For example, what characteristics of an object class has the model learned for locations in the training set, and would these generalize to locations present in the test set?
3. Scale dependence of geodiversity measures – At what smaller spatial scales, if any, does apparent heterogeneity in global image features for an object class become more homogenous?
4. Spatial dependency of object class geodiversity – To what extent does the geodiversity of object representations in a region influenced by that in neighboring regions?
5. Spatial nonstationarity of image feature variance – How smooth is the spatial transition of object class appearance across a study area?

Whereas measures to assess the above aspects of *geo*diversity in large training datasets are absent in the remote sensing domain, a variety of measures of general, nonspatial data diversity have been proposed in the broader computer science and machine learning literature. The following sections review research which model concepts such as heterogeneity, representativeness, and fairness for categorical and continuous data, highlighting methods with high potential for adaptation to RSI data.

D. CATEGORY-BASED MEASURES OF DATA DIVERSITY

A measure of *fairness* in ML classification, in terms of mitigating ML algorithm reliance on group- and individual-level protected attributes (e.g., race, ethnicity) during the classification process, was developed by [66] to obfuscate attributes which may identify a record as belonging to a protected subgroup. Reference [48] posits *fairness* and *diversity* as the degree of categorical bias present in a dataset, and applies a distribution rebalancing technique to ImageNet image annotations to achieve statistical parity among representations within dataset categories. Reference [66] explores the theoretical relationship between fairness and diversity and argues that metrics for data category separation, independence, and sufficiency are needed to fully model dataset fairness. There have been attempts to measure the complexity or variety of whole-scene images in terms of the number and type of discriminable phenomena detected in the image. For the most part, these entail algorithms which model the uncertainty inherent in assigning an image to a single object class based on the apparent number of object classes present in the image. For example, [67] propose a method for image complexity assessment for generic, nongeotagged images based on Shannon entropy of handcrafted image features. More recently, interest in assessing the quality of synthetic images created by deep (i.e., CNN-based) generative models has increased, especially with respect to generative adversarial network (GAN) output. Along with this trend has come key implementations of measures of whole-image content variety for generative models trained on multicategory datasets.

One of these measures, the Inception Score (*IS*) [68], posits that the quality of generated images increases as a) the certainty of object class membership within individual synthetic images increases (i.e., each generated image for a certain object class contains a distinct and easily distinguishable example of the object class), and b) as the categorical variety of generated object classes in a batch of synthetic samples increases (i.e., the number of object classes detected by a classifier approaches the total number of object classes learned by that classifier). Both criteria can be extracted from the softmax function output of a CNN pre-trained on $j \in J$ classes arrived at by passing a batch of generated images $x_i \in X$ through the network. This output can be conceptualized as an $|X| \times |J|$ matrix, each row i containing numeric class membership probabilities (conditional label distribution) that sum to 1.0 for each respective image. Object class certainty criteria can then be modeled as the Shannon entropy of the conditional label distribution $p(y = j|x_i)$, (where p denotes the probability of correctly predicting the label $y \in Y$ of an image given training sample x_i), and the categorical variety criteria can be modeled as the Shannon entropy of the integral of the conditional label distribution $\int p(y = j|x_i) = p(y)$. Given these assumptions, a batch of images can be passed through a pretrained CNN [68] to compute the conditional and marginal distributions. The *IS* is then computed as a measure of the magnitude of the difference between the two distributions as in (1).

$$IS = \exp\left((1/X) \sum\nolimits_{i \in X} D_{KL}\left(p(y = j|x_i)||p(y)\right)\right) \tag{1}$$

Kullback-Leibler divergence D_{KL} is used to assess the difference between the distributions and the exponential of the average D_{KL} over all images $x \in X$ can be computed to ensure *IS* falls within the range $[1, |J|]$ [69]. An 'ideal' result (i.e., $IS = |J|$) is reached when: a) minimum entropy/uncertainty exists in the conditional distribution, and b) maximum entropy/uncertainty exists in the marginal distribution. The effect on *IS* of less than perfect divergence between the conditional and marginal distributions is illustrated in Fig. 2. Importantly, the detection of multiple objects per image and/or imbalanced class representation between images results in lower *IS* scores, to be interpreted as ‘low quality’ synthetic images.

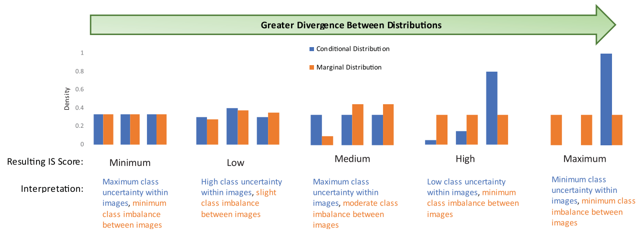

FIGURE 2. An illustration of resulting *IS* values as divergence between the conditional (blue) and marginal (orange) distributions increases.

E. FEATURE DISTANCE-BASED MEASURES

The concept of similarity of data samples has a long history of usage in data mining and pattern recognition [70], often measured in terms of relative distance of data points within the space of their attributes. The compactness (clustering) or evenness (dispersion) of data points makes classification possible [71] and directly relates to conceptualizations of entropy, information content, and diversity. Less distance among clusters (or individual points) equates to more similarity, less entropy, and less diversity, whereas more distance equates to less similarity, more entropy, and more diversity [72], [73]. Algorithms to summarize dissimilarity among clusters of high-dimensional points using ‘diversity maximization’ are described by [74], [75] for datasets of arbitrary type in the context of data mining in multimedia internet databases. Pairwise feature distance dissimilarity metrics are detailed by [76], [77] in the context of assessing nongeotagged image dataset task suitability and crowd-sourced worker opinion diversity, respectively, outside of the DL framework. Data collection-phase assessment of feature diversity which takes into account data noise by applying average distance to probability density transformations of feature space are proposed by [78] and applied to ImageNet. Besides directly analyzing datasets to measure feature diversity, image features contained within trained models can be the focus of analysis. For example, [79] addresses the problem of redundant image features in CNNs by applying a cosine similarity-based feature diversity regularization term to the objective function. Some approaches have sought to characterize the representativeness of sampled RS-derived measurements relative to a ‘global’ representation (e.g., a reference dataset) using feature distance metrics at the pixel level [51], [80]–[82]. However, such methods have not been applied to object or whole image analysis of feature similarity (or dissimilarity) in RS datasets in the DL framework.

In addition to *IS*, other metrics for assessing GAN quality have recently been developed to directly compare high-dimensional image features present in an ‘actual’ training dataset with those in synthetic GAN output. For instance, the Fréchet Inception Distance (*FID*) involves comparing learned image features between a training set and a batch of GAN output, extracting these from feature vectors contained in the final pooling layer of the Inceptionv3 CNN architecture [83]. The model can contain pre-trained features from a comparable dataset (e.g., ImageNet) or the features can be constructed from scratch directly from the input data. Specifically, *FID* is an adaptation of the Wasserstein-2 distance (d^2) [84] between the multivariate Gaussian $G(\mu_a, C_a)$ created from the mean μ_a and covariance C_a of feature vectors of the ‘actual’ image data a and the multivariate Gaussian $G(\mu_s, C_s)$ created from the mean μ_s and covariance C_s of the synthetic GAN output s. *FID* is the minimum linear divergence of the first and second order moments (mean and covariance) of the real and synthetic images arrived at by adding the Euclidean norm of means μ_a and $\mu_s(||\mu_a - \mu_s||_2^2)$ with a trace operation (Tr) of matrices C_a and C_s as in (2):

$$\begin{aligned} FID &= d^2 G(\mu_a, C_a) G(\mu_s, C_s) \\ &= \|\mu_a - \mu_s\|_2^2 + Tr(C_a + C_s - 2(C_a C_s)^{1/2}) \end{aligned} \quad (2)$$

If the image sets are identical, there will be no distance between the two, resulting in a *FID* = 0.0. Conversely, if the image sets contain completely dissimilar image features, the *FID* score approaches the total length of the feature space (e.g., an upper limit of ~768.0 for the 8-bit, 3-channel images commonly used in deep learning applications). However, the interpretation of a ‘good’ *FID* score (in the context of measuring synthetic image quality) is somewhat subjective and with no universally accepted threshold score. GAN results for benchmark datasets such as CIFAR-10, LSUN, and Celeb-A [85], [86] demonstrate that synthetic samples can become visually indistinguishable from real samples when *FID* scores are less than ~50.0. Feature dissimilarity and/or image distortions become more pronounced as *FID* scores increase from ~50.0 to ~300.0, and the fidelity and structure of objects in synthetic images can become difficult to visually discern as *FID* scores increase past ~300.0.

Both *FID* and *IS* are promising candidates for scalable spatial exploration of the multiple facets of object class geodiversity in present and future RSI datasets. The next section introduces geospatial extensions of *IS* and *FID* tailored to the remote sensing domain.

III. METHODS

A. GEOSPATIAL INCEPTION SCORE

The Geospatial Inception Score ($GeoIS_r$) is proposed as a triplet of indices (φ, μ, β) for evaluating the number and variety of co-occurring objects and background contexts surrounding object class labels in a geographic region $r \in R$ within a RSI dataset (see Fig. 3 for an illustration).

