THE COLLEGE OF WILLIAM & MARY

DISSERTATION PROSPECTUS

Spatially Aware Convolutional Networks: Architectures, Algorithms and Applications

---

Author:
Heather BAIER

Advisor:
Dan RUNFOLA

A prospectus submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in the

Computational Geography Specialization
Department of Applied Science

August 28, 2024

Contents

1 Introduction **1**

2 Literature Review **5**

2.1 Deep Learning and Satellite Imagery to Predict Sociodemographic Data 5

2.1.1 A Brief History of Satellite Imagery Sources 7

2.1.2 Data Sources Leveraged in Deep Learning Models 8

2.2 Spatially-Aware Convolutional Neural Networks 9

2.3 Application Area: United Nations Sustainable Development Goals 10

2.3.0.1 Structure of the SDGs 10

2.4 Benchmark Datasets for Deep Learning with Satellite Imagery . . 11

3 Quantitative Analysis **14**

3.1 Introduction to Quantitative Analysis 14

3.2 Background Literature 15

3.2.1 Socioeconomic Indicator Prediction using Satellite Imagery and Deep Learning 15

3.2.2 Spatially-Aware Deep Learning Models 16

3.3 Data 17

3.3.1 USAID Demographic and Health Survey 17

3.3.2 Planet Basemaps Satellite Imagery 20

3.4 Methodology 21

3.4.1 Framework Overview 21

3.4.2 Feature Extraction 23

3.4.2.1 Proposed Candidates 23

3.4.2.2 Cross-Border Candidate Selection 23

3.4.2.3 Source and Candidate Feature Extraction 23

3.4.3 Contextual Embedding 24

3.4.4 Spatial Embedding 24
3.4.5 Fully Connected Layers for Final Prediction 25
3.5 Experiments 25
3.5.1 Implementation Details and Model Selection 25
3.5.2 Baseline & SOTA Models for Comparison 26
3.5.2.1 Baseline: DeepAll 27
3.5.2.2 Domain generalization using a mixture of multiple latent domains (DGMMLD) 27
3.5.2.3 Spatially-Aware Network 28
3.6 Results 28
3.6.1 Evaluation 28
3.7 Discussion 29
3.7.1 Limitations 30

4 Proposed Structure for Dissertation **33**
4.1 Chapter 1: Proximity-driven embedding of Visual, Spatial and Contextual features for the estimation of wealth in data-sparse regions 33
4.1.1 Motivation and Synopsis 33
4.1.2 Limitations and Plan for Improvement 34
4.1.2.1 Proposed Candidate Images 34
4.1.2.2 Alternative Candidate Selection Mechanisms . . 37
4.1.2.3 Regularization 37
4.2 Chapter 2: Spatially Adaptive Convolutional Networks with Coordinate-Conditioned Convolutional Layers 39
4.2.1 Motivation and Synopsis 39
4.2.2 Methodology 40
4.2.2.1 HyperNet 41
4.2.2.2 Adaptive Convolutional Layer 42
4.2.2.3 Layer Placement 42
4.2.3 Experiments 44
4.2.4 Preliminary Results 44
4.2.5 Discussion 45
4.2.6 Limitations and Plan for Improvement 46

4.2.6.1 Layer Placement 47
4.3 Chapter 3: torchSDG 48
4.3.1 Motivation and Synopsis 48
4.3.2 Data 49
4.3.2.1 SDG4 Tabular Datasets 49
4.3.2.2 Meta High Resolution Population Density Maps 52
4.3.2.3 Public School Data 52
4.3.2.4 Satellite Imagery 53
4.3.3 Methods 53
4.3.3.1 Grid Level 54
4.3.3.2 School Level Predictions 55

5 Timeline **56**

Bibliography **58**

List of Figures

3.1 Dataset coverage map. 7 countries (in purple) represent our source countries and 2 countries (in red) represent our target countries. . 19
3.2 Comparative histograms of the mean-normalized Wealth Index between source data (aggregated from multiple countries) and target data (aggregated from selected countries). 20
3.3 Histograms displaying the distributions of the mean-normalized Wealth Index in the four target countries: Honduras, Rwanda, Tajikistan, and Timor-Leste. Each histogram shows the frequency of data points at different Wealth Index levels 21
3.4 Examples of imagery clips 22
3.5 The proposed GeoEmbed framework 22
3.6 ResNet18 architecture 26
3.7 DGMMLD architecture 27

4.1 Each row in both this figure and figure 4.1 is an input image and it's 8 proposed candidates. The input image in both figures is circled in orange, the candidate images the model selected are in green, and the remaining candidates are in red. As demonstrated, the candidates (anything in green or red) are all very similar in their content, representing very little feature diversity. 35
4.2 Outside country example 36
4.3 Example candidates based on style statistics 36
4.4 HyperNet Architecture 41
4.5 ResNet18 Architecture 42
4.6 School imagery examples 55

5.1 Timeline 57

List of Tables

2.1 Satellites Commonly Used in Deep Learning Models 7

3.1 Comparison of Models 28
3.2 Comparison of R2 Scores Across Target Countries 29

4.1 Comparison of Models 34
4.2 Experiment Results Summary 44
4.3 Educational Subindicators, Data Sources, and Prediction Granularity 50

List of Abbreviations

CNN	Convolutional Neural Network
SDG	Sustainable Development Goal
DHS	Demographic and Health Surveys
LSMS	Living Standards Measurement Study
NDVI	Normalized Difference Vegetation Index
GAN	Generative Adversarial Network
SVANN	Spatial Variability Aware Deep Neural Networks
OSFA	One-Size-Fits-All
SPAGNN	Spatially-Aware Graph Neural Networks
SHGNN	Spatial Heterophily Graph Neural Network
USGS	U.S. Geological Survey
ESA	European Space Agency
MODIS	Moderate Resolution Imaging Spectroradiometer
VIIRS	Visible Infrared Imaging Radiometer Suite
PCA	Principal Component Analysis
USAID	United States Agency for International Development
MLP	Multi-Layer Perceptron
DGMMLD	Domain Generalization using a Mixture of Multiple Latent Domains
SOTA	State Of The Art
GANs	Generative Adversarial Networks

Chapter 1

Introduction

Over the last decade, the use of convolutional neural networks (CNNs) with satellite imagery to predict sociodemographic data has become increasingly popular. This trend began with Jean et al's [32]'s work, which used satellite imagery to predict poverty in African nations. Since then, deep learning and satellite imagery methods have proven effective in predicting a variety of variables, such as migration patterns, road quality, conflict, and educational achievement [44, 21, 42, 65].

The effectiveness of CNNs in using satellite imagery to predict sociodemographic data lies in the fact that human living conditions are reflected on the Earth's surface [12]. For instance, the material of a household's roof can indicate it's inhabitant's income level, while the presence of a playground at a school might be indicative of more resources, and thereby better student outcomes [60]. These methods have demonstrated potential not only in academic research but also in practical applications in sectors such as housing, finance, and security [3, 65].

Despite the potential of satellite imagery in advancing various technologies, there has been a notable shortfall in harnessing its spatial characteristics effectively [58]. Technologies often overlook factors like spatial heterogeneity and spatial autocorrelation and the exploration of spatially-aware deep learning models, which integrate coordinate information into their designs, remains minimal [77]. Additionally, there's been a lack of concerted effort to benchmark progress—a common practice in other domains [10, 39]. The introduction of torchGeo marked a significant step within the satellite imagery and deep learning community by providing baseline accuracies and benchmark datasets [68].

However, these resources predominantly focus on land cover classification, ecological conservation, building segmentation, and similar areas. They do not cover the prediction of sociodemographic data, leaving a significant gap in the field [71]. This absence of foundational benchmarks means that methods often overlap without clear understanding of their relative expected accuracies. Only recently has research begun to explore which types of sociodemographic data can be effectively predicted from space, indicating a nascent stage in this area of study [61].

Another critical aspect that has been underexplored in leveraging satellite imagery is the transferability of models across different geographical regions, particularly the challenge of applying models trained on data-rich regions to data-sparse regions [29]. This gap significantly undermines the potential of satellite imagery analytics, as models optimized for regions with abundant data may underperform or fail entirely when applied to areas where data is scarce or of lower quality [45]. The scarcity of efforts to develop models on diverse, globally representative datasets exacerbates this issue, leading to a stark disparity in model performance across different parts of the world [5]. Consequently, regions most in need of the insights provided by satellite imagery—for instance, remote areas lacking in robust ground-based monitoring infrastructure—remain underserved. This limitation not only restricts the practical utility of satellite imagery in addressing regional challenges but also highlights a pressing need for more adaptable and robust models capable of performing well across a wide range of geographical contexts, especially in data-sparse environments [80, 67].

In addressing the challenge of model transferability and the underutilization of spatial characteristics in satellite imagery, the first chapter of my dissertation will introduce GeoEmbed, a method designed to enhance model adaptability across diverse geographic landscapes. GeoEmbed incorporates visual, contextual, and spatial information from satellite imagery into its feature vectors. This approach not only leverages the inherent image features for visual cues but also integrates spatial data through the inclusion of coordinates, enriching the model's spatial understanding. Furthermore, GeoEmbed utilizes a unique selection mechanism for incorporating candidate neighbor imagery, which aids in capturing contextual relevance by identifying neighboring images that significantly influence wealth prediction. This integration aims to overcome the

limitations of existing models by ensuring more robust and transferable predictions across varied and data-sparse regions, enhancing the model's ability to generalize to new areas without sacrificing accuracy.

My second chapter will explore the development of spatially-conditioned convolutional and fully connected layers within a hypernetwork architecture, aimed at addressing the need for models that can dynamically adapt to new and different regions. This approach allows the model to tailor its internal parameters, specifically the kernels of convolutional networks and the weights of fully connected layers, based on coordinate information input. By enabling the model to adjust its functioning according to the geographical context of the input data, I expect this architecture to significantly improve the model's adaptability and effectiveness across diverse geographic areas.

In my third and final chapter, I will introduce torchSDG, a benchmarking package specifically designed for deep learning applications using satellite imagery to predict sociodemographic data. TorchSDG is intended to be a resource for the research community, offering baseline trained models and access to publicly available Sentinel imagery curated for each SDG subindicator, starting with a pilot limited to SDG4. By focusing on SDG4 as an initial test case, torchSDG not only facilitates the development and evaluation of models aimed at predicting educational attainment and quality but also establishes a framework for expanding to other sociodemographic indicators. This comprehensive benchmarking tool will address the previously identified gap in the field by setting standardized metrics for performance and progress, encouraging the development of models that are both accurate and universally applicable across various regions, including those that are data-sparse. Through torchSDG, my dissertation will contribute to advancing the field by offering a structured approach to model evaluation and encouraging further exploration into the prediction of sociodemographic data from satellite imagery.

My dissertation aims to set forth a comprehensive framework for addressing some of the most pressing challenges in the application of satellite imagery for predicting sociodemographic data across diverse geographical regions. Through the approaches presented in each chapter, including the development of GeoEmbed for enhanced model transferability, the implementation of spatially-conditioned layers within a hypernetwork architecture for adaptable model performance,

and the introduction of torchSDG for benchmarking models on a global scale, my dissertation seeks to drive not only scientific advancement in the field, but provide societal impacts by improving our ability to predict and address complex sociodemographic challenges on a global scale.

Chapter 2

Literature Review

2.1 Deep Learning and Satellite Imagery to Predict Sociodemographic Data

Understanding the distribution of socioeconomic characteristics around the world is an important goal for many communities, with applications including disaster response, commercial development, scheduling social services, or targeting poverty [2, 19, 37]. In the last decade, the remote sensing community has engaged with this problem, developing methods utilizing deep learning models and satellite imagery to predict sociodemographic data being popularized around 2016 [33, 79, 4, 55].

The application of satellite imagery in deep learning models for predicting sociodemographic data gained significant attention around 2016 [33]. Since then, it has become a central theme in an expanding body of literature [16, 8, 27, 31]. Initially, these methods often relied on nighttime light data as a proxy for economic development, combined with traditional satellite imagery sources to enhance model performance [33, 48, 53]. Additionally, studies utilized DHS and LSMS survey data, as well as census data, to predict variables such as LSMS poverty scores, purchasing power, roof materials, and other dimensions of wealth [8].

One of the principal challenges with these methods has been the difficulty in interpreting their results, a hurdle commonly encountered with many deep learning models. For example, Yeh et al. [79] demonstrated the effectiveness of Convolutional Neural Networks (CNNs) in predicting poverty estimates for villages across Africa, underscoring their potential for real-world applications.

Nonetheless, they acknowledged the interpretability challenges that accompany deep learning models.

One existing vein of inquiry is into the external generalizability of these classes of models [38, 6]. While our ability to extend some satellite analysis approaches across domains is well understood, (i.e. NDVI), there is a much more limited body of literature in the context of deep learning [30, 43]. For example, while Babenko et al. [5] illustrated strong performance in a subset of Mexico municipalities, they found a rapid degradation of accuracy in other municipalities, despite their geographic proximity.

Many groups have begun to explore how different network architectures may be valuable in the context of satellite imagery. Tile2Vec, introduced by Jean et al. [34], employed unsupervised representation learning to geospatial data, achieving a 49.6% explanation of poverty variance in Uganda, outperforming transfer learning methods. Perez et al. [56] explored generative adversarial networks (GANs), achieving notable validation accuracy. Ni et al. [52] tested various deep learning architectures for daytime satellite imagery and found that the DenseNet model with a squeeze and excitation module performed best. Zhao et al. [81] utilized a Vgg16 model trained on various data sources to predict household wealth in Bangladesh and Nepal, emphasizing the importance of proximity to urban areas in explaining poverty. Similar techniques extended to Mexico, the Philippines, Thailand, India, and Brazil.

