Rethink Geographical Generalizability with Unsupervised Self-Attention Model Ensemble: A Case Study of OpenStreetMap Missing Building Detection in Africa

Hao Li
Technical University of Munich
Germany
hao_bgd.li@tum.de

Jiapan Wang
Technical University of Munich
Germany
jiapan.wang@tum.de

Johann Maximilian Zollner
Technical University of Munich
Germany
maximilian.zollner@tum.de

Gengchen Mai
University of Georgia
USA
gengchen.mai25@uga.edu

Ni Lao
Mosaix.ai
USA
noon99@gmail.com

Martin Werner
Technical University of Munich
Germany
martin.werner@tum.de

ABSTRACT

The recent advance of adapting pre-trained task-agnostic artificial intelligence (AI) models leads to great successes in downstream tasks via fine-tuning, or low-resource (i.e., few-shot and zero-shot) learning. However, when adapting such pre-trained AI models to geographical applications, it is still challenging to find the "sweet spot" of the model's generalizability and specializability (e.g., geographic generalizability v.s. spatial heterogeneity). For instance, a building detection task may require vision models with different parameters across different geographic areas of the world. In this paper, we rethink this interesting topic, namely *Geographical Generalizability* of GeoAI models, with a case study of detecting OpenStreetMap (OSM) missing buildings across different countries in sub-Saharan Africa. We consider a real-world scenario, in which we first train a Single-Shot Multibox Detection (SSD) base model for OSM missing building detection in Kakola, Tanzania, where a previous humanitarian mapping project of OSM was organized to map all possible buildings. Then we extrapolate this base model using Few-Shot Transfer Learning (FSTL) to a set of areas in the proximity of the test area in Cameroon. Here, we develop a Geographical Weighted Model Ensemble (GWME) method to improve *Geographical Generalizability* of GeoAI models. Moreover, we compare four unsupervised model ensemble weighting strategies: 1) Average weighting, 2) Image similarity weighting, 3) Geographical distance weighting, and 4) Self-attention-based weighting. Experiments show promising results of the proposed GWME method, which implicitly generates model weights from their location embedding and image feature embedding in an unsupervised manner. More specifically, the self-attention-based model ensemble achieves the highest performance. The results shed inspiring light on improving the generalizability and replicability of GeoAI models across geographic areas. Data and code are available at https://github.com/tum-bgd/GWME.

CCS CONCEPTS

• **Information systems → Geographic information systems**; • **Computing methdologies → Artificial intelligence**.

KEYWORDS

GeoAI, OpenStreetMap, Vision Transformer, Model Ensemble, Humanitarian Mapping, Self-Attention

ACM Reference Format:
Hao Li, Jiapan Wang, Johann Maximilian Zollner, Gengchen Mai, Ni Lao, and Martin Werner. 2023. Rethink Geographical Generalizability with Unsupervised Self-Attention Model Ensemble: A Case Study of OpenStreetMap Missing Building Detection in Africa. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3589132.3625598

1 INTRODUCTION

As one of the fundamental principles of GIScience, **spatial heterogeneity** refers to the phenomenon that the expectation of a random variable varies across the Earth's surface [1]. In social and environmental science, we may observe spatial heterogeneity in both the relevant variables and confounding variables of the discovery process. Because of these phenomena, in many geospatial artificial intelligence (GeoAI) studies, one often encounters difficulties in replicating the results of the study to other areas that may or may not overlap with the original area without a significant performance decrease[16]. For example, a common scenario in GeoAI research is that a deep learning model pre-trained in a specific area may perform poorly in other geographic areas. Meanwhile, spatial autocorrelation, as stated in Tobler's First Law of Geography, may limit the applicability of knowledge learned from a certain training area only to their geographic proximity, which again hinders a seamless model transfer across space [15]. Herein, the ability of a GeoAI model to replicate or generalize the model's prediction ability across space is called **geographical generalizability** [30, 32], which is also coined as **replicability across space** [16].

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11…$15.00
https://doi.org/10.1145/3589132.3625598

Meanwhile, in the AI domain, the discussion about **the generalization and specialization** capability of AI models has a long history [2]. While it is desired that a model can effectively learn and solve a specific task, i.e., specializing a model to a certain extent, a model may also be over-specified, so-called overfitting. To avoid this, numerous efforts have been made toward better generalized models for example by improving the gradient descent optimization procedure [17], creating large-scale labeled datasets (e.g., ImageNet [7]), developing a more powerful AI architecture like the Transformer family [11, 43], or meta-learning techniques [37]. Although higher generalizability is preferred in most AI applications, spatial phenomena (e.g., spatial heterogeneity and autocorrelation) bring an additional scenario for AI applications in which one needs to find the balance between generalizability and specialization of GeoAI models, especially when aiming at large-scale applications across space. However, finding this "sweet spot" of geographical generalizability is a challenging task.