In situations where regional changes in density of co-occurring objects is the focus of the geodiversity analysis, $GeoIS_r^{\varphi}$ can be calculated for region $r \in R$ given the

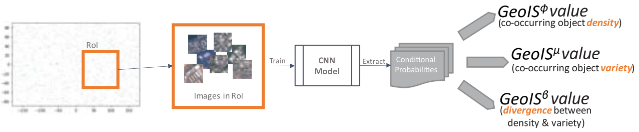
RoI
Images in RoI
Train
CNN Model
Extract
Conditional Probabilities
GeoIS^Φ value
(co-occurring object density)
GeoIS^μ value
(co-occurring object variety)
GeoIS^β value
(divergence between density & variety)

FIGURE 3. Extraction of image chips of an object class within an arbitrary region of interest and computation of the three *GeoIS* indices representing different facets of contextual geodiversity: co-occurring object density, co-occurring object variety and the magnitude of divergence between density and variety.

conditional distribution $f_i = p(y = j|x_i)$ as in (1). $GeoIS_r^{\varphi}$ can be structured as the Hill number of the average Shannon entropy of f_i over samples $x_i \in X$ as in (3).

$$GeoIS_r^{\varphi} = \exp\left((1/X) - \sum\nolimits_{i \in X} p(\varphi_i) \log p(\varphi_i)\right) \tag{3}$$

$GeoIS_r^{\varphi}$ can be interpreted as the average number of co-occurring objects/contexts (i.e., object density) near the target label in region r. In situations where regional changes in the number of distinct classes of co-occurring objects around the target label is the focus on the geodiversity analysis, $GeoIS_r^{\mu}$ can be calculated for region $r \in R$ given the marginal distribution $m_i = p(y)$ as in (1). $GeoIS_r^{\mu}$ can be structured as the Hill number of the average Shannon entropy of the marginal distribution as shown in (4).

$$GeoIS_r^{\mu} = \exp\left((1/X) - \sum\nolimits_{i \in X} p(m_i) \log p(m_i)\right) \tag{4}$$

$GeoIS_r^{\mu}$ can be interpreted as the average number of distinct classes of co-occurring objects/contexts (i.e., 'object variety') near the target label in region r. Given the maximum number of distinct classes is equal to the $|J|$ classes on which the CNN model was pre-trained, the output index of $GeoIS_r^{\mu}$ can be further scaled by the number of classes (e.g., $GeoIS_r^{\mu}/|J|$) such that the final score is the proportion of possible co-occurring object variety allowed by the CNN.

In situations where the total *bias* of co-occurring object density versus co-occurring object variety is the focus of the dataset geodiversity analysis, $GeoIS_r^{\beta}$ is proposed as in the original *IS* algorithm (1) for an arbitrary region r as shown in (5).

$$GeoIS_r^{\beta} = \exp\left((1/X) \sum\nolimits_{i \in X} D_{KL}(\varphi_i || m_i)\right) \tag{5}$$

Optionally, (5) can be scaled as $(GeoIS_r^{\beta}/|J|)$ so that the final score is the proportion of possible divergence between the conditional φ and marginal μ distributions.

Whereas the *IS* algorithm integrates the concepts of within image object class certainty (conditional distribution) and between image object class variety (marginal distribution) into a single index (via K-L divergence), $GeoIS_r$ allows for the separate evaluation of both components depending on the type of geodiversity analysis required. This is necessary because the original *IS* tends to penalize similar distributional patterns in the conditional and marginal with low scores, whereas in the RSI context these patterns could sometimes signify *greater* geodiversity and warrant higher scores. Fig. 4 illustrates two such situations in which a low *IS* index score can mask high levels of *contextual* geodiversity in RSI datasets based on object co-occurrence. The components of $GeoIS_r$ ($GeoIS_r^{\varphi}$, $GeoIS_r^{\mu}$, and $GeoIS_r^{\beta}$) can be applied individually or together to any georeferenced image dataset which links a class label for a geographic phenomenon (e.g., object, land cover, activity, etc.) to a specific location on the Earth. This includes popular benchmark RSI datasets that combine overhead imagery swaths with text annotations of georeference and metadata (e.g., [19], [87]) as well as datasets created on-the-fly from queries to geodata APIs provided by online services such as OpenStreetMap or Google Maps.

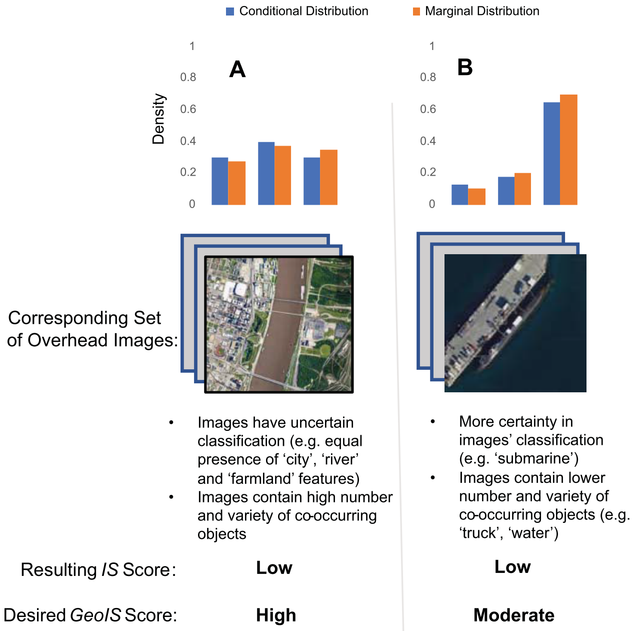

FIGURE 4. Expected result (A) of applying the original *IS* computation toward the RSI geodiversity analysis problem versus the desired result (B) for *GeoIS*.

Spatial analysis of how $GeoIS_r$ changes across dataset subregions can reveal patterns of relative heterogeneity or homogeneity in the locations chosen for an object classes' samples. Therefore, it can be useful in reasoning about a) the diversity of geographic contexts in which a particular object class is situated, b) changes to this diversity throughout the spatial extent of an RSI dataset, c) the density of associated (co-occurring) objects near the target label, and d) changes to this density throughout the spatial extent of an RSI dataset.

The following describes the general workflow to apply $GeoIS_r$ to a locally-stored RSI dataset which includes text files (e.g., json format) that reference X-Y coordinate locations within georeferenced image files (e.g., geotiff format). First, the component(s) of the $GeoIS_r$ computation are determined based on the requirements of analysis. Then, a CNN backbone pretrained on j object class labels is selected to detect the co-occurring objects/contexts of interest near the target label locations.[2] Image chips within a prescribed distance of label locations are then extracted to be fed into the CNN for co-occurring object detection. The desired unit(s) of spatial aggregation for binning chipped images is then determined and created, which can be single or multiple geographic regions within the spatial extent of dataset labels, or a tessellation of the entire extent (e.g., gridded overlay of the minimum bounding geometry of the dataset). For each region r, the desired $GeoIS_r$ components are calculated for the set of image chips x located within that region (e.g., those whose centroids fall within polygon r) by way of Eq. 3-5. The resulting scores are appended as attributes to each polygon r for follow-on spatial analysis.

B. GEOSPATIAL FRÉCHET INCEPTION DISTANCE

The Geospatial Fréchet Inception Distance ($GeoFID^k_{rq}$) is proposed to compare image features (e.g., texture, shape, association, etc.) of an object class $k \in K$ learned in two geographic subsets of RSI datasets (see Fig. 5 for an illustrative example). As the *FID* algorithm directly compares image features learned in two 'bins' of images (i.e., actual and synthetic), its adaptation to comparing spatial bins of RSI label samples is straightforward and, like $GeoIS_r$, can be applied to any georeferenced image dataset which links a class label for a geographic phenomenon to a specific location on the Earth.

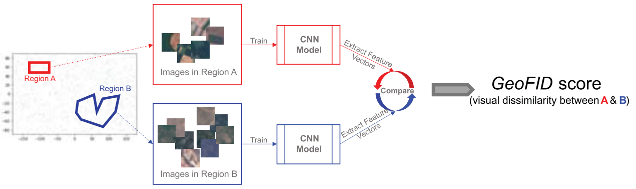

FIGURE 5. Illustration of *GeoFID* workflow for computing the dissimilarity of appearance of samples of an object class between two geographic regions.

Specifically, the $GeoFID^k_{rq}$ takes as input a batch of image samples $x_i \in X^k_r$ of class k from region r and a batch of image samples $v_i \in V^k_q$ of class k from region q and outputs the dissimilarity between the two as shown in (6).

$$\begin{aligned} GeoFID^k_{rq} &= d^2 G(\mu^k_r, C^k_r) G(\mu^k_q, C^k_q) \\ &= \left\| \mu^k_r - \mu^k_q \right\|^2_2 + Tr(C^k_r + C^k_q - 2(C^k_r C^k_q)^{1/2}) \end{aligned} \quad (6)$$

This is accomplished through comparison of the multivariate Gaussians $G(\mu, C)$ of samples X from region r and samples V from region q as in (2). Specifically, instead of comparing image features learned in bin a of actual images and bin s of synthetic images as in (2), spatial regions $r \in R$ and $q \in R$ are compared. As with the *FID*, $GeoFID^k_{rq}$ values range from 0.0 (signifying identical image features for the object class between the two geographic regions) to a limit value equal to the length of the feature space (signifying maximum dissimilarity of image features for the object class between the two geographic regions).