Measurement imprecision in the underlying instruments used to train deep learning models is also relatively poorly understood, with only a small number of authors exploring this topic [31, 31, 41]. Recent research has noted limitations in current deep learning approaches, particularly when facing spatial perturbations in poverty indicators [31]. While CNNs traditionally relied on nighttime lights, Liu et al. [41] confirmed the validity of this proxy by detecting visual patterns related to economic development. Liu et al. trained a VGG-16 model on county-level GDP data in mainland China with high predictive success.

Recent work has emphasized the use of open-source data to reduce costs associated with imagery and model training. Lower-resolution Landsat 7 imagery, supplemented with hyperspectral bands, has shown promise in surpassing previous benchmarks [56]. Additionally, crowd-sourced geospatial information has been effective for real-time poverty mapping in the Philippines [72].

TABLE 2.1: Satellites Commonly Used in Deep Learning Models

Satellite Name	**Resolution Type**	**Temporal Resolution**	**Spatial Resolution**	**Radiometric Resolution**	**Owner**
WorldView-3	High	Days	0.31m	11-bit	DigitalGlobe
GeoEye-1	High	Days	0.41m	11-bit	DigitalGlobe
Landsat 8	Medium	16 days	15m - 100m	12-bit	USGS NASA
Sentinel-2	Medium	5 days	10m - 60m	12-bit	ESA
MODIS	Low	1-2 days	250m - 1km	12-bit	NASA
VIIRS	Low	1 day	375m - 750m	14-bit	NASA NOAA

Other approaches to reduce the costs of imagery acquisition are also being explored, such as Ayush et al's [4] study that proposes a reinforcement learning algorithm to reduce the need for high-resolution imagery, making poverty prediction more accessible, particularly for smaller-budget teams, without compromising accuracy. This algorithm achieved an 80% reduction in the necessary high-resolution imagery while maintaining model precision.

2.1.1 A Brief History of Satellite Imagery Sources

Satellite images are taken by satellites orbiting the Earth, which capture and relay data about the planet's surface and atmosphere [36]. Satellite images span a broad range of resolutions, from high spatial resolution images that can discern individual buildings and vehicles, to low spatial resolution imagery that covers large areas of land in single pixels [59]. Satellite imagery can be used for monitoring and managing natural resources, forecasting weather patterns, planning urban areas or enhancing agricultural practices and assessing the impacts of natural disasters [75]. The temporal resolution of satellite imagery, how frequently images are taken of a location on the earth's surface, can vary, enabling both real-time insights and long-term trend analysis [9]. Radiometric resolution refers to a satellite sensor's ability to distinguish between fine differences in energy (or brightness), enabling the identification of subtle variations in the reflectance or emission of features on the Earth's surface [35]. When combined with methods like deep learning, satellite imagery can offer insights into Earth's processes, human-made structures, and the interplay between them [82].

2.1.2 Data Sources Leveraged in Deep Learning Models

The source of satellite imagery employed in deep learning models varies depending on task. High resolution data such as that from DigitalGlobe [15], Planet Scope [57] and Google Basemaps [22] is used often in papers that predict more fine-grained data such as crime, housing prices or crop yield [64]. In contrast, low resolution imagery is used to make more broad scale predictions such as large-scale trends migration and conflict [61, 51].

WorldView satellites (like WorldView-3 and WorldView-4) and GeoEye [47] are known for their high-resolution imagery capabilities, allowing users to capture very fine details from space. These satellites are often used for precise mapping, surveillance, and various applications where detailed imagery is crucial [54].

The Landsat program began in the 1970's and is one of the longest-running efforts to collect moderate-resolution publicly available satellite imagery of Earth. Jointly managed by NASA and the USGS (U.S. Geological Survey), Landsat satellites (from Landsat 1 to Landsat 9) have provided a continuous record of Earth's surface, aiding in studies of deforestation, urbanization, and climate change, among others [49].

Part of the Copernicus Programme operated by the European Space Agency (ESA), the Sentinel satellites are a family of satellites designed to deliver comprehensive and medium-to-high-resolution data about our planet. There are various Sentinel missions (e.g., Sentinel-1, Sentinel-2), each equipped with different sensors for applications ranging from land and ocean monitoring to atmospheric and climate studies [17].

Aboard the Terra and Aqua satellites, MODIS is a key instrument that captures data in a wide range of spectral bands. It provides low-resolution images, allowing for large-scale observations of the entire Earth every 1 to 2 days. This frequent revisiting capability makes MODIS important for tracking dynamic phenomena like wildfires, cloud formations, and phytoplankton blooms [50].

Found on the Suomi-NPP and NOAA-20 satellites, VIIRS collects visible and infrared imagery and radiometric measurements of the land, atmosphere, cryosphere, and oceans. It's known for its nighttime lights dataset, which captures artificial light on Earth's surface, offering insights into human activity and development patterns [50].

2.2 Spatially-Aware Convolutional Neural Networks

Here, I define Spatially-Aware convolutional neural networks as those which incorporate spatial information (i.e., spatial networks or geographic place identifiers) into architectures which take advantage of such information [24, 25]. Two distinct approaches to the integration of spatial information into deep learning models have emerged from the recent literature: addressing spatial variability [24, 25] and employing graph-based methodologies for modeling spatial relationships and interactions [7, 76].

The work on spatial variability aware deep neural networks (SVANN) highlights a shift from models that do utilize any geographic information, or one-size-fits-all (OSFA) models, towards models that consider the unique spatial characteristics of different geographic areas [24, 25]. Studies published by Gupta et al. in 2020 and 2021 demonstrate that SVANN models, which are tailored to specific regions, significantly outperform OSFA models in tasks such as urban garden mapping and wetland mapping.

An alternative approach, spatially-aware graph neural networks (SPAGNN and SHGNN) focuses on capturing the interactions and heterophily within spatial data [7, 76]. These models excel in urban applications by formulating cities as urban graphs, where nodes represent urban objects, and edges encapsulate the relationships between these objects. Specifically, SHGNN addresses the unique spatial heterophily property observed in urban graphs, demonstrating that handling the diversity of spatial distances between nodes can significantly improve the model's effectiveness in urban applications [76]. Similarly, SPAGNN's approach to relational behavior forecasting, using message passing to iteratively update actor states, showcases the ability of graph-based models to incorporate spatial transformations and probabilistically model uncertainty in scenarios like autonomous driving [7].

Both the SVANN and graph-based methodologies present compelling advancements in making deep neural networks more spatially aware. While SVANN models focus on adapting to spatial variability across different geographic locations, graph-based approaches like SPAGNN and SHGNN emphasize modeling the web of relationships and interactions within spatial data.

2.3 Application Area: United Nations Sustainable Development Goals

In my research, I will propose leveraging satellite imagery to capture variables of interest under the Sustainable Development Goals (SDGs) definitions promoted by the United Nations [73]. The Sustainable Development Goals were established as a global framework aimed at addressing a broad spectrum of sustainability issues ranging from poverty and inequality to climate change and environmental degradation. This comprehensive framework consists of 17 goals, each with a set of targets and indicators designed to measure progress toward achieving these goals.

2.3.0.1 Structure of the SDGs

The Sustainable Development Goals (SDGs) are a framework within academic and non-governmental organization (NGO) sectors, guiding research, policy analysis, and intervention strategies aimed at addressing global challenges. Rooted in the principles of economic growth, social inclusion, and environmental protection, the SDGs offer a comprehensive agenda for promoting sustainable development worldwide by 2030 [23].

In academia, the SDGs serve as a reference point for interdisciplinary research, shaping studies across environmental science, economics, public health, and social sciences [40]. Researchers utilize the SDGs to frame investigations, assess the impact of sustainable development practices, and contribute to a broader understanding of how global challenges can be addressed effectively [11]. The goals' structure—encompassing 169 targets and 232 unique indicators—provide a blueprint for measuring progress and outcomes in sustainable development [23].

Meanwhile, NGOs leverage the SDGs to design and implement development projects, monitor progress, and evaluate the impact of their interventions [73]. The goals' allow organizations to align their missions with global priorities, fostering collaboration and coordination among various stakeholders [63].

2.4 Benchmark Datasets for Deep Learning with Satellite Imagery

In computer vision, benchmark datasets play a pivotal role in developing and evaluating algorithms for image analysis [26]. Table 2.2 highlights notable benchmark datasets utilizing satellite imagery, highlighting their unique contributions to the domain.

Name	**Task**	**Type of Imagery**	**Description**	**Citation**
BigEarthNet	Land Cover Classification, Monitoring	Multispectral Satellite Imagery	A large-scale benchmark archive, consisting of Sentinel-1 and Sentinel-2 image patches, designed for land cover classification and monitoring.	[70]
Merced Land Use Dataset	Land Use Classification	High-Resolution Satellite Imagery	Comprises 2100 land-use images taken from the USGS National Map, suitable for tasks like pattern recognition and classification.	[78]
SpaceNet	Object Detection, Segmentation	High-Resolution Satellite Imagery	Focuses on geospatial object detection, offering a series of challenges with datasets for building and road detection, and much more.	[74]

SEN12MS	Multitask Learning (Classification, Segmentation)	Multispectral and SAR Satellite Imagery	A dataset that pairs Sentinel-1 SAR and Sentinel-2 multispectral imagery for 12 tasks, including land cover classification and scene understanding.	[66]
EuroSAT	Land Use and Land Cover Classification	Multispectral Satellite Imagery	Consists of Sentinel-2 satellite images covering 13 spectral bands and is designed for land use and land cover classification.	[28]
DeepGlobe Satellite Challenge	Road Extraction, Building Detection, Land Cover Classification	High-Resolution Satellite Imagery	A challenge that features several satellite imagery tasks for road extraction, building detection, and land cover classification for mapping purposes.	[13]
So2Sat LCZ42	Urban Typology Classification (Local Climate Zones)	Multispectral and SAR Satellite Imagery	Combines Sentinel-1 SAR and Sentinel-2 multispectral imagery for the classification of 42 local climate zones, relevant for urban planning and development.	[83]

As shown in table 2.2, there is a wide range of satellite image repositories available for benchmarking purposes. However, when it comes to estimating socioeconomic variables, these repositories present significant limitations. For instance, of all the benchmarks currently known, only one, DeepGlobe, could potentially be employed to determine variables pertinent to socioeconomic status, yet its applicability is likely confined primarily to road infrastructure. In

Chapter 3 of my dissertation, I will tackle this challenge by introducing a new set of benchmarks specific to socioeconomic prediction in torchSDG.

Chapter 3

Quantitative Analysis

3.1 Introduction to Quantitative Analysis

The estimation of socioeconomic features in data-sparse regions, particularly those affected by conflict or lacking historical survey data, is an important goal for both academic research and policy-making [32]. Recent advancements have focused on addressing these challenges by leveraging deep learning techniques in conjunction with contemporary and historical satellite imagery [62, 61]. These methods have demonstrated substantial potential in providing estimates for various factors like income, education, health, and other socioeconomic indicators that would otherwise be challenging, or even impossible, to derive [60, 5, 4].

One of the primary challenges in utilizing satellite imagery for estimating socioeconomic information lies in the variability of features across different geographic locations. For example, the density and type of housing, infrastructure development, and land use can significantly differ between urban, suburban, and rural areas, each providing different indicators of economic status. In urban areas, densely packed high-rise buildings may indicate a higher cost of living and potentially higher incomes, whereas in rural areas, the extent of cultivated land or the presence of modern agricultural facilities might suggest wealth levels. This problem is exacerbated when trying to measure socioeconomic factors across different countries.

This lack of stationarity in relevant features poses a significant obstacle to generalizing models trained on a single region to out-of-sample locales [5]. As

models that leverage satellite information predominantly focus on specific regions (i.e., countries), the non-uniformity of predictive features hinders the transferability of these models to new geographic areas [33].

To overcome this challenge, we propose GeoEmbed. GeoEmbed integrates a pre-trained ResNet18 architecture for initial feature extraction from satellite images. Recognizing the importance of geographic proximity in determining wealth similarity, we then introduce a proximity-based similarity selection mechanism. This mechanism evaluates neighboring candidate images which are geographically close to the estimation target image and selects the top three candidates that share the highest level of similarity in features. The information in these images are then encoded using a two-layer Multi-Layer Perceptron (MLP), which integrates the extracted features with the geographic coordinates of the source image to enhance the prediction accuracy. This encoded vector ultimately includes information on the neighboring images, their geographic location, and contextual information about the latitude and longitude. This information is then concatenated with an image of the target location (i.e., where we want to predict income or related information), providing the model with a more holistic view of the context of a given target.

This section is structured as follows. In section 3.2, we review related work to situate our approach within the current landscape of wealth estimation using satellite imagery and deep learning. Section 3.4 and 3.5 detail the methodology of the proposed GeoEmbed method and the details of our experiment implementation. Finally, Sections 3.6 and 3.7 present our experimental results, discuss their implications, and conclude with the potential impacts of our work on socioeconomic analysis and policy-making.

3.2 Background Literature

3.2.1 Socioeconomic Indicator Prediction using Satellite Imagery and Deep Learning

The fusion of deep learning with satellite imagery has become a powerful tool for predicting sociodemographic data, a breakthrough that gained significant momentum around 2016. This method, utilizing satellite imagery alongside

proxies such as nighttime light data, has shown promising results in estimating wealth indices and other socioeconomic variables [32, 48]. Despite its success, challenges such as the interpretability of deep learning models and their generalizability across diverse geographic regions remain [79, 5].

Model generalizability in the context of satellite imagery and deep learning presents a complex challenge that intersects with the spatial and contextual diversity of the globe. As models excel in specific geographic locales, extending these successes to new regions requires nuanced understanding and adaptation of the models to local conditions. This challenge is exemplified by studies where model performance varied significantly across different settings, despite geographic and socio-economic similarities [5, 38].

Innovations in network architectures and the application of various machine learning techniques have enhanced the predictive capabilities of these models. Techniques like unsupervised representation learning and generative adversarial networks (GANs) have expanded the analytical potential of satellite imagery in socioeconomic analysis [34, 56]. Furthermore, addressing the precision of training data and exploring proxies have been critical in refining model accuracy and reliability [31, 41].

Efforts to reduce the financial barriers associated with high-resolution satellite imagery have led to the exploration of open-source data and more cost-effective algorithmic approaches. These strategies aim to make socioeconomic predictions more accessible and practical for broader applications, including in resource-limited environments [56, 4]. The continued innovation and exploration in this field underscore the evolving landscape of remote sensing and deep learning, driving forward the capabilities of sociodemographic predictions through satellite imagery.

3.2.2 Spatially-Aware Deep Learning Models

Spatially-Aware Convolutional Neural Networks (CNNs) represent integrate geographical information into deep learning architectures. This integration manifests in two primary methodologies: the adaptation to spatial variability and the employment of graph-based approaches to model spatial relationships [24,

25, 7, 76]. The former, exemplified by spatial variability aware deep neural networks (SVANN), marks a departure from one-size-fits-all models, tailoring neural networks to the unique spatial characteristics of distinct geographic areas. Studies by Gupta et al. have shown that such region-specific models yield superior performance in applications like urban garden and wetland mapping [24, 25].

Concurrently, spatially-aware graph neural networks (e.g., SPAGNN and SHGNN) focus on encapsulating the complex web of interactions within spatial datasets. By treating cities as urban graphs where nodes represent urban elements and edges define their interconnections, these models adeptly capture the dynamic interactions and spatial heterogeneity inherent in urban environments. SHGNN, in particular, addresses spatial heterophily—varied spatial distances between nodes—enhancing model efficacy in urban planning and analysis [76]. Similarly, SPAGNN leverages message passing techniques to refine relational forecasting, a crucial capability in scenarios like autonomous driving [7].

These approaches underline the evolving landscape of spatially-aware deep learning, where SVANN models optimize for spatial variability across geographies, and graph-based methodologies, through the modeling of spatial relationships, offer detailed insights into spatial interactions.

3.3 Data

3.3.1 USAID Demographic and Health Survey

We use data collected from the United States Agency for International Development (USAID) Demographic and Health Surveys (DHS) [1], which provides geocoded information about individuals socioeconomic status, such as wealth, household characteristics, educational attainment, and critical health indicators including infant and child mortality, fertility rates, family planning utilization, maternal health. The survey data is publicly accessible, and the indicators are collected so as to be comparable across countries and over time.

In this study, we leverage the DHS's Wealth Index, which is a composite measure designed to capture the relative socioeconomic status of households. It is calculated using a principal component analysis (PCA) technique on data for

assets owned by the household and housing characteristics, such as the availability of electricity, water supply, flooring material, and ownership of various consumer goods (e.g., television, bicycle). The first principal component derived from PCA is used to generate the wealth scores for each household. These scores are then standardized and divided into quintiles, representing different wealth levels from the poorest to the wealthiest households. To improve comprability across geographies, we mean-normalize these scores on a per country basis following equation 3.1:

$$x_{\text{norm}} = \frac{x - \mu}{\max(x) - \min(x)} \tag{3.1}$$

where

- x_{norm}: The mean-normalized value of the data point x.
- x: An individual data point within the dataset.
- μ: The mean of all the data points in the dataset.
- $\max(x)$: The maximum value in the dataset.
- $\min(x)$: The minimum value in the dataset.

The presented analysis includes information from 9 countries, with 7 serving as the source dataset (153,875 data points) and 2 withheld for validation (19,803): Honduras and Guatemala. The countries included in training are shown in 3.1.

Figure 3.2 illustrates the differences between the wealth distribution labels in our source and target datasets. The range of values is similar across both partitions.

Figure 3.3 shows a more detailed breakdown of the statistics within each of our target countries.

We aggregated the Standard DHS survey data to the household level within each DHS cluster. The DHS program offsets coordinates for rural clusters by up to 5km and for urban clusters by up to 2km. Following established methodologies [32], we sampled 10 random points within a 5km buffer around each rural cluster's centroid and 10 random points within a 2km buffer around each urban cluster's centroid. These points served as the locations from which Planet

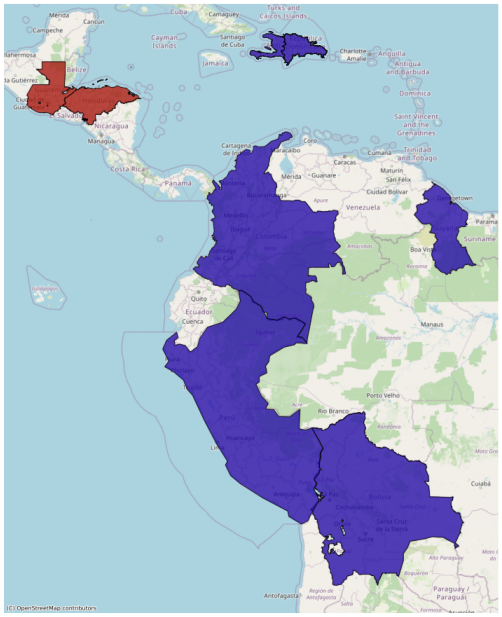

FIGURE 3.1: Dataset coverage map. 7 countries (in purple) represent our source countries and 2 countries (in red) represent our target countries.

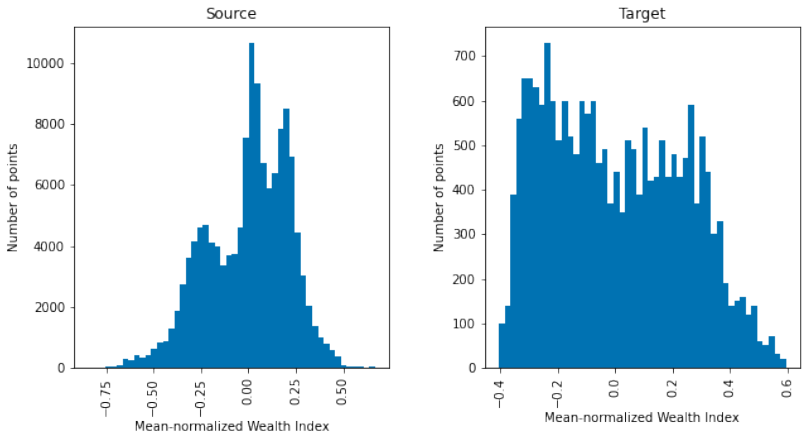

FIGURE 3.2: Comparative histograms of the mean-normalized Wealth Index between source data (aggregated from multiple countries) and target data (aggregated from selected countries).

Basemaps imagery was extracted for our final dataset, as outlined in section 3.3.2.

3.3.2 Planet Basemaps Satellite Imagery

Planet Basemaps [57], produced by the commercial satellite imagery provider Planet, leverage a network of Earth observation satellites equipped with a variety of sensors, including the DOVE, DOVE+, and DOVEULTRAPLUS series. These satellites generate high-resolution imagery that forms the backbone of Planet's global composite basemaps. The imagery has a spatial resolution of 3 meters and is available in both monthly and quarterly composites.

We clipped the 2023 third quarter global composite Planet Basemaps image to a .08km buffer around each point in our dataset as shown in Figure 3.4, resulting in 153,875 total imagery tiles each labeled with the associated DHS mean-normalized Wealth Index value.

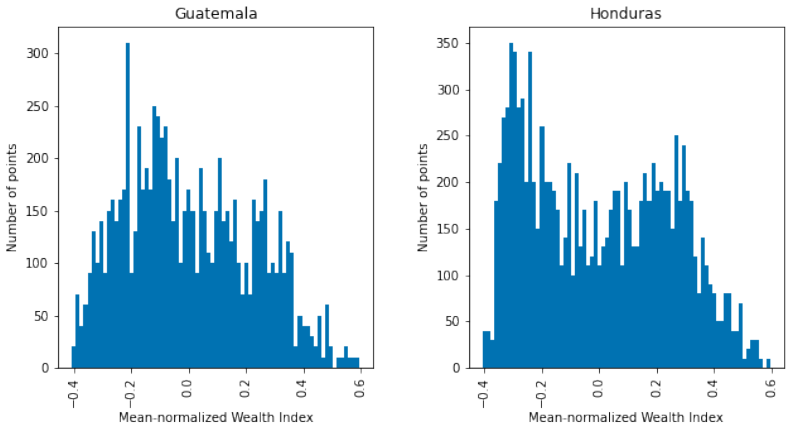

FIGURE 3.3: Histograms displaying the distributions of the mean-normalized Wealth Index in the four target countries: Honduras, Rwanda, Tajikistan, and Timor-Leste. Each histogram shows the frequency of data points at different Wealth Index levels

3.4 Methodology

3.4.1 Framework Overview

Our approach uses proximity-driven embedding of visual, spatial and contextual information to train a model that generalises to data-sparse regions following the steps outlined below.

1. **Feature Extraction** (Section 3.4.2: In this stage, we pair each source image (i.e., an image of a household surveyed by the DHS) with eight candidate images from proximate areas. For each image, we then use a DeepAll ResNet-18 model to create a 512-dimensional vector for each source and candidate image, with weights frozen before training.

2. **Contextual embedding** (Section 3.4.3): For proximity-based embedding, we compare the target and candidate images' vectors and select the top three based on similarity. These selected vectors are then aggregated into a set referred to as the 'nearest neighbor set', which encapsulates the closest matches based on cosine similarity.

FIGURE 3.4: Examples of imagery clips

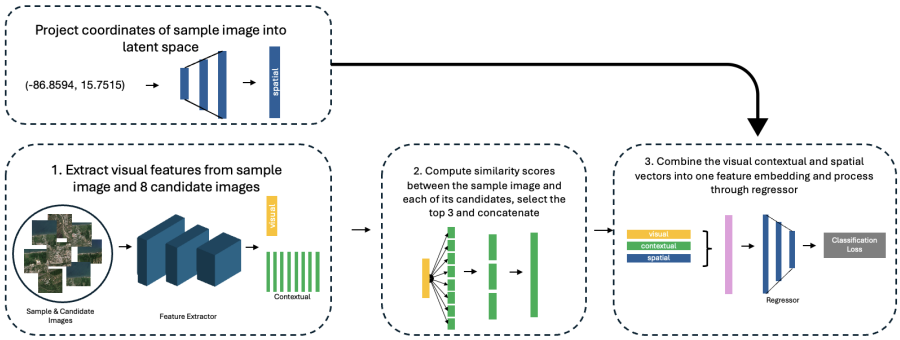

FIGURE 3.5: The proposed GeoEmbed framework

3. **Spatial Embedding** (Section 3.4.4): Process geographic coordinates of each source image through a two-layer MLP to produce a 128-dimensional vector, allowing the model to capture spatial heterogeneity.

4. **Fully Connected Layers for Final Prediction** (Section 3.4.5): We merge the embeddings from coordinates, proximity-based analysis, and source images into a single 1-dimensional vector, which is then processed by a series of affine layers to produce the final prediction.

3.4.2 Feature Extraction

We use a pretrained DeepAll model, as described in section 3.5.2.1, as our initial feature extractor. The fully connected Layer of the DeepAll model is replaced with an Identity layer, resulting in a 512-dimensional feature vector output for each image.

3.4.2.1 Proposed Candidates

GeoEmbed proposes eight candidate images for each input image to use for contextual feature embedding. The candidate slection method is outlined below.

3.4.2.2 Cross-Border Candidate Selection

Our candidate selection method expands the search for candidate imagery beyond the confines of geographic and national borders, aiming to introduce a rich set of features into the selection pool. For each source image, instead of limiting the search to the nearest neighbors within the same country, this method identifies the eight geographically closest images located outside of the source image's country of origin. This cross-border approach is designed to capture a broader spectrum of environmental, architectural, and cultural diversity.

This method recognizes that similar or relevant features can exist across national borders, especially in regions where ecological zones or cultural landscapes span multiple countries. By including imagery from neighboring countries or regions, this approach aims to enrich the dataset with a more varied set of characteristics, potentially enhancing the robustness and generalizability of analyses or models developed using this imagery.

3.4.2.3 Source and Candidate Feature Extraction

We pair each source image with its candidates based on the method described in section 3.4.2.1. Each image in a mini-batch, along with its associated candidate images, is then processed through the DeepAll model, which serves as a feature extractor. This process generates a 512-dimensional feature vector for each image, capturing its high-level features. All of the weights of this feature extractor are frozen prior to training.