For example, the image features of samples within a small local subregion can be compared to the 'global' class distribution, or subregions of similar scale can be compared to each other. Spatial analysis of pairwise region-level comparisons of an object class' appearance can reveal locations in which the object class appears significantly different from the global distribution, 'feature boundaries' in which object class appearance changes abruptly in space, and other insights. Thus, the $GeoFID^k_{rq}$ could be used to provide insight as to a) the magnitude of visual variation in the target label throughout the spatial extent of the dataset to detect homogeneity / representational bias, b) differences in target label appearance between two regions to empirically test assumptions of that difference, and c) visual outlier locations with unexpectedly high dissimilarity in the target label vis-a-viz the global distribution.

The following describes the general workflow to apply $GeoFID^k_{rq}$ to a locally stored RSI dataset which includes text files (e.g., json format) that reference X-Y coordinate locations within georeferenced image files (e.g., geotiff format). First, an appropriately pre-trained CNN backbone is selected to extract image features. Depending on the format of the dataset labels (i.e., point or polygon labels), as well as the goal of the analysis, additional image context within a prescribed distance of the target labels can be included. Alternatively, image chips of only the label extents themselves can be extracted. The desired unit(s) of spatial aggregation for binning chipped images and comparing regions is then determined and created; this could simply be two regions, or a tessellation of the entire extent (e.g., gridded overlay of the minimum bounding geometry of the dataset) to facilitate comprehensive pairwise, focal, or zonal comparisons. For each intended pairwise region comparison rq, $GeoFID^k_{rq}$ is calculated between the set of image chips $x_i \in X^k_r$ of class k from region r (those whose centroids are within polygon r) and the set of image chips $v_i \in V^k_q$ of class k from region q (those whose centroids fall within polygon q) as in Eq. 6. The resulting scores are relations defining the visual dissimilarity of class k between r and q and can be appended as attributes to each polygon r or q for follow-on spatial analysis, or they can be stored in digraph matrix format for further computation.

IV. APPLICATION

A. SYNTHETIC DATASET

$GeoFID^k_{rq}$ and $GeoIS_r$ are applied iteratively over several spatial scales on a synthetic dataset with image features that vary geographically according to known parameters. This is done to evaluate the methods for stability, scale dependence (i.e., effects of Modifiable Areal Unit Problem), sensitivity,

[2] Care should be taken to choose a CNN model which reflects the goal of the geodiversity/bias analysis. For narrow-scope analysis attempting to account for a small number of specific co-occurring objects of interest, a CNN trained on only those object classes would be appropriate. For more comprehensive analysis, a CNN trained on a large number of object and land cover classes may be warranted.

and any model-specific sample size requirements. Two sets of 1000 samples of 3 synthetic object classes representing simple shapes (ellipse, star, polygon) were created using the *pycairo* Python package so that image features for each object class, including size, shape, color, rotation, translation (x-y displacement), ellipse ratio (for ellipse features), number of sides (for polygon features), and number of points (for star features), can be controlled parametrically. Both sets of shape classes are assigned a random geolocation over the latitude/longitude domain with bounds of (+90.0, -90.0) for latitude and (+180.0, -180.0) for longitude to simulate the GCS 1984 coordinate system. One set of shape objects is created with image features that are randomized (i.e., location independent). The other set of shape objects is created with image features that vary isometrically in magnitude according to their assigned (x, y) coordinate location (i.e., location-dependent). All datasets are available in a *figshare* repository (https://doi.org/10.6084/m9.figshare.13318271.v1). An example of the polygon object class' feature variation over the geospatial domain for the random and controlled sets is illustrated in Figs. 6 and 7, respectively. To facilitate application of $GeoFID^k_{rq}$ and $GeoIS_r$ at different geographic scales, the study area is partitioned into 4, 9, and 16 equal area subregion grids as geojson polygon objects using the shapely Python package (see Figs. 8A-D). Synthetic images are then indexed to these grids based on their location.

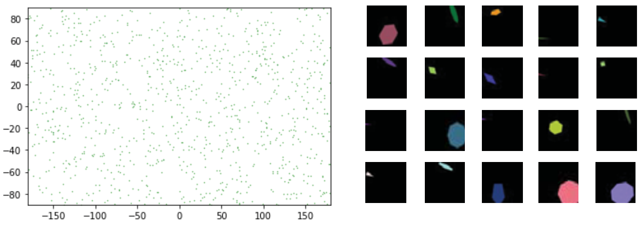

FIGURE 6. Example of random image features for the polygon class. The location of each sample (green dots) is assigned an image with randomly generated polygon features (for size, number of sides, displacement, rotation, skewness, and color attributes).

All experiments are performed in Python 3.5 using Google Cloud Services with a Linux virtual machine allocated 1x Tesla P100 GPU, 25.51GB RAM and 147.15GB HDD storage.

B. EXPERIMENTAL SETUP FOR GEOFID

The task of comparing an arbitrary subregion of images with those in the entire study region is simulated by calculating the $GeoFID^k_{rq}$ for the entire region R (Fig. 8A) relative to itself $GeoFID^k_{RR}$ (to ensure unity) as well as to each subregion $GeoFID^k_{Rr} \forall r \in R$ in order to analyze variations in those values within and between scales. In the following experiments, the Inception v3 CNN network pretrained on the 1000-class ImageNet dataset is used and the learned features on the synthetic dataset are fine-tuned. Finally, the learned features are extracted from the final pooling layer of the CNN to be used as inputs to the $GeoFID^k_{rq}$ index.

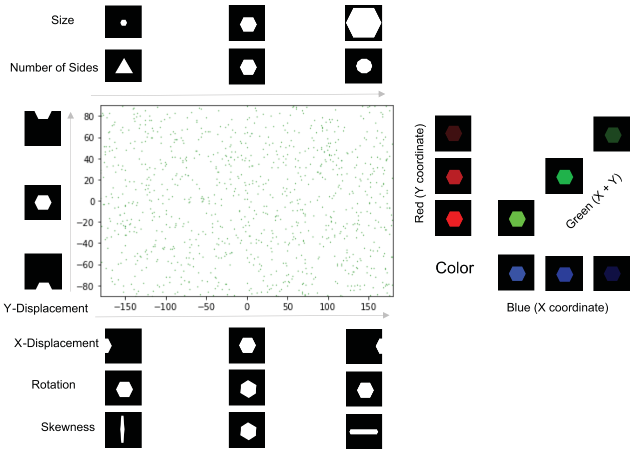

FIGURE 7. Example of geographically controlled image features for the polygon class. Each location (green dot) is assigned an image with polygonal features having isometrically constrained magnitudes.

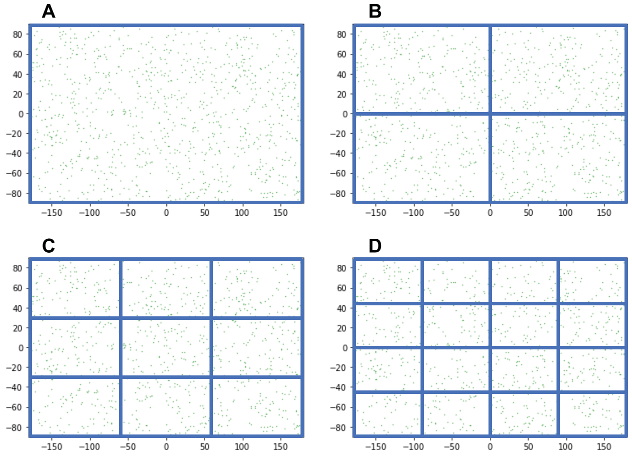

FIGURE 8. The 'global' region (A) of the WGS 1984 coordinate plane, along with subregions evaluated in the synthetic dataset experiments: (B) 4 subregions, (C) 9 subregions, and (D) 16 subregions.

Criteria for measuring stability, scale dependence, and sensitivity of $GeoFID^k_{rq}$ in the experiment series are now established. For the set of shape objects constructed with random image features, it is known that the region as a whole and each subregion have features drawn from the same (random) distribution and should be visually indistinguishable. Therefore, any variation in $GeoFID^k_{Rr}$ values in the random feature set between the region as a whole and subregions of the same scale is an indicator of $GeoFID^k_{Rr}$ stability within that scale. Stability is measured and reported by the standard deviation of $GeoFID^k_{Rr}$ values within subregion grids (e.g., 9 values for the 3 × 3 grid and 16 values for the 4 × 4 grid). Likewise, any variation in $GeoFID^k_{Rr}$ values in the random feature set between the global distribution and subregions of *different* scale is an indicator of the MAUP's effect on $GeoFID^k_{rq}$ across scales. Scale dependence is measured and reporting by noting changes in average $GeoFID^k_{Rr}$ across the 3 subregion grids of the random-feature set. Finally, for the set of shape objects constructed with geographically controlled features, it is known that visual variation should exist between the global distribution and any subregion, with the magnitude of that variation directly related to the scale of the subregion (because image features are more homogenous within smaller subregions). Therefore, it is expected that $GeoFID^{k}_{Rr}$ values in the controlled feature set increase as smaller subregions are compared to the region as a whole, with characteristics of this trend being an indicator of $GeoFID^{k}_{Rr}$ sensitivity to small changes of feature variation. Sensitivity is measured and reported by the trend in average $GeoFID^{k}_{Rr}$ across the 3 subregion grids of the controlled feature set.