3.4.3 Contextual Embedding

Post feature extraction, the feature vectors of candidate (i.e., all proximate images for which features were extracted) and the target image are projected into latent space. This is achieved by passing the feature vectors individually through separate linear layers, each reducing the input vectors from 512 to 256 dimensions. We compute the cosine similarity between the feature vector of the target image and each of the individual candidate vectors, selecting the top three candidates based on similarity scores. This has the effect of identifying, of the proposed candidates, the three which share the most similar image characteristics.

We employ a softmax weighting strategy to calculate how similar each candidate image is to the target image, enabling a weighted analysis of their importance. The softmax layer to converts their raw similarity scores into percentages, transforming them into a distribution that highlights the relative similarity of each candidate to the target. Subsequently, we adjust the vectors of these top three candidates by their respective softmax percentages. This weighted adjustment ensures that the influence of each candidate on the final vector is proportional to its similarity to the target image. By applying their corresponding percentages, we refine the embedding process to emphasize the contributions of the most relevant images.

After calculating the softmax similarities, we select the top 3 most similar candidates, each being represented by their 256-dimensional feature vector. The feature vectors are then multiplied by their respective similarity scores. Finally, we concatenate the adjusted vectors of the three candidates into a single 768-dimensional vector. This concatenated vector is a comprehensive representation that amalgamates the most relevant candidates. This vector is then fed through a linear layer, which reduced it down to a 256-dimensional vector.

3.4.4 Spatial Embedding

For every source image, the geographic coordinates of the household it represents are input into a two-layer Multi-Layer Perceptron (MLP). The MLP consists of two hidden layers with dimensions 64 and 128, respectively, and ReLU

activation functions are leveraged after each hidden layer to introduce non-linearity and aid in learning spatial relationships. This process outputs a 128-dimensional feature vector that encapsulates the spatial context of the source image.

By integrating geographic coordinate information directly into our model, we aim to enhance the model's sensitivity to spatial variance across different geospatial domains, thereby improving the accuracy and generalizability of our wealth estimation predictions.

3.4.5 Fully Connected Layers for Final Prediction

The final prediction stage involves concatenating the feature vectors of the top three similar candidate images with the source image's feature vector and the vector representing geographical coordinates. The combined feature vector serves as a comprehensive representation that includes visual, spatial, and contextual information, which is then used for the final wealth prediction through the primary two-layer MLP. This MLP consists of two linear layers of sizes 128 and 32, and outputs the final wealth prediction for the source image.

3.5 Experiments

In this section, we begin by outlining our experimental framework in Section 3.5.1, detailing our dataset of satellite imagery and the specifics of our model selection strategies.

3.5.1 Implementation Details and Model Selection

Our method is implemented via PyTorch's Distriubted Data Parallel. We utilize a standard ResNet18 architecture, shown in Figure 3.7, pretrained on ImageNet. The network is composed of four main blocks, each containing convolutional layers (denoted by 'conv') with varying kernel sizes and filters, followed by batch normalization and ReLU activation. The black arrows represent the standard flow of the convolutional operations, while the red dashed arrows signify

skip connections, a hallmark of ResNet architecture that helps alleviate the vanishing gradient problem by allowing the direct flow of gradients. Skip connections essentially enable the network to learn identity mappings, ensuring that deeper layers can perform at least as well as shallower ones. The final output is obtained after an average pooling layer and a fully connected layer, which together generate the classification predictions.

We utilize an L1 loss function with 2 fold validations for each model, with 25 epochs per fold. Each image has 3 RGB channels and is randomly clipped to an input size of $224 \times 224 \times 3$. The images are normalized to the global mean and distribution during input transformations.

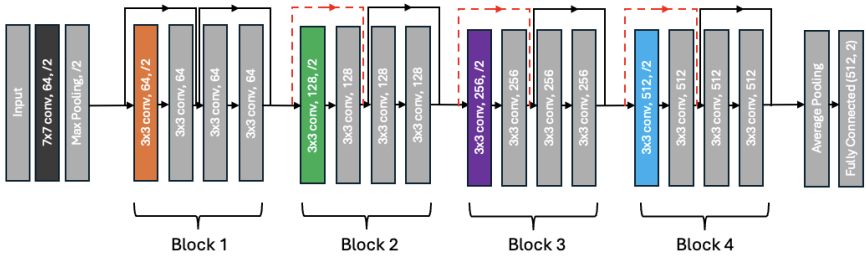

FIGURE 3.6: ResNet18 architecture

Model's were contrasted in their ability to predict accurate wealth values in two target countries, which are withheld during training. The performance of the methods was evaluated using two metrics: the R2 score, which measures the proportion of variance in the dependent variable predictable from the independent variable(s), and the Mean Absolute Error (MAE), which quantifies the average magnitude of errors in a set of predictions, without considering their direction. The model with the highest R2 was selected as the most performant.

3.5.2 Baseline & SOTA Models for Comparison

Using the training and validation datasets described in section 3.3, we test a series of models in terms of their performance when trained on a series of (153,87 images (from 7 countries), and then applied to 19,803 images from 2 out-of-domain countries. This section describes our model implementations in each case.

3.5.2.1 Baseline: DeepAll

We train a DeepAll model by aggregating data from all our source countries, employing a ResNet18 architecture as shown in Figure 3.7. This model serves as our primary benchmark for comparison.

3.5.2.2 Domain generalization using a mixture of multiple latent domains (DGMMLD)

DGMMLD [46] posits that training data, collected under varied conditions, can be segmented into numerous latent domains through unsupervised learning, using clustering or latent variable models to deduce the domain structure. It then models the data as a mixture model, where each mixture component corresponds to a latent domain, with data points associated with domains via learned weights. For each domain, domain-specific feature representations are learned through specialized layers, while simultaneously learning a shared feature representation to ensure generalization across unseen domains. This is facilitated by adversarial training methods that obfuscate domain distinctions to the domain classifier, promoting domain-invariant feature extraction. The training objective encompasses both domain-specific and domain-invariant loss terms, supplemented by regularization to align the feature distributions across domains, utilizing metrics like Maximum Mean Discrepancy and fostering consistent predictions across domains.

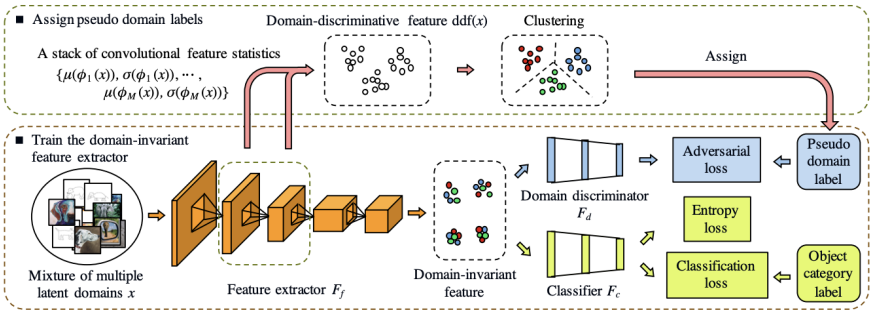

FIGURE 3.7: DGMMLD architecture

3.5.2.3 Spatially-Aware Network

As a baseline method of comparison, we constructed a Spatially-Aware (SAN) network. This network utilizes a ResNet18 architecture pretrained on ImageNet and a 2-layer MLP that projects the coordinates of every image into latent space using 64 and 128 element linear layers. The 128-element output of the coordinate projector and the 512 element input to the ResNet18 fully connected layer are concatenated into a 640 element vector that is then fed through the fully connected layer for a final wealth prediction. We test GeoEmbed method against SAN to ensure that the introduction of the coordinates to our feature extractor was not sufficient to attain the boost in accuracy that GeoEmbed provides, and that the combination of visual, spatial and contextual information is what provides the improvement in accuracy.

3.6 Results

TABLE 3.1: Comparison of Models

Model	**r2**		**MAE**	
	Source	**Target**	**Source**	**Target**
DeepAll	0.505	0.26	0.108	0.17
GeoEmbed	0.289	0.283	0.138	0.169
DGMMLD	0.439	0.363	0.122	0.156
SAN	0.461	0.261	0.113	0.172

3.6.1 Evaluation

DGMMLD achieved the highest R2 score of .363 across the target countries, outperforming DeepAll (.26), GeoEmbed (0.2869) and SAN (.261), indicating that DGMMLD was more effective in capturing the variance in wealth prediction across different geographic locations.

Consistent with the R2 score, DGMMLD also led in terms of MAE with an average error of 0.156, compared to DeepAll (0.17), GeoEmbed (0.169) and SAN (0.172)

For each target country, the R2 score and MAE were computed to assess the model's predictive accuracy, as shown in Figure 3.2.

TABLE 3.2: Comparison of R2 Scores Across Target Countries

Model	**Honduars**	**Guatemala**
DeepAll	0.275	0.234
GeoEmbed	0.375	0.132
DGMMLD	0.39	0.323
SP-Aware	0.28	0.23

The per-country results demonstrate varied performance in predictive accuracy highest r2 score among the compared models, with values of 0.389789 for Honduras and 0.323 for Guatemala, indicating the strongest predictive ability across both categories.

Conversely, the GeoEmbed model, while performing significantly better in Honduras with an r2 score of 0.375, shows a notable decrease in performance in Guatemala, with the lowest score of 0.132 among the models evaluated. This stark contrast suggests that the GeoEmbed model's effectiveness is more pronounced in Honduras but may not generalize as well across the imagery and geographic features in Guatemala.

The DeepAll and SP-Aware models show more consistent, yet moderate performance across both categories. DeepAll achieves an r2 of 0.274794 for Honduras and 0.234 for Guatemala, while SAN records scores of 0.28 and 0.23 for Honduras and Guatemala, respectively. These results indicate a level of reliability across different conditions, although they do not reach the predictive accuracy of the DGMMLD model.

3.7 Discussion

In the comparative analysis of predictive models, the DGMMLD model demonstrates superior performance, particularly evident in its application to the Target dataset. It achieves an r2 score of 0.363 and a MAE of 0.156. This level of accuracy positions the DGMMLD model as a primary benchmark for future iterations of GeoEmbed.

While GeoEmbed (my proposed algorithm) is underperforming relative to DGMMLD, it outperforms both DeepAll and SAN. Further, GeoEmbed is the only model that exhibits consistency across both the Source and Target datasets, with r2 scores of 0.289 and 0.283, respectively. This consistent performance is indicative of the model's robustness and its general applicability across varied datasets. Furthermore, the GeoEmbed model maintains competitive MAE values of 0.138 for the Source and 0.169 for the Target datasets.

SAN demonstrates a solid performance with an r2 score of 0.461 for the Source dataset and 0.261 for the Target dataset, alongside MAE values of 0.113 and 0.172, respectively. Compared to DeepAll's performance however, with r2 scores of 0.505 and 0.26 for the Source and Target datasets, respectively, and MAE scores of 0.108 and 0.17, the DeepAll model demonstrate a greater proficiency in capturing the nuances of the Source dataset. The only difference between the two models was the inclusion of coordinates, which indicates that making the model spatially-aware was actually detrimental to performance among the source dataset.

3.7.1 Limitations

There are a number of limitations to the GeoEmbed approach, may of which I will seek to test or overcome as a part of future iterations of my dissertation.

1. **Geospatial Data Dependency:** GeoEmbed's reliance on geospatial data for its functioning underscores a significant limitation—its applicability is confined to scenarios where precise coordinate information is available. This requirement restricts its utility to datasets that include accurate geographic coordinates, sidelining its potential application in contexts where such data is sparse, obfuscated, or entirely absent. The intrinsic value of GeoEmbed lies in leveraging spatial relationships, which, while powerful, narrows its use cases to well-mapped domains, leaving a gap in contexts where geospatial data integrity is compromised or non-existent.

2. **Geographic Imprecision Effects**: The impact of geographic imprecision, such as the DHS offset issues, introduces a layer of complexity in the model's predictive accuracy. These offsets, intended to protect survey respondent

privacy, can distort the true spatial relationships and contextual relevance among data points. The extent to which this imprecision affects model performance remains an area ripe for exploration. It raises concerns about the robustness of GeoEmbed and similar models in accurately interpreting and integrating spatial data, especially in applications where precise geographic information is crucial for understanding and predicting socioeconomic indicators.

3. **Neighbor Selection Limitations:** The efficacy of GeoEmbed is inherently tied to the density and availability of neighboring observations. In regions with sparse data, the model's capacity to leverage contextual information from proximate observations is significantly diminished. This limitation mirrors challenges faced in traditional imputation methods, where the absence of nearby data points can lead to decreased performance and increased uncertainty in predictions. The model's dependency on a rich dataset for neighbor selection highlights a vulnerability in data-sparse environments, potentially undermining its utility in the very contexts it aims to serve.

4. **Spatial Autocorrelation Requirement**: GeoEmbed's performance advantage is contingent on the presence of spatial autocorrelation within the dataset—where similar values cluster geographically. In scenarios lacking this spatial correlation, particularly where the geographic distribution of features does not align with the distribution of wealth or other socioeconomic variables, GeoEmbed's methodology may not offer substantial benefits over non-spatially aware models. This limitation delineates a boundary condition for the model's applicability, suggesting that its advantages are context-dependent, rooted in the nature of the underlying spatial data patterns.

5. **Computational Cost Concerns** The process of passing neighboring images through the feature extractor amplifies the computational demands of the GeoEmbed framework, marking a significant limitation in terms of efficiency and scalability. This computational overhead, stemming from the

additional processing required for each neighboring image, poses challenges for deployment in resource-constrained environments or in real-time applications where rapid processing is essential. The increased computational load necessitates a careful consideration of the trade-offs between the model's spatial insights and the practical constraints of computational resources.

Chapter 4

Proposed Structure for Dissertation

4.1 Chapter 1: Proximity-driven embedding of Visual, Spatial and Contextual features for the estimation of wealth in data-sparse regions

4.1.1 Motivation and Synopsis

Section 3 details the groundwork and progress I have made towards my first proposed chapter: Proximity-driven embedding of Visual, Spatial and Contextual features for the estimation of wealth in data-sparse regions.

In my quantitative section, I proposed a framework, GeoEmbed, for estimating the distribution of household wealth in data-sparse regions using satellite imagery and deep learning. For each household, the methodology employs a pre-trained ResNet18 architecture for initial visual feature extraction, both for the target of the estimation as well as it's geographic neighbors. On the basis of an image similarity function, a subset of these geographic neighbors are then selected for use as inputs into the network to aid in the estimation of wealth for the target household. This approach enables the network to take advantage of spatially-explicit contextual factors - for example, a household that is proximate to other wealthy households may be more likely to also be wealthy.

Experimental results demonstrate the potential of this approach within central and south America, however the method is still under performing DG-MMLD, as shown in Table 4.1.

In the following section 4.1.2, I will outline my plan for improvement and ablation studies based on the current limitations of GeoEmbed, for the purpose

TABLE 4.1: Comparison of Models

Model	**r2**		**MAE**	
	Source	**Target**	**Source**	**Target**
DeepAll	0.505	0.26	0.108	0.17
GeoEmbed	0.289	0.283	0.138	0.169
DGMMLD	0.439	0.363	0.122	0.156
SpAware	0.461	0.261	0.113	0.172

of inclusion into my dissertation.

4.1.2 Limitations and Plan for Improvement

4.1.2.1 Proposed Candidate Images

In its current implementation, GeoEmbed proposes 8 candidate images for potential use as "neighboring" cases, and then selects the three with the most similarities (via a cosine distance method). The process of creating the dataset, however, involved selecting 10 points within a confined buffer zone, which results in the chosen candidate images being notably similar in terms of features - i.e., they are frequently drawn from the same country (see Figure 4.1). We would ultimately like to be able to apply this technique in contexts where no data exists within a given country - i.e., using information from Mexico to improve estimates within Honduras - but this will necessarily result in less image similarity (see Figure 4.2).

To overcome this limitation, I propose to test two other candidate selection processes. For the first test, I intend to randomly sample candidates from the nearest 32 neighbors, either within or outside of the country within which the imagery was taken. I expect this adjustment to introduce a broader spectrum of candidate features, thereby enriching the dataset and enabling the model to base its decisions on a more diverse and representative feature set.

The second approach I will investigate eliminates reliance on spatial dependencies and instead selects candidate images based on their style statistics alone. In domain adaptation literature, style statistics are often deemed superior in representing the domains of imagery compared to features derived from fully connected layers. This perspective suggests that style statistics capture essential

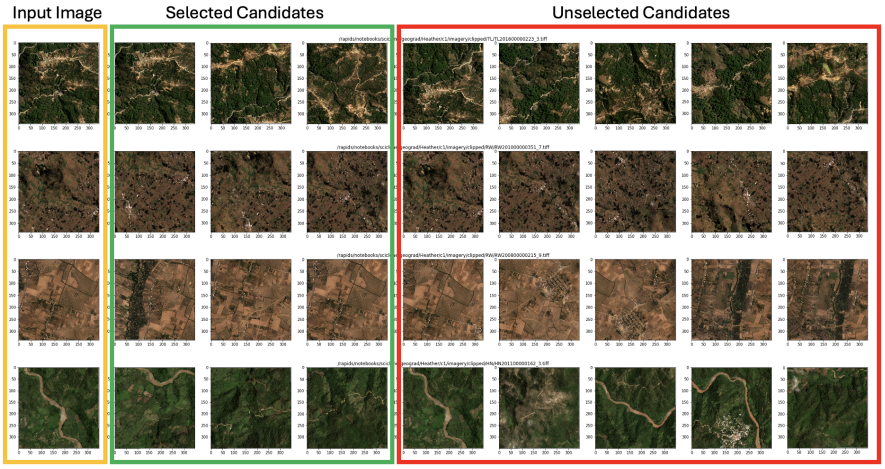

FIGURE 4.1: Each row in both this figure and figure 4.1 is an input image and it's 8 proposed candidates. The input image in both figures is circled in orange, the candidate images the model selected are in green, and the remaining candidates are in red. As demonstrated, the candidates (anything in green or red) are all very similar in their content, representing very little feature diversity.

visual elements that more accurately reflect the underlying domain characteristics of images.

To extract style statistics from every image, each image will be processed through a pretrained DeepAll model to obtain the feature maps from the ResNet18's block 4. This tensor will then be reshaped and flattened along its spatial dimensions to calculate a Gram matrix by multiplying its tensor with its transpose. This process computes the correlation between different feature maps, capturing the image's style [20].

Following this style computation, the candidates of every image are those with the smallest euclidean distances between their style matrices, regardless of geographic distance. This approach is based on the theory that even across disparate geographical locations, regions with similar styles may share underlying socioeconomic and cultural characteristics. For instance, urban areas worldwide might exhibit similar patterns of development, infrastructure, and housing that correlate with certain economic levels. Similarly, rural areas with comparable agricultural practices or natural landscapes might reflect analogous economic conditions. By leveraging these visual similarities, a model should be able to apply insights gained from one region to predict the wealth of another region

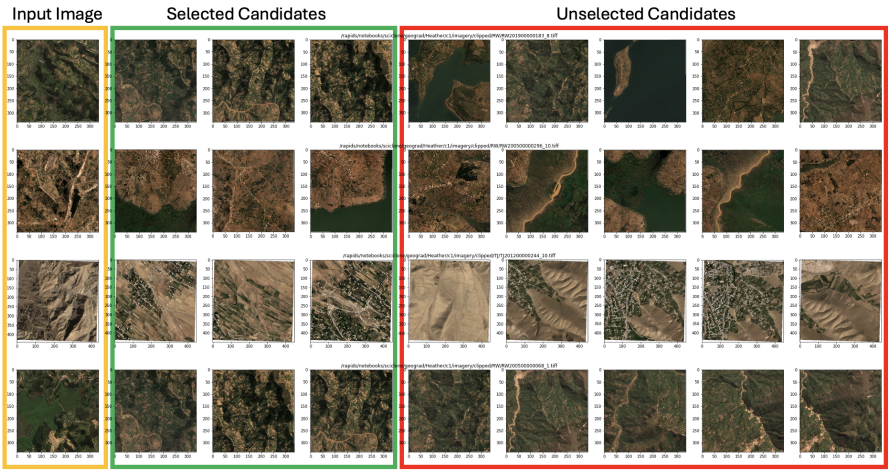

FIGURE 4.2: Outside country example

that it resembles visually, despite their physical separation.

An example of candidate images following this methodology is shown in Figure 4.3. The first image is each row is an example input image and the remaining images in the row are the first 4 most similar images based on style statistics. Each image is titled with the country from which is is taken.

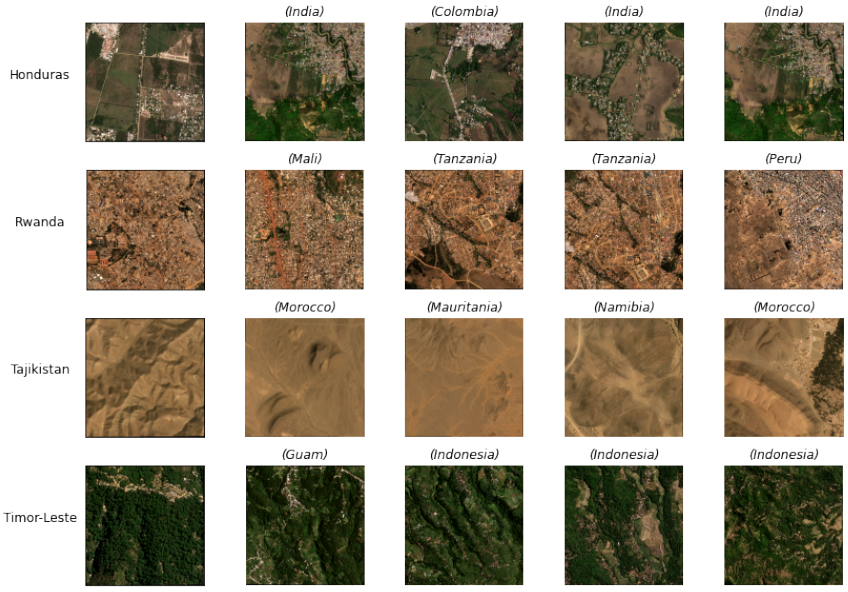

FIGURE 4.3: Example candidates based on style statistics

4.1.2.2 Alternative Candidate Selection Mechanisms

In its current implementation, GeoEmbed uses a similarity mechanism to select candidate images based on the theory that spatial entities close to one another are more likely to exhibit similar characteristics due to geographic proximity and spatial autocorrelation. This approach assumes that neighboring images, by virtue of being geographically close, are likely to contain similar contextual and visual information relevant for the model's predictions.

However, a secondary theory that attention-based mechanisms can better capture complex interdependencies between images by weighing the importance of different candidates based on the context of the source image, might lead to more robust candidate features and thereby stronger model generalization. Attention mechanisms, unlike simple similarity measures, allow the model to dynamically focus on the most relevant aspects of candidate images in relation to the source image, potentially uncovering subtle but informative patterns that are missed by the current approach. This shift towards an attention-based selection process could enhance the model's ability to discern and leverage critical information from a broader set of candidate images, improving its predictive performance across diverse geographic and contextual scenarios.

4.1.2.3 Regularization

GeoEmbed currently demonstrates a large degree of overfitting. I believe this is due to the limited augmentation and regularization techniques applied during training. To improve it, I plan to enhance the model's generalization capabilities by introducing more advanced data augmentation techniques to increase the diversity of the training dataset and implementing dropout and L2 regularization within the MLP to prevent overfitting. Additionally, experimenting with varying the number of candidate images based on contextual relevance rather than a fixed number may also help the model learn more robust embeddings. These steps aim to address the overfitting issue by both broadening the scope of learnable features and preventing the model from relying too heavily on the training data's specific characteristics.

Similarly, GeoEmbed freezes the weights of the DeepAll model before training, however I would like to explore if selectively unfreezing and fine-tuning

some of the top layers during training might allow the model to adapt more specifically to the task at hand.

4.2 Chapter 2: Spatially Adaptive Convolutional Networks with Coordinate-Conditioned Convolutional Layers

4.2.1 Motivation and Synopsis

In recent years, deep learning has significantly impacted the analysis of remote sensing and satellite imagery, enhancing our understanding of Earth's geography, environment, and socio-economic factors. Convolutional Neural Networks (CNNs) have been particularly effective in this field due to their capacity to automatically identify and learn features from large and complex datasets. However, conventional CNN models do not account for the geographical differences inherent in satellite images, such as the effects of local infrastructure, vegetation, and land use. This limitation can affect the models' accuracy in interpreting spatial data from various geographic areas.

To address this gap, my proposed chapter 2 will introduce an approach that incorporates spatially conditioned convolutional layers into existing CNN architectures. This methodology aims to enhance the model's sensitivity to geographic discrepancies by dynamically adjusting convolutional operations based on spatial metadata, thereby improving feature extraction relevancy and prediction accuracy across varied regions. The premise of this approach is rooted in the understanding that the significance of certain features in satellite imagery such as road quality, roof materials, crop productivity, and urban landscapes, differs vastly from one geographic location to another. By integrating spatial conditioning into a network, the model can modulate the feature extraction process to better align with region-specific characteristics.

My proposed second chapter will be structured as follows. I will begin with a description of the adaptive convolutional process in section 4.2.2. Next, I will relay our preliminary tests in section 4.2.4 and then discuss each of them in section 4.2.5. Finally, while the preliminary results show promise, the exploration reveals several limitations and areas necessitating further improvement; I will discuss my plan for improvement in section 4.2.6.

4.2.2 Methodology

The architecture of our coordinate-conditioned network is distinguished by incorporating adaptive convolutional layers into a ResNet18 framework (Figure 4.4), creating a contrast with traditional convolutional approaches.

In a traditional convolutional setup, earlier layers generally capture basic, low-level features such as edges and textures, while layers higher up extract more complex, abstract features that represent broader aspects of the input data. For instance, in the context of high-resolution satellite imagery, conventional higher layers might identify patterns indicative of urban versus rural areas without dynamically adjusting to the specific context of the image.

In contrast, adaptive convolutional layers, as implemented in our network, modify this paradigm by utilizing input-dependent weights. Early in the network, these adaptive layers can dynamically focus on features directly relevant to the specific input image's geographic context. For example, in regions where agricultural productivity is closely tied to wealth, an adaptive layer can emphasize features like crop types and vegetation health right from the initial stages, unlike standard convolutional layers that apply the same filters regardless of the input.

Similarly, adaptive layers placed lower in the network continue this tailored approach but focus on fine-tuning the detection of detailed, localized features with direct relevance to the image's specific geographic and socio-economic context. This could mean enhancing the recognition of infrastructure quality in urban settings or the density of housing, which standard convolutional layers might overlook or treat uniformly across diverse regions.