Some expectations and criteria for success in a $GeoFID^{k}_{rq}$ experiment series could include:

1. Unity – the $GeoFID^{k}_{RR}$ value between any region and itself (in this case, the region as a whole and itself) is at or near 0.0 for all object classes, signifying identical image statistics and near-perfect similarity (subject to small perturbations due to stochastic nature of CNN model training).
2. Stability – The standard deviations of $GeoFID^{k}_{Rr}$ values within subregion grids in the random feature set are low (e.g., $\leq$10.0% relative standard deviation) AND significantly different than their control-feature counterparts at the same scale (as measured by two-tailed F-test).
3. Scale dependence – Following the observation that the strength of multivariate image statistics tends to degrade with smaller spatial scales due to the MAUP [88], [89], it is expected that $GeoFID^{k}_{Rr}$ will *increase* as spatial scale decreases (signaling increasing visual differences between the region as a whole and smaller subregions). However, for $GeoFID^{k}_{rq}$ to be considered a reliable discriminator, the effects of the MAUP on $GeoFID^{k}_{Rr}$ at smaller scales for the random synthetic shape set should not exceed the qualitative threshold considered as 'visually similar' in the literature (i.e., the two random distributions being compared are not misclassified as distinct distributions solely because of the MAUP).
4. Sensitivity – Whereas higher $GeoFID^{k}_{Rr}$ values (i.e., dissimilarity) are expected between the region as a whole (R) and smaller subregions r therein, a limit exists in the detected dissimilarity (though distant or of different scale, regional samples of the same object class have some underlying similarity that justifies membership in the same class).

C. EXPERIMENTAL SETUP FOR GeoIS

In the next set of experiments, the bias property of $GeoIS_r$ ($GeoIS^{\beta}_{r}$) is tested on the same synthetic shapes dataset. The Inception v3 network pretrained on the 1000-class ImageNet dataset [90] is again used to calculate $GeoIS^{\beta}_{r}$ for each subregion grid of the random and control feature sets. Given that the dataset was constructed with one shape object per image, high entropy in both the conditional and marginal distributions is expected and should result in low $GeoIS^{\beta}_{r}$ scores for any region across all 3 shape classes. However, analyzing the extent of score variation within and between scales permits assessment of criteria such as stability, scale dependence and sensitivity for $GeoIS_r$, much like $GeoFID^{k}_{rq}$. Because only one set of image samples is used in $GeoIS_r$, stability of the measure is evaluated on a per-region basis with 10-fold cross-validation (on the random feature set). Scale dependence is measured by noting changes in average $GeoIS_r$ across the 3 subregion grids of the random feature set. Sensitivity is measured by noting the trend in average $GeoIS_r$ across the 3 subregion grids of the controlled feature set.

Some expectations and criteria for success in the $GeoIS_r$ experiments series include:

1. Stability – For the random feature set, 10-fold cross-validation $GeoIS_r$ values for subregions at the same scale are expected to be very similar (low standard deviations between scores as measured by $\leq$10.0% relative deviation from the mean $GeoIS_r$ value), because each spatial bin of images at the same scale has very similar distributions.
2. Scale Dependence – Following the same MAUP assumption as $GeoFID^{k}_{rq}$ (that the strength of multivariate image statistics tends to degrade with smaller spatial scales), it is expected that for the random feature set, $GeoIS_r$ will *decrease* as spatial scale decreases.
3. Sensitivity – For the controlled feature set, the global sample should contain the highest object variety (marginal distribution), and this variety should diminish at smaller spatial scales due to a combination of MAUP effects and an increasing homogeneity of the objects at smaller scales.

V. RESULTS

A. GeoFID

Table 1 shows unity tests for the synthetic shape datasets where the values are the $GeoFID^{k}_{RR}$ reflecting a comparison of the region as a whole to itself for each of the three shape classes. Scores were consistently below -1.0E^{-4} for all classes in both the randomized and geographically controlled feature image sets, sufficiently close to the expected score (0.0) and demonstrating reliable unity of the model.

Table 2 shows percent RSD of $GeoFID^{k}_{Rr}$ scores within subregion grids across the synthetic shape classes for the random and control set. RSD of scores for the random set within subregions remained consistently $<$10.0% of the means, demonstrating stability of measurements within the same scale. Additionally, F-tests comparing $GeoFID^{k}_{Rr}$ scores of the random and control sets for each subregion grid show a significant difference between scores for the two sets, indicating that the model has successfully discriminated the low variance random images from the higher variance control images. Together, these findings indicate that $GeoFID^{k}_{rq}$ may be confidently used to detect even small differences in visual appearance between regions of the same scale.

Fig. 9 shows the $GeoFID^{k}_{Rr}$ score curves for the random set across the three subregion grids, demonstrating a linear,

TABLE 1. ***GeoFID*** **calculation between the global region of samples and itself for each of the 3 synthetic shape classes. Values close to 0.0 are expected.**

	Star	Polygon	Ellipse
Random Feature Set	-4.6E-05	-4.7E-05	-3.2E-05
Control Feature Set	-4.8E-05	-7.2E-05	-4.1E-05

TABLE 2. **Relative standard deviation (%) of** ***GeoFID*** **scores within subgrids for each synthetic shape class showing (A) low RSD for each sub-scale of the random-feature set and (B) higher RSD of each sub-scale of the controlled-feature set.**

A		Random Set		
		Star	*Polygon*	*Ellipse*
Sub-Grid	2x2	6.7	3.1	4.7
	3x3	8.2	7	7.4
	4x4	10	7.8	5.6
B		**Control Set**		
		Star	*Polygon*	*Ellipse*
Sub-Grid	2x2	34.3	34.8	12.3
	3x3	15.9	21.8	17.9
	4x4	22.2	22.5	16.7

increasing effect of the MAUP on GeoFID_{Rr}^{k} across scales. As discussed previously, there is no true difference (i.e., dissimilarity) between the random images from the region as a whole and any subregion therein, therefore we extrapolate the effect of the MAUP on all three synthetic shape classes to determine the scale at which GeoFID_{Rr}^{k} degrades past the qualitative threshold of 50.0 (the threshold of noticeable visual dissimilarity). This threshold is reached at or near the 6th level subgrid (6 × 6 division of the WGS84 coordinate system in this case), suggesting that GeoFID_{Rr}^{k} measures of dissimilarity for features of class k between a global region R and one of its subregions r may become unreliable when r is at a scale smaller than a 6 × 6 division of R. The nature and magnitude of scale dependence of GeoFID_{rq}^{k} is therefore as expected and permits confident usage of GeoFID_{rq}^{k} to compare regions of moderately different scale.

Conversely to the random set, degrees of visual dissimilarity between the control images from the region as a whole and its subregions clearly exist, and smaller-scale spatial subsets of the control set will have increasingly distinct features compared to those in the region as a whole. Therefore, it is expected that GeoFID_{rq}^{k} will be sensitive to these distinctions.

Fig. 10 shows the GeoFID_{rq}^{k} score curves for the control set of synthetic shape classes, demonstrating a logarithmic increase which exceeds the qualitative threshold of visual dissimilarity of $\mathit{GeoFID}_{Rr}^{k} = 50.0$ at the first level of division (2 × 2 subregion). Importantly, extrapolating the increasing trend of GeoFID_{rq}^{k} scores at smaller scales shows that the qualitative threshold of 'extreme visual dissimilarity' of $\mathit{GeoFID}_{Rr}^{k} \geq 300.0$ (i.e., breakdown of features indicating shared class membership) is not reached until the global sample of control shapes of class k in R is compared with subregions r smaller than the 11th level subgrid (11 × 11 division of the WGS coordinate system in this case). These findings indicate that GeoFID_{rq}^{k} is sufficiently sensitive to true differences in image features between regions, even when those regions are of highly different scale.