This distinction between adaptive and normal convolution emphasizes the adaptive layers' ability to modulate their processing based coordinates, enabling a more nuanced and context-aware feature extraction process throughout the network. This adaptability allows for a more precise and relevant analysis of spatial data, enhancing the model's performance in tasks requiring an understanding of spatial heterogeneity.

We implement the coordinate-conditioned convolutional layers using two modules: the HyperNet and the Adaptive Convolutional Layer. The HyperNet takes as input coordinates of samples and generates weights (and optionally biases) for the Adaptive Convolutional Layer. The Adaptive Convolutional Layer

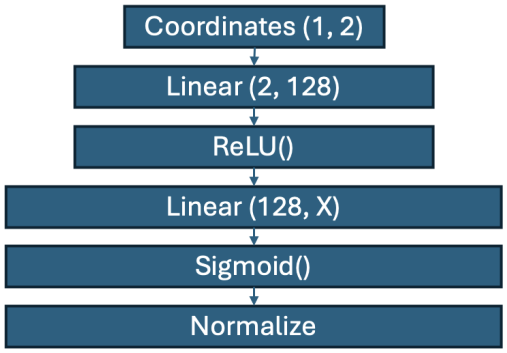

FIGURE 4.4: HyperNet Architecture

then uses these adaptive weights to perform convolution operations that are specifically tailored to the input data. The tasks and components of each module are further described in sections 4.2.2.1 and 4.2.2.2 respectively, while their current placement within the architecture is described in section 4.2.2.3.

4.2.2.1 HyperNet

The primary function of the HyperNet is to produce kernel weights for convolutional layers, adapting them based on specific input coordinates. This process begins with a linear layer that maps a 2-element coordinate input into a 128-dimensional latent space. This mapped output is then processed through an activation function, leading to a second fully connected layer that further maps these latent features to a targeted output dimension suitable for creating convolutional kernels. This vector is then reshaped to conform to the precise kernel shape needed for convolution, enabling the adaptive layer to apply these custom-tailored kernels to the input data effectively.

For instance, consider a convolutional layer characterized by dimensions [64, 64, 3, 3]. This layer possesses 64 output channels, 64 input channels, and utilizes a 3x3 kernel for convolution operations. Consequently, the output from the HyperNet would be a tensor with dimensions [1, 36864], containing the requisite number of elements to reshape into the convolutional layer's adaptive weights and biases. This tensor is then reshaped back into the original convolutional layer's structure, [64, 64, 3, 3], thereby integrating the dynamically generated parameters for subsequent processing.

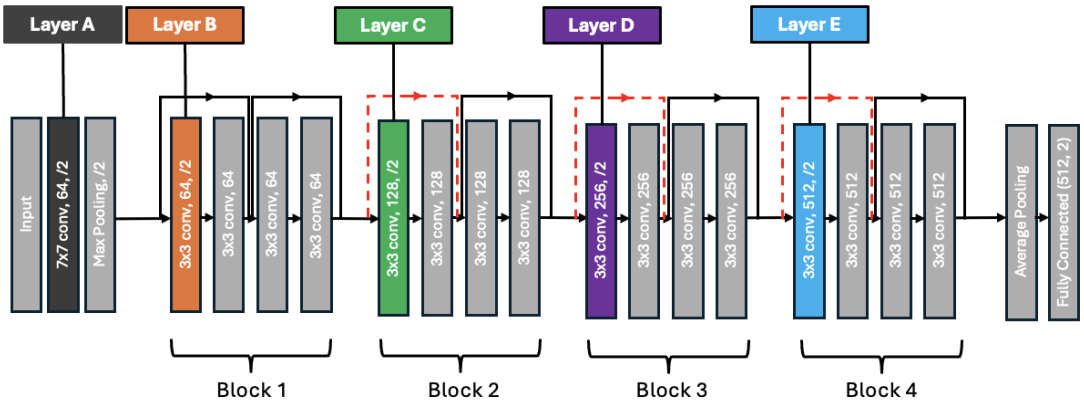

FIGURE 4.5: ResNet18 Architecture

4.2.2.2 Adaptive Convolutional Layer

Utilizing the dynamically generated kernel from the HyperNet, the adaptive convolutional layer performs a convolution operation on the image.

The adaptive convolutional layer is designed to perform the same function as a traditional convolutional layer, which is to identify patterns in the image that are relevant for the classification of a given target. In standard convolutional layers, weights for each kernel are refined through back-propagation to enhance pattern recognition. However, our adaptive convolutional layer diverges from this standard methodology by utilizing weights directly provided by the HyperNet, which are specifically tailored for the current input image.

When the adaptive convolutional layer receives these custom weights from the HyperNet, it first reshapes the one-dimensional vector into a convolutional kernel, as detailed in 4.2.2.1. Following this, a convolutional layer is set up with these adaptive weights, enabling it to process input features with a level of specificity and relevance not achievable through traditional, statically weighted layers.

4.2.2.3 Layer Placement

In the current best performing implementation, we include two HyperNet and adaptive convolutional layers within a standard ResNet18 architecture. Specifically, ResNet18 contains 4 blocks after an initial set of convolutional, batch normalization and activation layers. We replaced the first convolutional layer (layer

A in Figure 4.5) and the first convolutional layer in the last block (layer E in Figure 4.5) with HyperNet and adaptive convolutional layers.

The network's operation is structured into three distinct stages. In the first stage, the network leverages an adaptive convolutional layer for initial feature extraction. The second stage involves standard convolutional layers for further feature reduction, and the third stage incorporates a final adaptive layer for nuanced feature modulation. These steps are expanded on further in the below paragraphs.

The first stage of the network constitutes a convolutional phase, where, unlike traditional approaches that derive weights through back-propagation, we utilize weights computed by the HyperNet, as detailed in section 4.2.2.1. This adaptability allows the network to fine-tune its convolutional filters to the specific characteristics of the input data, enabling a more targeted extraction of features relevant to wealth indicators within the geographic context. For example, in regions where wealth is closely linked with agricultural productivity, the network, through its adaptive layer, focuses on pertinent features such as vegetation coverage or crop types. In contrast, urban settings prompt adjustments towards features indicative of housing density, infrastructure quality, or commercial activities.

In stage two, the process transitions to using standard convolutional layers for additional feature reduction. This choice stems from the understanding that once the initial, crucial set of features has been selectively emphasized through the adaptive process, the focus should shift towards condensing these features into a more compact representation. This phase is designed to refine the extracted features, aiming to reduce dimensionality.

The network's third stage introduces a second adaptive layer within the final block, positioned just before the concluding set of feature reduction convolutions. The rationale for this late-stage adaptive intervention is based on the hypothesis that the feature maps at this juncture, having been distilled through prior processing, encapsulate a concentrated yet diverse array of wealth indicators that could benefit from one last round of adaptive fine-tuning. This final adaptive layer allows the network to make adjustments to the feature maps, ensuring the network's focus is optimally aligned with the most salient features for wealth estimation. This strategic placement facilitates final modifications based

on the condensed feature set, enhancing the network's ability to prioritize the most predictive elements.

4.2.3 Experiments

We utilize the same USAID DHS and Planet Imagery datasets as described in sections 3.3.1 and 3.3.2, limiting the geographic scope to the Western Africa, including Burkina Faso, Cote d'Viore, Guinea, The Gambia, Guinea-Bissau, Liberia, Mali, Sierra Leone, Senegal, Togo. The scope of our implementation was limited in order to develop our method more efficiently on a small scale, and we plan to incorporate the remaining global data at a later stage.

4.2.4 Preliminary Results

The results of our preliminary testing using the spatially-conditioned convolutional layers - HyperNets - are shown in Table 4.2

TABLE 4.2: Experiment Results Summary

Description	**Adaptive Layers 4.5**	**Batch Size**	**R^2**	**MAE**	**ID**
Standard ResNet18 architecture	N/A	-	0.4925	**0.0931**	A
Spatial-Aware: Coordinates into latent space, FC layer	N/A	256	0.4813	0.1048	B
Coordinate-conditioned convolutional layer in first conv. layer and block 4; Conditional weights are normalized before they are used	B & E	64	**0.5229**	0.0969	C
Coordinate-conditioned convolutional layers in blocks 1 & 4	B & E	64	0.4678	0.0987	D
Coordinate-conditioned convolutional layer in block 1	B	256	0.3728	0.1075	E

Continued on next page

Table 4.2 – *Continued from previous page*

Description	**Adaptive Layers 4.5**	**Batch Size**	**R^2**	**MAE**	**ID**
Coordinate-conditioned convolutional layer in block 4	E	256	0.4153	0.1066	F
Coordinate-conditioned convolutional layer in first conv. layer and block 4	A & E	64	0.4738	0.0989	G
Fully Connected (512-feature) layer uses dynamic weights; No Coordinate-conditioned convolutional layers	N/A	64	0.3911	0.1072	H

4.2.5 Discussion

While the spatially-conditioned convolutional architecture is still in development, Table 4.2 shows the current progress of each perturbation to the architecture and training parameters. Below, I offer some explanations as to the relative performance of each perturbation.

- **DeepAll (Model A in Table 4.2.4):** The basic configuration with no modifications achieved solid performance, suggesting the robustness of the standard ResNet18 architecture. This acts as a strong baseline, indicating the model's capacity to generalize without any specialized adaptation.
- **SpAware (Model B in Table 4.2.4):** Incorporating spatial awareness slightly improved the R^2 but increased MAE, suggesting that making the model spatially-aware offers improvements, but may not be as powerful as method of spatial awareness compared to
- **Coordinate-conditioned convolutional layer - Layers 1 & 4 (Model C in Table 4.2.4):** Positioning adaptive layers in the first and fourth blocks of ResNet18 had varying impacts.

- **Coordinate-conditioned convolutional layer - Layer 1 (Model D in Table 4.2.4):** This setup, with adaptive convolution only in the first layer, showed a decrease in performance, possibly indicating that early adaptation might not capture the depth of spatial features needed for accurate predictions.
- **Coordinate-conditioned convolutional layer - Layer 4 (Model E in Table 4.2.4):** Placing an adaptive layer in the fourth block alone performed better than just the first block, suggesting that later-stage spatial adaptation is beneficial but still not as effective as the more integrated approaches.
- **Coordinate-conditioned convolutional layer - First Conv & Layer 4 (Model F in Table 4.2.4):** Introducing adaptive convolution both at the very beginning and in the fourth block balanced initial spatial awareness with deeper feature refinement, reflected in improved performance, especially in terms of MAE.
- **Fully Connected Predictor (Model G in Table 4.2.4):** This approach underperformed, indicating that dynamic weight prediction for the fully connected layer alone may not sufficiently capture the spatial complexities of the data.
- **Coordinate-conditioned convolutional layer - First Conv & Layer 4 - Normalized Coordinates/Weights (Model H in Table 4.2.4):** Tests with normalization of weights alone led to the best performance among all tests. This suggests that appropriate normalization can significantly enhance model sensitivity to spatial features by aligning the scale of the adaptive weights with the network's internal representations.

4.2.6 Limitations and Plan for Improvement

While the preliminary implementation of spatially conditioned convolutional layers is promising, there are many limitations surrounding its current implementation and improvements needed.

4.2.6.1 Layer Placement

There are two different strategies I intend to pursue in terms of improving layer placements. The first is in the context of Bottleneck Layers. Bottleneck layers are designed to reduce the dimensionality of inputs before applying a more computationally expensive operation, such as a 3x3 convolution, and then restoring dimensionality. Placing adaptive layers here could allow the network to dynamically adjust the information flow based on spatial context at these critical points.

The second strategy is to take advantage of residual connections. Placing adaptive layers just before residual connections may allow the network to dynamically influence how information is combined and propagated through the network. This could potentially enhance the network's ability to incorporate spatially relevant features into its deeper layers.

4.3 Chapter 3: torchSDG

4.3.1 Motivation and Synopsis

As machine learning, particularly deep learning, advances and becomes more integrated into various fields, the importance of benchmarking the performance of new methods grows. This is crucial for tracking progress within specific domains. ImageNet for example, was on of the first publicly available benchmark datasets for computer vision, and is still used as a dataset for benchmarking methodological progress [14]. Deep learning applications in analyzing satellite imagery to predict sociodemographic data represent a niche yet rapidly evolving area. Traditionally, this field has attracted more applied practitioners and has lacked standardized benchmarks and methodologies, making comparative evaluation of progress challenging.

In my third paper, I will introduce torchSDG, a package designed to benchmark the performance of machine learning models using satellite imagery to predict indicators relevant to the Sustainable Development Goals (SDGs). torchGeo was the first package released for the benchmarking of deep learning models utilizing satellite imagery in 2022 [69]. torchSDG will build on the foundations of torchGeo without replicating resources such as geospatial samplers and loss functions. TorchSDG will be a prototype for a large application to all SDG's, and will start with the benchmarking of SDG 4. It will provide open-source access to Sentinel datasets, tabular geocoded datasets, and models for comparison.

For researchers, this will offer a common ground for comparing algorithmic innovations, facilitating the validation of new ideas. The availability of benchmark models and datasets will lower the entry barrier for new researchers and practitioners into this field, democratizing access to state-of-the-art tools and data. For policymakers and stakeholders focused on sustainable development, torchSDG will provide evidence-based insights and models that are directly applicable to monitoring and achieving SDG targets. These models, trained and evaluated against standardized benchmarks, can serve as reliable tools for decision-making, planning, and resource allocation.