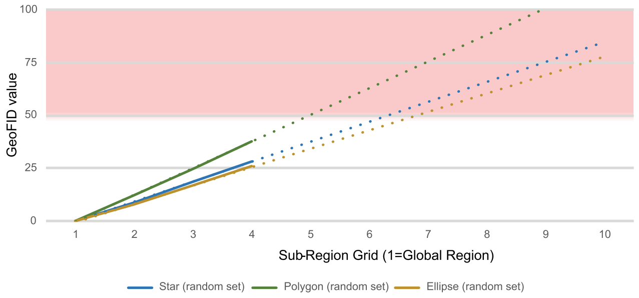

FIGURE 9. **Random feature** ***GeoFID*** **results on the synthetic shapes dataset.** ***GeoFID*** **values between subregions and the global region are not expected to exceed the threshold associated with 'noticeable visual dissimilarity' (*****GeoFID*** **≥ 50.0, red shaded region).**

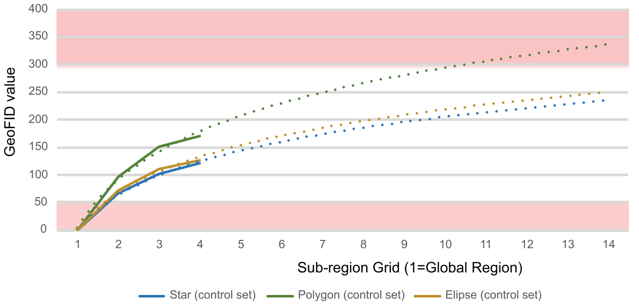

FIGURE 10. **Control feature** ***GeoFID*** **results on the synthetic shapes dataset.** ***GeoFID*** **values are expected to remain above the threshold associated with 'noticeable visual dissimilarity' (*****GeoFID*** **≥ 50.0, lower red shaded region) yet below the threshold associated with 'extreme visual dissimilarity' (*****GeoFID*** **≥ 300.0, upper red-shaded region).**

B. GeoIS

Table 3 shows results of $\mathit{GeoIS}_{r}^{\beta}$ stability tests as relative standard deviation (RSD) of values computed for regions within the same scale on the random feature dataset. RSD of values within subregions remained consistently below 7.0% of the means for all synthetic shape classes, demonstrating a very high level of stability of GeoIS_{r} measurements within the same scale.

Fig. 11 shows results for $\mathit{GeoIS}_{r}^{\beta}$ scale dependence on the random feature set (dark-colored lines) across the three shape classes of the synthetic datasets. As expected, divergence between φ and μ is very low at all spatial scales, resulting in low $\mathit{GeoIS}_{r}^{\beta}$ values. The MAUP tends to have a linearly-degrading effect on calculated $\mathit{GeoIS}_{r}^{\beta}$ values at smaller spatial scales, extrapolated to reach the minimum score (1.0) between the 9×9 and 10×10 subscales across the three shape classes. This suggests that only small deviations

TABLE 3. ***GeoIS*** **stability test results showing percent relative standard deviation (% RSD) of** ***GeoIS*** **values per subgrid region for the random feature synthetic shapes dataset.**

		Star	Polygon	Ellipse
Sub-Grid	2x2	1.69	1.8	1.33
	3x3	2.89	3.43	4.61
	4x4	2.01	6.58	4.31

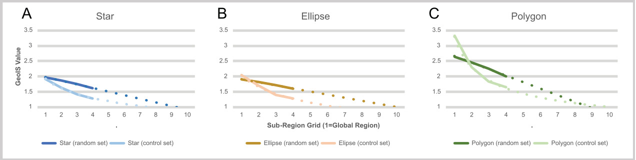

FIGURE 11. ***GeoIS*** **(β) results for (A) star, (B) ellipse and (C) polygon synthetic shapes showing scale dependence (dark lines) and sensitivity (light lines). Low values were expected at all scales due to a combination of low-dimensional features and low object variety.**

in $GeoIS_r^\beta$ values for the same distribution can be expected when repeating the computation at smaller scales, except for very small spatial divisions of the region as a whole.

Sensitivity tests, in the case of $GeoIS_r^\beta$, measure the effect of the baked-in feature variation in the control dataset on resulting $GeoIS_r^\beta$ across spatial scales. Results (Fig. 11, light-colored lines) show a rapid, exponential decline in certainty-variety divergence at smaller spatial scales, extrapolated to reach the minimum score (1.0) between the 6 × 6 and 10 × 10 subscales across the three shape classes.

Interestingly, the highest amount of divergence detected across the three shapes classes occurred in the 'polygon' class. This could have been due to a larger number of object categories detected in the polygon dataset given it contains triangles, squares, pentagons, octagons, and other common shape patterns. Therefore, those shape patterns may have been associated with multiple classes present in the 1000-class ImageNet dataset on which the $GeoIS_r$ CNN is pretrained. This result demonstrates that even a synthetic image dataset specifically constructed with singular object classes may be interpreted as having multiple, distinct classes, depending on the characteristics of the dataset on which the backbone CNN model is pretrained.

VI. DISCUSSION

Results of tests for unity, stability, scale dependence, and sensitivity of $GeoFID_{rq}^k$ on synthetic data bode well for its use as a tool for geospatial analysis of diversity of object class appearance in RSI datasets. Importantly, tests demonstrate that $GeoFID_{rq}^k$ can confidently be extended beyond the use case simulated in the previous experiments (comparing global object class appearance to subregional appearance) and into more holistic and comprehensive pairwise comparisons of object class representation within and between scales. This includes geostatistical modeling of $GeoFID_{rq}^k$ computations at sample locations in a study area (e.g., estimating the effect of geographic distance, direction and/or spatial covariates on image feature similarity as computed by the model), making a range of predictive analyses feasible.

Likewise, tests of the bias/divergence component of $GeoIS_r$ demonstrate a high level of stability within scales and low levels of between scale variation due to the MAUP. These results indicate $GeoIS_r$ is a reliable tool for geospatial analysis of contextual diversity of object classes over the extent of RSI datasets. Key to $GeoIS_r$ utility in the RS domain is its disaggregation of the original Inception Score into components to measure co-occurring object density ($GeoIS_r^\varphi$), co-occurring object variety ($GeoIS_r^\mu$) and the bias/divergence between the two ($GeoIS_r^\beta$).

As a result of the rigorous battery of experiments to verify the reliability of $GeoFID_{rq}^k$ and $GeoIS_r^\beta$, it can be stated that both algorithms clearly perform within the criteria for success (as noted in Sections IV and V) as spatial dataset analysis methods.

Whereas the *FID* and *IS* algorithms as initially introduced in the literature are often applied unaltered to the evaluation of synthetic image quality, thoughtful model choices will be required for $GeoFID_{rq}^k$ and $GeoIS_r$ to be appropriately applied to geodiversity related analyses in the RS domain. This includes the choice of CNN architecture as well as the dataset on which the architecture is pre-trained. For example, image features and object classes transferred from overhead imagery training datasets may be more appropriate than those found in ImageNet or similar generic image datasets [84]. Additionally, spatial size and shape of image chips around target labels will control the amount of geographic context that is learned by the model. In the case of $GeoFID_{rq}^k$, this parameter choice should reflect the spatial footprint of the object class/phenomena of interest, and in the case of $GeoIS_r$, it should reflect a spatial extent congruent with the co-occurring objects of interest. Moreover, in the case of $GeoFID_{rq}^k$, extraction of feature vectors from layers other than the final pooling layer of the chosen CNN architecture may be warranted if the focus of analysis is on low-level image features (e.g., edges, textures, colors) or if comparison of a comprehensive set of high- and low-level features is desired. Finally, the components of geodiversity modelled by $GeoIS_r$ and $GeoFID_{rq}^k$ can also be modelled by other types of algorithms that deserve exploration, especially in light of the recognized shortcomings of both *IS* and *FID* [86], [91]–[93]. For example, CNN-based object detectors could be used to quantify co-occurring object density and variety, whereas different feature comparison measures (e.g., topology- [94] or geometry-based [95] distance metrics) could quantify changes in object class appearance over the landscape. The pros and cons of applying these related methods to the RS domain have yet to be tested.

Although this paper focuses on the utility of the proposed methods for exploratory geodiversity analysis to uncover and understand geographic bias in RSI datasets in the context of

TABLE 4. The novelty of deep learning-based geodiversity metrics for RSI data is apparent in the multidisciplinary geoscience and remote sensing literature. Here, attributes of the most closely related works are compared to the current manuscript.

Reference (Year)	Focus on dataset analysis?	Inputs are images?	Inputs contain precise (i.e., lat / lon) georeference?	Aggregates inputs to arbitrary spatial scales / regions?	Algorithms tailored for RSI?	Deep learning-based algorithm / analysis?	Object-level analysis?	Includes entropy / category-based metric?	Includes similarity / feature distance-based metric?
[67] (2017)	■	■						■	
[104] (2007)	■		■					■	
[105] (2019)		■					■	■	■
[78] (2019)	■	■					■		■
[106] (2019)		■				■	■		■
[81] (2009)	■	■	■		■				■
[107] (2020)	■	■				■	■	■	■
Current Manuscript (2021)	■	■	■	■	■	■	■	■	■

improving discriminative modeling, applications to generative modeling and geospatial simulation are apparent. Generative modeling in the RS domain is in its infancy, yet the consequences of using datasets and/or models with unevaluated geographic bias for tasks such as RSI dataset augmentation, style transfer and image interpolation/ completion can be as problematic as those uncovered in discriminative applications. The extent and characteristics of visual variation and contextual diversity in RSI datasets, discoverable using the methods introduced here, can be used for geographically informed generative modeling and lead to more realistic and un-biased RSI data synthesis. Similarly, the prospect of extending image-based deep feature geodiversity measures to analyze attribute geodiversity of nonimage geodata types (i.e., attributed points, polylines, and polygons) may prove fruitful.