Furthermore, by publishing open-source Sentinel datasets and tabular geocoded

datasets alongside these benchmarks, torchSDG enables a transparent and reproducible research environment. This transparency is crucial for trust in machine learning applications, particularly in areas as impactful as policy making and resource distribution for sustainable development. The collective advancement in accurately predicting and understanding sociodemographic metrics from satellite imagery, as facilitated by torchSDG, will support more informed, data-driven policy development. Ultimately, this will contribute to the global effort to achieve the SDGs, ensuring that interventions are guided by reliable data and robust analytical methods.

There are 231 unique subindicators included in the SDG's. Given the enormity of the task required to benchmark every indicator, I have selected SDG4 as a prototype task for torchSDG. Therefore, for the remainder of this proposal, I will limit my discussion to methods (section 4.3.3), data (section 4.3.2) and preliminary results (section **??**) for for SDG4.

4.3.2 Data

4.3.2.1 SDG4 Tabular Datasets

The list of datasets and benchmarks models that I am proposing to publish as part of the pilot are included in table 4.3.2.1.

TABLE 4.3: Educational Subindicators, Data Sources, and Prediction Granularity

Sub-indicator Number	**Sub-indicator Description**	**Data Sources**	**Level of Granularity of prediction**
4.1.1	Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex	This Dissertation	6720x6720 meter Grid
4.2.2	Participation rate in organized learning (one year before the official primary entry age), by sex	Meta High Resolution Population Density Maps; This Dissertation	6720x6720 meter Grid
4.3.1	Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, by sex	Meta High Resolution Population Density Maps; This Dissertation	6720x6720 meter Grid
4.4.1	Proportion of youth and adults with information and communications technology (ICT) skills, by type of skill	This Dissertation	6720x6720 meter Grid

Continued on next page

Table 4.3 – continued from previous page

Subindicator Number	**Subindicator**	**Data Sources**	**Level of Granularity of prediction**
4.5.1	Parity indices (female/male, rural/urban, bottom/top wealth quintile and others such as disability status, indigenous peoples and conflict-affected, as data become available) for all education indicators on this list that can be disaggregated	This Dissertation	6720x6720 meter Grid
4.6.1	Percentage of population in a given age group achieving at least a fixed level of proficiency in functional (a) literacy and (b) numeracy skills, by sex	This Dissertation	6720x6720 meter Grid
4.a.1	Proportion of schools with access to: (a) electricity; (b) the Internet for pedagogical purposes; (c) computers for pedagogical purposes; (d) adapted infrastructure and materials for students with disabilities; (e) basic drinking water; (f) single-sex basic sanitation facilities; and (g) basic handwashing facilities (as per the WASH indicator definitions)	This Dissertation	School

Continued on next page

Table 4.3 – continued from previous page

Subindicator Number	**Subindicator**	**Data Sources**	**Level of Granularity of prediction**
4.c.1	Proportion of teachers in: (a) pre-primary; (b) primary; (c) lower secondary; and (d) upper secondary education who have received at least the minimum organized teacher training (e.g., pedagogical training) pre-service or in-service required for teaching at the relevant level in a given country	This Dissertation	6720x6720 meter Grid

4.3.2.2 Meta High Resolution Population Density Maps

The high-resolution population density maps involves a collaboration between Facebook Connectivity Lab and CIESIN at Columbia University [18]. It combines machine vision AI, satellite imagery, and census information to generate detailed population density maps. Additionally, demographic data related to age and gender are integrated, providing insights into the population's location and demographics across most countries globally. These datasets were chosen for this project because they provide population estimates broken down by age groups relevant to the prediction of public school indicators.

4.3.2.3 Public School Data

We have been collected open access data on public schools for the past year. These data collection efforts have largely targeted open government data portals and other international agency sources such as the Red Cross and the United Nations Office for the Coordination of Humanitarian Affairs (OCHA). The data that is collected is cleaned and standardized into 4 main tables within a PostreSQL

database: spatial (including school coordinates and administrative unit information), school personnel (including student enrollment and number of teachers broken down by gender), school resources (including access to internet, electricity) and school outcomes (including national testing data and matriculation rates). This database will serve as the primary source of education data for torchSDG models.

4.3.2.4 Satellite Imagery

While high-resolution imagery, such as that from Planet, is the preferred source for predicting socioeconomic indicators, it is not available for redistribution and, therefore, cannot be used in this project. As a result, we have chosen Sentinel-2 imagery as our primary data source [17]. The Sentinel-2 satellite, a part of the European Space Agency's Copernicus program, is renowned for its global monitoring capabilities, offering high-resolution, multispectral imagery. Equipped with the Multispectral Imager (MSI), Sentinel-2 provides images with resolutions of 10 m, 20 m, and 60 m across 13 spectral bands and has a revisit time of 10 days, making it well-suited for our project's needs.

In order to enhance comprability to other sources, the imagery in our model will use the red, green and blue bands stacked on natural color order and each input image will be clipped to a standard input size of 3 channels, each 256x256 pixels. Each image will be standardized upon input to the model to the global mean and standard deviation of each channel in order to maintain uniform standards for input across all models.

4.3.3 Methods

The objective of torchSDG is to establish benchmark models and datasets for Sustainable Development Goal (SDG) indicators, starting with a pilot using SDG 4. The subindicators of SDG 4 are categorized into two main types based on the nature of the information they contain. The first type encompasses data that is spatially continuous and not explicitly linked to schools. This category includes the majority of subindicators, which will be analyzed using a grid-level

approach detailed in Section 4.3.3.1. Conversely, Subindicator 4.a.1, which pertains to resources available at schools, necessitates predictions to be made on an individual school basis, as outlined in Section 4.3.3.2.

4.3.3.1 Grid Level

In this section, I discuss the creation of grid-based products for a benchmarking dataset focused on school quality. The output from this methodology will be a global grid, in which for each grid cell a satellite image and subindicator value are made available. The majority of subindicators in Table 4.3.2.1 will be predicted within a grid, because they reference school catchment areas as opposed to individual schools. Due to the myriad of different methods for creating school catchment zones across the globe, for instance parent's choice or geographic proximity, and in an effort to standardize how predictions are made, we will follow the following methodology to create school catchment grids based on geographic proximity:

1. The standard width and height of an input image to a CNN model is 256x256 pixels. Sentinel has a 10 meter resolution, so in order to generate a grid where every cell will represent an input image of the standard size, the cells need to each be $10 * 256$ meters, which equates to a 2,560x2,560 meter grid covering each country in which we have data for a subindicator.

2. Overlay school points on top of the generated grid.

3. Assign each grid cell to the closest school point based on proximity. If a cell has multiple schools that are equidistant to it, the school assigned to it will be randomly chosen amongst the closest schools.

4. Overlay additional relevant data (e.g., population density) on the grid, assigning values to each grid cell based on overlap.

5. For each point of interest, calculate specific metrics using the data assigned to the grid cells, such as summing population values and dividing by relevant enrollment figures to calculate participation rates.

FIGURE 4.6: School imagery examples

6. Download satellite imagery for each grid cell and label it with the calculated metric (e.g., participation rate).

4.3.3.2 School Level Predictions

School level predictions for subindicator 4.a.1 will utilize the following methodology:

1. Generate a 256meter x 256meter buffer around each school point in the spatial dataset described in section 4.3.2.3.
2. Download Sentinel imagery within each buffer.
3. Label the image with its school's associated metric (i.e. access to electricity)

Examples imagery clips in for schools in the Philippines are shown in Figure 4.6.

Chapter 5

Timeline

In the initial phase of my dissertation research, I will work to overcome the limitations of my current implementations of chapter 1 outlined in 4.1.2. In task 1 I will test different candidate proposal strategies for my GeoEmbed model. In task 2, I will test different candidate selection strategies and in task 3 I will test the implementation of regularization techniques to prevent overfitting.

Following the completion of my GeoEmbed model in Chapter 1, I will shift my attention to the adaptive convolutions proposed in Chapter 2 and their current limitations, as described in section 4.2.6. The majority of my effort will be spent on identifying the ideal adaptive layer placement within my adaptive convolutional network. This work will culminate in June 2024.

The concluding phase of my dissertation project will involve a series of data collection and prediction tasks for the implementation of torchSDG, as describe in section 4.3.1. While this work has already begun and will be carried out throughout the next year, the bulk of the work is projected to begin in August 2024 and be carried out intermittently through June 2025. In parallel, I will begin drafting of my dissertation and preparing for my defense in early spring 2025, for a defense date in early summer.

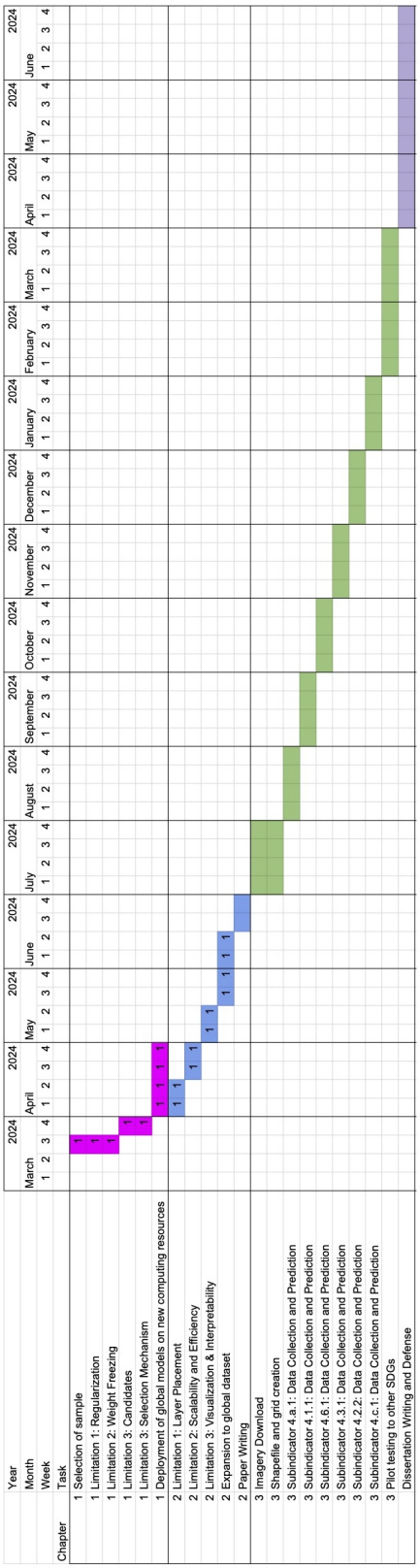

FIGURE 5.1: Timeline

Bibliography

[1] 2023. URL: https://www.usaid.gov/global-health/demographic-and-health-surveys-program.

[2] A. P. S. C. Almeida et al. "Socioeconomic determinants of access to health services among older adults: a systematic review". In: *Revista De Saúde Pública* 51 (0 2017). DOI: 10.1590/s1518-8787.2017051006661.

[3] S. Amit et al. "Analysis of satellite images for disaster detection". In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2016, pp. 5189–5192. DOI: 10.1109/IGARSS.2016.7730352. URL: https://consensus.app/papers/analysis-satellite-images-disaster-detection-amit/c7d53ef80de1554fa854a1935235d58d/?utm_source=chatgpt.

[4] Kumar Ayush et al. "Efficient poverty mapping from high resolution remote sensing images". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 1. 2021, pp. 12–20.

[5] Boris Babenko et al. "Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico". In: *arXiv preprint arXiv:1711.06323* (2017).

[6] Nadir Bengana and Janne Heikkilä. "Improving land cover segmentation across satellites using domain adaptation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2020), pp. 1399–1410.

[7] Sergio Casas et al. "SpAGNN: Spatially-Aware Graph Neural Networks for Relational Behavior Forecasting from Sensor Data". en. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. Paris, France: IEEE, May 2020, pp. 9491–9497. ISBN: 978-1-72817-395-5. DOI: 10.1109/

ICRA40945.2020.9196697. URL: https://ieeexplore.ieee.org/document/9196697/ (visited on 03/15/2024).

[8] Guanghua Chi et al. "Microestimates of wealth for all low-and middle-income countries". In: *Proceedings of the National Academy of Sciences* 119.3 (2022), e2113658119.

[9] E. Chuvieco. *Fundamentals of Satellite Remote Sensing: An Environmental Approach*. CRC Press, 2016.

[10] Isaac Corley et al. "Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters". In: *arXiv preprint arXiv:2305.13456* (2023).

[11] Robert Costanza, Lorenzo Fioramonti, and Ida Kubiszewski. "The UN Sustainable Development Goals and the dynamics of well-being". In: *Frontiers in Ecology and the Environment* 14.2 (2016), pp. 59–64. DOI: 10.1002/fee.1231.

[12] National Research Council et al. *People and pixels: Linking remote sensing and social science*. National Academies Press, 1998.

[13] Ilke Demir et al. "Deepglobe 2018: A challenge to parse the earth through satellite images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 172–181.

[14] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[15] *DigitalGlobe Satellite Imagery*. https://www.maxar.com/. Accessed: 2023-09-30. 2023.

[16] Ryan Engstrom, Jonathan Samuel Hersh, and David Locke Newhouse. "Poverty from space: using high-resolution satellite imagery for estimating economic well-being". In: *World Bank Policy Research Working Paper* 8284 (2017).