VII. CONCUSION

This paper argues that a gap of understanding exists in the level of meaningful and representative geographic variation present in the growing corpus of large RSI datasets, and that this gap risks perpetuation of biased data and models and undermines the application of AI-based vision systems to geospatial object detection and classification tasks. Methods were introduced for measuring various facets of geodiversity in RSI dataset object classes (spatial comparison of object class appearance in the case of $GeoFID^k_{rq}$, and measures of contextual heterogeneity in the case of $GeoIS_r$). Several experiments were devised to explore the stability, scale dependence, and sensitivity of $GeoIS_r$ and $GeoFID^k_{rq}$. To this end, synthetic random and control datasets were developed to benchmark the ability of these metrics to discriminate among regionalized phenomena. It was found that both metrics were stable, sensitive to small feature changes and not overly affected by the Modifiable Areal Unit Problem.

Given the analytical characteristics of $GeoIS_r$ and $GeoFID^k_{rq}$, they have tremendous practical utility in application to large RSI datasets, including those containing labels for discrete objects [18], [19], [87], land cover patches [96], [97], activity/change annotations [22], [98] and other observable phenomena. Providing baseline feature geodiversity metadata for RSI datasets with metrics like $GeoIS_r$ and $GeoFID^k_{rq}$ would add significant interpretative and diagnostic value and would help develop reflexive and transparent data practices [99], [100] in the remote sensing domain. However, further work will be needed to refine both $GeoFID^k_{rq}$ and $GeoIS_r$ considering the shortcomings of *FID* and *IS* algorithms referenced earlier. More methodologies will also be needed to fully model the concept of geodiversity in RSI data, including those that take into account temporal variation and/or exploit radiometric and spectral properties of different imaging systems. Moreover, computational solutions to other diversity-related RSI dataset analyses are needed to help researchers understand biases in both input data and trained models. For example, visualization of CNN-extracted features for an object class' samples aggregated on a per-region basis (country, administrative unit, etc.) could reveal the 'average' appearance of that class within each region, thereby detecting collection bias *before any final models are trained and deployed.* Above all, multidisciplinary perspectives as well as input from a variety of stakeholders in academia, government and industry will be required to guide future directions of geodiversity research in the geoscience and remote sensing domains.

REFERENCES

[1] J. Piaget, *Child's Conception of Space: Selected Works*, vol. 4. Evanston, IL, USA: Routledge, 2013.

[2] D. Marr, "The philosophy and the approach," in *Vision*. San Francisco, CA, USA: Freeman, 1982.

[3] E. L. Ullman and R. R. Boyce, *Geography as Spatial Interaction*. Seattle, WA, USA: Univ. of Washington Press, 1980.

[4] P. Klapka and M. Halás, ''Conceptualising patterns of spatial flows: Five decades of advances in the definition and use of functional regions,'' *Moravian Geograph. Rep.*, vol. 24, no. 2, pp. 2–11, Jun. 2016.

[5] R. L. Morrill, *The Spatial Organization of Society*. Pacific Grove, CA, USA: Duxbury, 1970.

[6] R. G. Golledge, *Spatial Behavior: A Geographic Perspective*. New York, NY, USA: Guilford, 1997.

[7] C. Lengen and T. Kistemann, ''Sense of place and place identity: Review of neuroscientific evidence,'' *Health Place*, vol. 18, no. 5, pp. 1162–1171, Sep. 2012.

[8] M. Li, S. Zang, B. Zhang, S. Li, and C. Wu, ''A review of remote sensing image classification techniques: The role of spatio-contextual information,'' *Eur. J. Remote Sens.*, vol. 47, no. 1, pp. 389–411, Jan. 2014, doi: 10.5721/EuJRS20144723.

[9] T. Blaschke, ''Object based image analysis for remote sensing,'' *ISPRS J. Photogram. Remote Sens.*, vol. 65, no. 1, pp. 2–16, 2010.

[10] W. R. Tobler, ''A computer movie simulating urban growth in the detroit region,'' *Econ. Geogr.*, vol. 46, no. 1, pp. 234–240, 1970.

[11] M. F. Goodchild, ''Geographical data modeling,'' *Comput. Geosci.*, vol. 18, no. 4, pp. 401–408, May 1992.

[12] N. W. Campbell, W. P. J. Mackeown, B. T. Thomas, and T. Troscianko, ''Interpreting image databases by region classification,'' *Pattern Recognit.*, vol. 30, no. 4, pp. 555–563, Apr. 1997.

[13] K. Aizawa, K. Sakaue, and Y. Suenaga, *Image Processing Technologies: Algorithms, Sensors, and Applications*. Boca Raton, FL, USA: CRC Press, 2004.

[14] W. Rawat and Z. Wang, ''Deep convolutional neural networks for image classification: A comprehensive review,'' *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, 2017.

[15] M. Goodchild, ''Citizens as sensors: The world of volunteered geography,'' *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

[16] G. B. Schmidt and W. M. Jettinghoff, ''Using amazon mechanical turk and other compensated crowdsourcing sites,'' *Bus. Horizons*, vol. 59, no. 4, pp. 391–400, Jul. 2016.

[17] *Airbus Announces Finalists of Multi-Data Challenge Within Copernicus Masters 2018*, Airbus, Leiden, The Netherlands, 2018.

[18] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, ''xView: Objects in context in overhead imagery,'' 2018, *arXiv:1802.07856*. [Online]. Available: https://arxiv.org/abs/1802.07856

[19] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, ''SpaceNet: A remote sensing dataset and challenge series,'' 2018, *arXiv:1807.01232*. [Online]. Available: https://arxiv.org/abs/1807.01232

[20] *Open AI challenge: Aerial Imagery of South Pacific Islands*, WeRobotics, Wilmington, DE, USA, 2018.

[21] S. Marconi, S. J. Graves, D. Gong, M. S. Nia, M. L. Bras, B. J. Dorr, P. Fontana, J. Gearhart, C. Greenberg, D. J. Harris, S. A. Kumar, A. Nishant, J. Prarabdh, S. U. Rege, S. A. Bohlman, E. P. White, and D. Z. Wang, ''A data science challenge for converting airborne remote sensing data into ecological information,'' *PeerJ*, vol. 6, p. e5843, Feb. 2019.

[22] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston, ''Creating xBD: A dataset for assessing building damage from satellite imagery,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 10–17.

[23] R. Cardillo, *Small Satellites: Big Data*. Springfield, VA, USA: National Geospatial-Intelligence Agency, 2017.

[24] X. Deng, P. Liu, X. Liu, R. Wang, Y. Zhang, J. He, and Y. Yao, ''Geospatial big data: New paradigm of remote sensing applications,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 10, pp. 3841–3851, Oct. 2019.

[25] Z. Jiang and S. Shekhar, *Spatial Big Data Science: Classification Techniques for Earth Observation Imagery*. Cham, Switzerland: Springer, 2017.

[26] J.-G. Lee and M. Kang, ''Geospatial big data: Challenges and opportunities,'' *Big Data Res.*, vol. 2, no. 2, pp. 74–81, 2015.

[27] T. Stryker and P. Colohan, ''The national plan for civil Earth observations,'' Office Sci. Technol. Policy, Executive Office President United States, Washington, DC, USA, Tech. Rep., 2014.

[28] J. Doshi, S. Basu, and G. Pang, ''From satellite imagery to disaster insights,'' 2018, *arXiv:1812.07033*. [Online]. Available: http://arxiv.org/abs/1812.07033

[29] S. Weichenthal, M. Hatzopoulou, and M. Brauer, ''A picture tells a thousand exposures: Opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology,'' *Environ. Int.*, vol. 122, pp. 3–10, Jan. 2019.

[30] J. Hruška, T. Adão, L. Pádua, P. Marques, A. Cunha, E. Peres, A. Sousa, R. Morais, and J. J. Sousa, ''Machine learning classification methods in hyperspectral data processing for agricultural applications,'' in *Proc. Int. Conf. Geoinform. Data Anal.*, Apr. 2018, pp. 137–141.

[31] E. Suel, J. W. Polak, J. E. Bennett, and M. Ezzati, ''Measuring social, environmental and health inequalities using deep learning and street imagery,'' *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.

[32] F. Maire, L. M. Alvarez, and A. Hodgson, ''Automating marine mammal detection in aerial images captured during wildlife surveys: A deep learning approach,'' in *Proc. Australas. Joint Conf. Artif. Intell.* Canberra, ACT, Australia: Springer, 2015, pp. 379–385.

[33] C.-O. Dufresne-Camaro, F. Chevalier, and S. I. Ahmed, ''Computer vision applications and their ethical risks in the global south,'' in *Proc. Graph. Interface Conf.*, 2020, pp. 1–10. [Online]. Available: https://openreview.net/forum?id=QLFSDNIvI

[34] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, ''Meta-learning in neural networks: A survey,'' 2020, *arXiv:2004.05439*. [Online]. Available: http://arxiv.org/abs/2004.05439

[35] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, ''Deep learning: New computational modelling techniques for genomics,'' *Nature Rev. Genet.*, vol. 20, no. 7, pp. 389–403, Jul. 2019.

[36] T. Young, D. Hazarika, S. Poria, and E. Cambria, ''Recent trends in deep learning based natural language processing,'' *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[37] L. Zhang, Y. Li, X. Xiao, X.-Y. Li, J. Wang, A. Zhou, and Q. Li, ''CrowdBuy: Privacy-friendly image dataset purchasing via crowdsourcing,'' in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 2735–2743.

[38] B. Zhao, S. Tang, X. Liu, X. Zhang, and W.-N. Chen, ''IronM: Privacy-preserving reliability estimation of heterogeneous data for mobile crowdsensing,'' *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5159–5170, Jun. 2020.

[39] H. Nguyen, L.-M. Kieu, T. Wen, and C. Cai, ''Deep learning methods in transportation domain: A review,'' *IET Intell. Transp. Syst.*, vol. 12, no. 9, pp. 998–1004, 2018.

[40] R. M. Sherman and S. L. Salzberg, ''Pan-genomics in the human genome era,'' *Nature Rev. Genet.*, vol. 21, pp. 243–254, Feb. 2020.

[41] G. Sirugo, S. M. Williams, and S. A. Tishkoff, ''The missing diversity in human genetic studies,'' *Cell*, vol. 177, no. 1, pp. 26–31, Mar. 2019.

[42] S. Ballouz, A. Dobin, and J. A. Gillis, ''Is it time to change the reference genome?'' *Genome Biol.*, vol. 20, no. 1, pp. 1–9, Dec. 2019.

[43] E. Dinan, A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston, ''Queens are powerful too: Mitigating gender bias in dialogue generation,'' 2019, *arXiv:1911.03842*. [Online]. Available: http://arxiv.org/abs/1911.03842

[44] N. Swinger, M. De-Arteaga, N. T. Heffernan, IV, M. D. Leiserson, and A. T. Kalai, ''What are the biases in my word embedding?'' in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jan. 2019, pp. 305–311.

[45] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley, ''No classification without representation: Assessing geodiversity issues in open data sets for the developing world,'' 2017, *arXiv:1711.08536*. [Online]. Available: http://arxiv.org/abs/1711.08536

[46] T. de Vries, I. Misra, C. Wang, and L. van der Maaten, ''Does object recognition work for everyone?'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 52–59.

[47] K. Vodrahalli, K. Li, and J. Malik, ''Are all training examples created equal? An empirical study,'' 2018, *arXiv:1811.12569*. [Online]. Available: http://arxiv.org/abs/1811.12569

[48] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, ''Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy,'' in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 547–558.

[49] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith, ''Diversity in faces,'' 2019, *arXiv:1901.10436*. [Online]. Available: http://arxiv.org/abs/1901.10436

[50] B. J. Hecht and M. Stephens, ''A tale of cities: Urban biases in volunteered geographic information,'' in *Proc. ICWSM*, 2014, vol. 14, no. 14, pp. 197–205.

[51] G. Zhang and A.-X. Zhu, ''The representativeness and spatial bias of volunteered geographic information: A review,'' *Ann. GIS*, vol. 24, no. 3, pp. 151–162, Jul. 2018.

[52] R. Kitchin, "Big data and human geography: Opportunities, challenges and risks," *Dialogues Hum. Geogr.*, vol. 3, no. 3, pp. 262–267, Nov. 2013.

[53] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. D. Langhans, M. Tegmark, and F. F. Nerini, "The role of artificial intelligence in achieving the sustainable development goals," *Nature Commun.*, vol. 11, no. 1, pp. 1–10, 2020.

[54] T. Salem, S. Workman, and N. Jacobs, "Learning a dynamic map of visual appearance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 12435–12444.

[55] Y. Nachmany and H. Alemohammad, "Detecting roads from satellite imagery in the developing world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 83–89.

[56] M. De-Arteaga, W. Herlands, D. B. Neill, and A. Dubrawski, "Machine learning for the developing world," *ACM Trans. Manage. Inf. Syst.*, vol. 9, no. 2, pp. 1–14, 2018.

[57] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[58] Amazon. (2018). *Geodiverse Open Training Data as a Global Good.* [Online]. Available: https://aws.amazon.com/blogdivers/publicsector/geo-diverse-open-training-data-as-a-global-public-good/

[59] D. Bollinger. (2018). Geo-diversity for better, fairer machine learning. Development Seed. [Online]. Available: https://medium.com/devseed/geo-diversity-for-better-fairer-machine-learning-5c64021708dd

[60] Office of the Director of National Intelligence. (2020). *Artificial Intelligence Ethics Framework for the Intelligence Community*. [Online]. Available: https://www.intelligence.gov/images/AI/AI_Ethics_Framework_for_the_Intelligence_Community_1.0.pdf

[61] T. Hellström, V. Dignum, and S. Bensch, "Bias in machine learning—What is it good for?" 2020, *arXiv:2004.00686*. [Online]. Available: http://arxiv.org/abs/2004.00686

[62] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2019, *arXiv:1908.09635*. [Online]. Available: http://arxiv.org/abs/1908.09635

[63] A. H. Strahler, C. E. Woodcock, and J. A. Smith, "On the nature of models in remote sensing," *Remote Sens. Environ.*, vol. 20, no. 2, pp. 121–139, Oct. 1986.

[64] G. Castilla and G. J. Hay, "Image objects and geographic objects," in *Object-Based Image Analysis*. Berlin, Germany: Springer, 2008, pp. 91–110.

[65] D. Arvor, L. Durieux, S. Andrés, and M.-A. Laporte, "Advances in geographic object-based image analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective," *ISPRS J. Photogram. Remote Sens.*, vol. 82, pp. 125–137, Aug. 2013, doi: 10.1016/j.isprsjprs.2013.05.003.

[66] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 325–333.

[67] P. Khanzadi, B. Majidi, and E. Akhtarkavan, "A novel metric for digital image quality assessment using entropy-based image complexity," in *Proc. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. (KBEI)*, Dec. 2017, pp. 0440–0445.

[68] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[69] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, no. 2, pp. 363–375, 2006.

[70] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[71] T. Li, S. Ma, and M. Ogihara, "Entropy-based criterion in categorical clustering," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 68.

[72] E. J. Dudewicz and E. C. Van Der Meulen, "Entropy-based tests of uniformity," *J. Amer. Stat. Assoc.*, vol. 76, no. 376, pp. 967–974, Dec. 1981.

[73] A. Holzinger, A. Holzinger, M. Hörtenhuber, C. Mayer, M. Bachler, S. Wassertheurer, A. J. Pinho, and D. Koslicki, "On entropy-based data mining," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Berlin, Germany: Springer, 2014, pp. 209–226.

[74] M. Ceccarello, A. Pietracaprina, and G. Pucci, "A general coreset-based approach to diversity maximization under matroid constraints," *ACM Trans. Knowl. Discovery Data*, vol. 14, no. 5, Aug. 2020, Art. no. 60, doi: 10.1145/3402448.

[75] M. Ceccarello, A. Pietracaprina, G. Pucci, and E. Upfal, "MapReduce and streaming algorithms for diversity maximization in metric spaces of bounded doubling dimension," *Proc. VLDB Endowment*, vol. 10, no. 5, pp. 469–480, Jan. 2017.

[76] A. Li, L. Zhang, J. Qian, X. Xiao, X.-Y. Li, and Y. Xie, "TODQA: Efficient task-oriented data quality assessment," in *Proc. 15th Int. Conf. Mobile Ad-Hoc Sensor Netw. (MSN)*, Dec. 2019, pp. 81–88.

[77] T. Wu, L. Chen, P. Hui, C. J. Zhang, and W. Li, "Hear the whole story: Towards the diversity of opinion in crowdsourcing markets," *Proc. VLDB Endowment*, vol. 8, no. 5, pp. 485–496, Jan. 2015.

[78] X. Xiao, L. Zhang, and X.-Y. Li, "Noisy data collection towards diversity maximization," in *Proc. 5th Int. Conf. Big Data Comput. Commun. (BIGCOM)*, Aug. 2019, pp. 283–287.

[79] B. O. Ayinde, T. Inanc, and J. M. Zurada, "Regularizing deep neural networks by enhancing diversity in feature extraction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2650–2661, Sep. 2019, doi: 10.1109/TNNLS.2018.2885972.

[80] F. Yang, A.-X. Zhu, K. Ichii, M. A. White, H. Hashimoto, and R. R. Nemani, "Assessing the representativeness of the AmeriFlux network using MODIS and GOES data," *J. Geophys. Res., Biogeosci.*, vol. 113, no. G4, Dec. 2008, Art. no. G04036.