[17] European Space Agency. *Sentinel-2 MSI: MultiSpectral Instrument, Level-1C*. https://scihub.copernicus.eu/. Accessed: INSERT-DATE-HERE. 2017.

[18] Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University. *High Resolution Settlement Layer (HRSL)*. Source imagery for HRSL © 2016 DigitalGlobe. Accessed 28 3 2024. 2016.

[19] E. Fussell, N. Sastry, and M. VanLandingham. "Race, socioeconomic status, and return migration to new orleans after hurricane katrina". In: *Population and Environment* 31 (1-3 2009), pp. 20–42. DOI: 10.1007/s11111-009-0092-2.

[20] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. "A neural algorithm of artistic style". In: *arXiv preprint arXiv:1508.06576* (2015).

[21] Seth Goodman, Ariel BenYishay, and Daniel M. Runfola. "A convolutional neural network approach to predict non-permissive environments from moderate-resolution imagery". In: *Transactions in GIS* 25 (2020), pp. 674 – 691.

[22] *Google Earth Engine Basemaps*. https://earthengine.google.com/. Accessed: 2023-09-30. 2023.

[23] David Griggs et al. "Policy: Sustainable development goals for people and planet". In: *Nature* 495 (2013), pp. 305–307. DOI: 10.1038/495305a.

[24] Jayant Gupta, Yiqun Xie, and Shashi Shekhar. *Towards Spatial Variability Aware Deep Neural Networks (SVANN): A Summary of Results*. en. arXiv:2011.08992 [cs]. Nov. 2020. URL: http://arxiv.org/abs/2011.08992 (visited on 03/15/2024).

[25] Jayant Gupta et al. "Spatial Variability Aware Deep Neural Networks (SVANN): A General Approach". en. In: *ACM Transactions on Intelligent Systems and Technology* 12.6 (Dec. 2021), pp. 1–21. ISSN: 2157-6904, 2157-6912. DOI: 10.1145/3466688. URL: https://dl.acm.org/doi/10.1145/3466688 (visited on 03/15/2024).

[26] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[27] Zhiyuan He et al. "Perceiving Commerial Activeness Over Satellite Images". In: *Companion Proceedings of the The Web Conference 2018*. 2018, pp. 387–394.

[28] Patrick Helber et al. "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.7 (2019), pp. 2217–2226.

[29] J. Iqbal and Mohsen Ali. "Weakly Supervised Domain Adaptation for Built-up Region Segmentation in Aerial and Satellite Imagery". In: *ArXiv* abs/2007.02277 (2020). DOI: 10.1016/j.isprsjprs.2020.07.001. URL: https://consensus.app/papers/weakly-supervised-domain-adaptation-builtup-region-iqbal/99ecf0bdd57b5d44a94b3bb2f6646ea0/?utm_source=chatgpt.

[30] Javed Iqbal and Mohsen Ali. "Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 167 (2020), pp. 263–275.

[31] Robin Jarry et al. "Assessment of CNN-based methods for poverty estimation from satellite images". In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 550–565.

[32] Neal Jean et al. "Combining satellite imagery and machine learning to predict poverty". In: *Science* 353.6301 (2016), pp. 790–794. DOI: 10.1126/science.aaf7894. eprint: https://www.science.org/doi/pdf/10.1126/science.aaf7894. URL: https://www.science.org/doi/abs/10.1126/science.aaf7894.

[33] Neal Jean et al. "Combining satellite imagery and machine learning to predict poverty". In: *Science* 353.6301 (2016), pp. 790–794.

[34] Neal Jean et al. "Tile2vec: Unsupervised representation learning for spatially distributed data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 3967–3974.

[35] John R Jensen and Dave C Cowen. "Remote sensing of urban/suburban infrastructure and socio-economic attributes". In: *Photogrammetric engineering and remote sensing* 65 (1999), pp. 611–622.

[36] John R Jensen and Kalmesh Lulla. "Introductory digital image processing: a remote sensing perspective". In: (1987).

[37] J. B. Kirby and T. Kaneda. "Neighborhood socioeconomic disadvantage and access to health care". In: *Journal of Health and Social Behavior* 46 (1 2005), pp. 15–31. DOI: 10.1177/002214650504600103.

[38] Yohei Koga, Hiroyuki Miyazaki, and Ryosuke Shibasaki. "A method for vehicle detection in high-resolution satellite images that uses a region-based object detector and unsupervised domain adaptation". In: *Remote Sensing* 12.3 (2020), p. 575.

[39] Alexandre Lacoste et al. "Geo-bench: Toward foundation models for earth monitoring". In: *Advances in Neural Information Processing Systems* 36 (2024).

[40] Walter Leal Filho et al. "Assessing the impacts of Sustainable Development Goals on global research: a bibliometric analysis". In: *International Journal of Sustainable Development & World Ecology* 27.7 (2020), pp. 595–607. DOI: 10.1080/13504509.2020.1819864.

[41] Haoyu Liu et al. "Nightlight as a proxy of economic indicators: Fine-grained gdp inference around chinese mainland via attention-augmented cnn from daytime satellite imagery". In: *Remote Sensing* 13.11 (2021), p. 2067.

[42] W. Liu. "Convolutional Neural Network Based Landcover Analysis of Satellite Images". In: *Journal of Physics: Conference Series*. Vol. 1345. 2019.

[43] Benjamin Lucas et al. "Unsupervised domain adaptation techniques for classification of satellite image time series". In: *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2020, pp. 1074–1077.

[44] Emmanuel Maggiori et al. "Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 55 (2017), pp. 645–657.

[45] Nikhil Makkar, Lexie Yang, and S. Prasad. "Adversarial Learning Based Discriminative Domain Adaptation for Geospatial Image Analysis". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 150–162. DOI: 10.1109/jstars.2021.3132259. URL:

https://consensus.app/papers/adversarial-learning-based-discriminative-domain-makkar/c473d5ce0747577ea019ec1e1daf6300/?utm_source=chatgpt.

[46] Toshihiko Matsuura and Tatsuya Harada. "Domain generalization using a mixture of multiple latent domains". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 11749–11756.

[47] Maxar Technologies. *Maxar Satellite Imagery Specifications*. https://www.maxar.com/. Accessed: 2023-09-30. 2023.

[48] Charlotta Mellander et al. "Night-time light data: A good proxy measure for economic activity?" In: *PloS one* 10.10 (2015), e0139779.

[49] NASA and USGS. *Landsat Program*. https://www.usgs.gov/land-resources/nli/landsat. Accessed: 2023-09-30. 2023.

[50] NASA MODIS Land Team. *MOD11A1 V006 Product User Guide*. Accessed: 2023-09-30. 2023. URL: https://lpdaac.usgs.gov/documents/118/MOD11_User_Guide_V6.pdf.

[51] Thanh Tam Nguyen et al. "Monitoring agriculture areas with satellite images and deep learning". In: *Applied Soft Computing* 95 (2020), p. 106565.

[52] Ye Ni et al. "An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction". In: *IEEE Geoscience and Remote Sensing Letters* 18.9 (2020), pp. 1545–1549.

[53] Abdisalan M Noor et al. "Using remotely sensed night-time light as a proxy for poverty in Africa". In: *Population Health Metrics* 6.1 (2008), pp. 1–13.

[54] Zhuokun Pan et al. "Deep learning segmentation and classification for urban village using a worldview satellite image based on U-Net". In: *Remote Sensing* 12.10 (2020), p. 1574.

[55] Sungwon Park et al. "Learning economic indicators by aggregating multi-level geospatial information". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 12053–12061.

[56] Anthony Perez et al. "Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty". In: *arXiv preprint arXiv:1902.11110* (2019).

[57] *PlanetScope Satellite Imagery*. https://www.planet.com/. Accessed: 2023-09-30. 2023.

[58] Daniel Rammer et al. "Small is Beautiful: Distributed Orchestration of Spatial Deep Learning Workloads". In: *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)*. 2020, pp. 101–111. DOI: 10.1109/UCC48980.2020.00029. URL: https://consensus.app/papers/beautiful-distributed-orchestration-spatial-deep-rammer/f7f97fcb68df5a98b2afe5c7?utm_source=chatgpt.

[59] D.P. Roy et al. "Landsat-8: Science and product vision for terrestrial global change research". In: *Remote Sensing of Environment* 145 (2014), pp. 154–172.

[60] D Runfola, A Stefanidis, and H Baier. "Using satellite data and deep learning to estimate educational outcomes in data-sparse environments". In: *Remote Sensing Letters* 13.1 (2022), pp. 87–97.

[61] Dan Runfola et al. "A multi-glimpse deep learning architecture to estimate socioeconomic census metrics in the context of extreme scope variance". In: *International Journal of Geographical Information Science* (2024), pp. 1–25.

[62] Daniel Runfola et al. "Deep learning fusion of satellite and social information to estimate human migratory flows". In: *Transactions in GIS* 26.6 (2022), pp. 2495–2518.

[63] Jeffrey D. Sachs. *The Age of Sustainable Development*. Columbia University Press, 2015. ISBN: 9780231173155.

[64] Vasit Sagan et al. "Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning". In: *ISPRS journal of photogrammetry and remote sensing* 174 (2021), pp. 265–281.

[65] Rahman Sanya, Gilbert Maiga, and Ernest Mwebaze. "Using Convolutional Networks and Satellite Imagery to Predict Disease Density in a Developing Country". In: *Preprints* (2019).

[66] Michael Schmitt et al. "SEN12MS–A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion". In: *arXiv preprint arXiv:1906.07789* (2019).

[67] Q. Shi et al. "Domain Adaption for Fine-Grained Urban Village Extraction From Satellite Images". In: *IEEE Geoscience and Remote Sensing Letters* 17 (2020), pp. 1430–1434. DOI: 10.1109/LGRS.2019.2947473. URL: https://consensus.app/papers/domain-adaption-finegrained-urban-village-extraction-shi/bdd73e0925245bd19ae51d47b0ac3ade/?utm_source=chatgpt.

[68] A. Stewart et al. "TorchGeo: deep learning with geospatial data". In: *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 2021. DOI: 10.1145/3557915.3560953. URL: https://consensus.app/papers/torchgeo-learning-data-stewart/0e408211d72e546196fa2cf0daf43?utm_source=chatgpt.

[69] Adam J Stewart et al. "Torchgeo: deep learning with geospatial data". In: *Proceedings of the 30th international conference on advances in geographic information systems*. 2022, pp. 1–12.

[70] Gencer Sumbul et al. "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding". In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2019, pp. 5901–5904.

[71] Michiaki Tatsubori et al. "A Programming Model for Geospatial Machine-Learning with Scalability in Hybrid Multiclouds". In: *EGU General Assembly Conference Abstracts*. 2023, EGU–3441.

[72] Isabelle Tingzon et al. "MAPPING POVERTY IN THE PHILIPPINES USING MACHINE LEARNING, SATELLITE IMAGERY, AND CROWD-SOURCED GEOSPATIAL INFORMATION." In: *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* (2019).

[73] United Nations Development Programme. "Supporting the 2030 Agenda for Sustainable Development: UNDP's Approach". In: *United Nations Development Programme Reports*. 2015. URL: https://www.undp.org/content/undp/en/home/librarypage/results/supporting-the-2030-agenda-for-sustainable-development.html.

[74] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. "Spacenet: A remote sensing dataset and challenge series". In: *arXiv preprint arXiv:1807.01232* (2018).

[75] Q. Weng. *Remote Sensing of Urban and Suburban Areas*. Springer, 2016.

[76] Congxi Xiao et al. "Spatial Heterophily Aware Graph Neural Networks". en. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Long Beach CA USA: ACM, Aug. 2023, pp. 2752–2763. ISBN: 9798400701030. DOI: 10.1145/3580305.3599510. URL: https://dl.acm.org/doi/10.1145/3580305.3599510 (visited on 03/15/2024).

[77] Yiqun Xie et al. "A Statistically-Guided Deep Network Transformation and Moderation Framework for Data with Spatial Heterogeneity". In: *2021 IEEE International Conference on Data Mining (ICDM)*. 2021, pp. 767–776. DOI: 10.1109/ICDM51629.2021.00088. URL: https://consensus.app/papers/statisticallyguided-deep-network-transformation-xie/bbb7e5344f0954369234dce8cfaeaa36/?utm_source=chatgpt.

[78] Yi Yang and Shawn Newsam. "Bag-of-visual-words and spatial extensions for land-use classification". In: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. 2010, pp. 270–279.

[79] Christopher Yeh et al. "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa". In: *Nature communications* 11.1 (2020), p. 2583.

[80] Fahong Zhang, Yilei Shi, and Xiaoxiang Zhu. "Self-supervised Domain-agnostic Domain Adaptation for Satellite Images". In: *ArXiv* abs/2309.11109 (2023). DOI: 10.48550/arXiv.2309.11109. URL: https://consensus.app/papers/selfsupervised-domain-adaptation-satellite-images-zhang/ed9fa876d2f4517aaa87c123f79524b5/?utm_source=chatgpt.

[81] Xizhi Zhao et al. "Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh". In: *Remote Sensing* 11.4 (2019), p. 375.

[82] X. Zhu and D. Tuia. "Deep learning in remote sensing: A comprehensive review and list of resources". In: *IEEE Geoscience and Remote Sensing Magazine* 5.4 (2017), pp. 8–36.

[83] Xiao Xiang Zhu et al. "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [Software and Data Sets]". In: *IEEE Geoscience and Remote Sensing Magazine* 8.3 (2020), pp. 76–89.