[81] M. O. Román, C. B. Schaaf, C. E. Woodcock, A. H. Strahler, X. Yang, R. H. Braswell, P. S. Curtis, K. J. Davis, D. Dragoni, M. L. Goulden, L. Gu, D. Y. Hollinger, T. E. Kolb, T. P. Meyers, J. W. Munger, J. L. Privette, A. D. Richardson, T. B. Wilson, and S. C. Wofsy, "The MODIS (collection V005) BRDF/albedo product: Assessment of spatial representativeness over forested landscapes," *Remote Sens. Environ.*, vol. 113, no. 11, pp. 2476–2498, Nov. 2009, doi: 10.1016/j.rse.2009.07.009.

[82] N. Schutgens, S. Tsyro, E. Gryspeerdt, D. Goto, N. Weigum, M. Schulz, and P. Stier, "On the spatio-temporal representativeness of observations," *Atmos. Chem. Phys.*, vol. 17, no. 16, pp. 9761–9780, Aug. 2017.

[83] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[84] L. N. Vasershtein, "Markov processes over denumerable products of spaces, describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–72, 1969.

[85] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[86] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? A large-scale study," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 700–709.

[87] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6172–6180.

[88] S. Openshow, "A million or so correlation coefficients, three experiments on the modifiable areal unit problem," *Stat. Appl. Spatial Sci.*, pp. 127–144, Jan. 1979.

[89] D. Josselin and R. Louvet, "Impact of the scale on several metrics used in geographical object-based image analysis: Does GEOBIA mitigate the modifiable areal unit problem (MAUP)?" *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 3, p. 156, Mar. 2019.

[90] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[91] S. Barratt and R. Sharma, "A note on the inception score," 2018, *arXiv:1801.01973*. [Online]. Available: http://arxiv.org/abs/1801.01973

[92] M. J. Chong and D. Forsyth, "Effectively unbiased FID and inception score and where to find them," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6070–6079.

[93] A. Borji, "Pros and cons of GAN evaluation measures," *Comput. Vis. Image Understand.*, vol. 179, pp. 41–65, Feb. 2019.

[94] D. Horak, S. Yu, and G. Salimi-Khorshidi, "Topology distance: A topology-based approach for evaluating generative adversarial networks," 2020, *arXiv:2002.12054*. [Online]. Available: http://arxiv.org/abs/2002.12054

[95] V. Khrulkov and I. Oseledets, "Geometry score: A method for comparing generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2621–2629.

[96] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[97] M. Brown, H. Goldberg, K. Foster, A. Leichtman, S. Wang, S. Hagstrom, M. Bosch, and S. Almes, "Large-scale public LiDAR and satellite image data set for urban semantic labeling," *Proc. SPIE*, vol. 10636, May 2018, Art. no. 106360P.

[98] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-action: An outdoor recorded drone video dataset for action recognition," *Drones*, vol. 3, no. 4, p. 82, Nov. 2019. [Online]. Available: https://www.mdpi.com/2504-446X/3/4/82

[99] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford, "Datasheets for datasets," 2018, *arXiv:1803.09010*. [Online]. Available: http://arxiv.org/abs/1803.09010

[100] M. Miceli, T. Yang, L. Naudts, M. Schuessler, D. Serbanescu, and A. Hanna, "Documenting computer vision datasets: An invitation to reflexive data practices," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Mar. 2021, pp. 161–172.

[101] D. McDuff, R. Cheng, and A. Kapoor, "Identifying bias in AI using simulation," 2018, *arXiv:1810.00471*. [Online]. Available: http://arxiv.org/abs/1810.00471

[102] Google, CNES/Airbus/Maxar Technologies, Voinjama Multilateral High School, Voinjama, Liberia. (2021). *Imagery 2021*. [Online]. Available: https://www.google.com/maps/place/Voinjama+Multilateral+High+School/@8.415964,-9.7518347,483m

[103] Google, Houston-Galveston Area Council/Maxar Technologies/Texas General Land Office/US Geological Survey, Waltrip High School, Houston, TX, USA. (2021). *Imagery 2021*. [Online]. Available: https://www.google.com/maps/place/Waltrip+High+School/@29.8177798,-95.4341409,684m

[104] C. Yesson, P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, R. J. White, A. C. Jones, F. A. Bisby, and A. Culham, "How global is the global biodiversity information facility?" *PLoS ONE*, vol. 2, no. 11, p. e1124, Nov. 2007.

[105] E. Creager, D. Madras, J. H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1436–1445.

[106] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jan. 2019, pp. 289–295.

[107] A. Wang, A. Narayanan, and O. Russakovsky, "REVISE: A tool for measuring and mitigating bias in visual datasets," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 733–751.

[108] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019.

[109] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020.

[110] A. Rajagopal, G. P. Joshi, A. Ramachandran, R. T. Subhalakshmi, M. Khari, S. Jha, K. Shankar, and J. You, "A deep learning model based on multi-objective particle swarm optimization for scene classification in unmanned aerial vehicles," *IEEE Access*, vol. 8, pp. 135383–135393, 2020.

[111] W. Li, H. Liu, Y. Wang, Z. Li, Y. Jia, and G. Gui, "Deep learning-based classification methods for remote sensing images in urban built-up areas," *IEEE Access*, vol. 7, pp. 36274–36284, 2019.

[112] X. Gao, X. Sun, Y. Zhang, M. Yan, G. Xu, H. Sun, J. Jiao, and K. Fu, "An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network," *IEEE Access*, vol. 6, pp. 39401–39414, 2018.

[113] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, 2018.

[114] Z. Bessinger and N. Jacobs, "A generative model of worldwide facial appearance," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1569–1578, doi: 10.1109/WACV.2019.00172.

[115] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38544–38555, 2018.

[116] Q. Shi, X. Liu, and X. Li, "Road detection from remote sensing images by generative adversarial networks," *IEEE Access*, vol. 6, pp. 25486–25494, 2017.

[117] Z. Wang, C. Zou, and W. Cai, "Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model," *IEEE Access*, vol. 8, pp. 71353–71363, 2020.

[118] A. Abdollahi, B. Pradhan, and A. Alamri, "VNet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data," *IEEE Access*, vol. 8, pp. 179424–179436, 2020.

[119] D. Hou, Z. Miao, H. Xing, and H. Wu, "V-RSIR: An open access web-based image annotation tool for remote sensing image retrieval," *IEEE Access*, vol. 7, pp. 83852–83862, 2019.

[120] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Mar. 2021, pp. 610–623.

[121] G. E. Hinton, "To recognize shapes, first learn to generate images," *Prog. Brain Res.*, vol. 165, no. 6, pp. 535–547, 2007.

[122] K.-T. Chang, A. Merghadi, A. P. Yunus, B. T. Pham, and J. Dou, "Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques," *Sci. Rep.*, vol. 9, no. 1, pp. 1–21, Dec. 2019.

[123] K. von Neumann-Cosel, E. Roth, D. Lehmann, J. Speth, and A. Knoll, "Testing of image processing algorithms on synthetic data," in *Proc. 4th Int. Conf. Softw. Eng. Adv.*, Sep. 2009, pp. 169–172.

[124] K. A. Redmill, J. I. Martin, and U. Ozgliner, "Virtual environment simulation for image processing sensor evaluation," in *Proc. IEEE Intell. Transp. Syst. (ITSC)*, Oct. 2000, pp. 64–70.

[125] F. E. Fassnacht, H. Latifi, and F. Hartig, "Using synthetic data to evaluate the benefits of large field plots for forest biomass estimation with LiDAR," *Remote Sens. Environ.*, vol. 213, pp. 115–128, Aug. 2018.

[126] L. Polidori and M. El Hage, "Digital elevation model quality assessment methods: A critical review," *Remote Sens.*, vol. 12, no. 21, p. 3522, Oct. 2020.

[127] Ø. Frette, S. R. Erga, J. J. Stamnes, and K. Stamnes, "Optical remote sensing of waters with vertical structure," *Appl. Opt.*, vol. 40, no. 9, pp. 1478–1487, 2001.

AARON M. WESLEY (Member, IEEE) received the M.A. degree in geography from the University of Missouri, Columbia, MO, USA, in 2014, where he is currently pursuing the Ph.D. degree in informatics with the Institute for Data Science and Informatics. He is also a Data Scientist with the National Geospatial-Intelligence Agency, Saint Louis, MO, USA, supporting the Agency's humanitarian assistance and disaster relief (HA/DR) mission.

TIMOTHY C. MATISZIW (Senior Member, IEEE) received the Ph.D. degree in geography from The Ohio State University, Columbus, OH, USA, in 2005. He is currently an Associate Professor with the Department of Civil and Environmental Engineering, the Department of Geography, and the Institute of Data Science and Informatics, University of Missouri, Columbia, MO, USA. He is a fellow of the Royal Geographical Society and a member of the American Association for the Advancement of Science (AAAS), the American Association of Geographers (AAG), the American Geographical Society (AGS), the American Geophysical Union (AGU), the American Society of Civil Engineers (ASCE), the Association for Computing Machinery (ACM), the Regional Science Association International (RSAI), the Sigma Xi, the Institute for Operations Research and the Management Sciences (INFORMS), the International Society for Computational Biology (ISCB), and the United States Geospatial Intelligence Foundation (USGIF).

• • •