In the existing works, there are mainly three ways of tackling this geographic generalizability problem. The first common practice is to partition the space into different regions based on the underlining data process and learn separated models for different regions [50, 54]. The drawback of this practice is the large number of model parameters required for different geographic areas. The second common way to solve this is to apply transfer learning across space, such as urban-to-rural transfer [12], city-to-city transfer [44], country-to-country transfer [23], and so on. In a third stream, early attempts seek to integrate representation learning methods (e.g., location encoding) into a range of GeoAI applications, e.g., place recognition [52], trajectory prediction [51], point cloud segmentation [36], and geo-aware image classification [29, 34, 35], where spatial locations are represented (or encoded) into a high-dimensional embedding space to capture the spatial heterogeneity across space in order to facilitate downstream learning tasks. For a review of location encoding in GeoAI, please refer to [33]. More recently, the concept of position/location encoding has achieved superior performance in general AI tasks along with the popularity of Transformer and Vision Transformer models [5, 11, 43], where a **self-attention** mechanism is adopted to capture the relationships between the different elements (e.g., words, image patches, video blocks) of the same sequence/frame (e.g., sentence or image). Inspired by these observations, we ask whether we can simultaneously benefit from transfer learning and representation learning to improve the generalizability of GeoAI models across space.

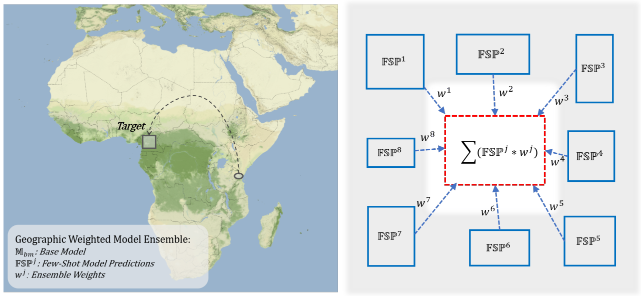

Figure 1: Overview of Geographic Weighted Model Ensemble (GWME). Left: an example target area (Cameroon) and a data-rich area (Tanzania), from which the base model $\mathbb{M}_{bm}$ is pre-trained. Right: a prediction in the target area is made by a weighted ensemble of predictions made by the base model few-shot transferred with examples from nearby reference areas.

In this paper, we rethink the geographical generalizability problem and propose to solve it by developing an unsupervised self-attention model ensemble method, namely Geographical Weighted Model Ensemble (GWME). An interesting case study of detecting OpenStreetMap (OSM) missing buildings across different counties in sub-Saharan Africa is conducted to demonstrate the effectiveness of GWME. The overall method is illustrated in Figure 1. Our general assumption is that although the target area is completely missing in OSM, it is still possible to find some nearby areas with a small number of training samples, which is useful to help improve the performance of the base model ($\mathbb{M}_{bm}$) trained in a target test area. We first train a base model $\mathbb{M}_{bm}$ on an OSM data-rich area A_{bm} (in Tanzania). And then we select a list of reference areas (A_j, where $j = 1, 2 \dots, T$) in the proximity of the test area (in Cameroon), which is geographically far away from Area A_{bm} and consists of very diverse landscapes and building structures. Next, we apply a Few-shot Transfer Learning technique [23] to extrapolate $\mathbb{M}_{bm}$ to those reference areas nearby the test area. After the FSTL, we now have a list of less accurate models close to the targeted test area, with which we seek to ensemble a more accurate model. Herein, a modern pre-trained ViT model with DINO (pre-trained on ImageNet) [5] is utilized to generate self-attention weights using an image patch ensemble method across different FSTL areas by implicitly considering their locations and image feature embeddings. Since self-attention can be calculated directly from a pre-trained model, the GWME is an unsupervised model ensemble method to improve the model's geographical generalizability for GeoAI applications.

We conduct intensive experiments on different weighting strategies with the proposed GWME method: 1) average weighting, 2) image similarity weighting, 3) geographical distance weighting, and 4) self-attention-based weighting. In this context, we elaborate on the potential of GWME as a promising avenue to improve the geographical generalizability of OSM missing building detection models and even beyond.

2 RELATED WORK

Herein, we discuss work related to the topic of replicating GeoAI models across space in the context of our case study with OSM missing building detection, which has recently gained increasing research interest among the broader geospatial community. The aim herein is to raise attention to a common problem of the model's geographical generalizability with an intuitive case study.

Humanitarian Mapping with OpenStreetMap - As an emerging spatial dataset, OpenStreetMap (OSM) has been extensively used to support humanitarian aid activities, especially in the Global South [18], where successful examples include the mapping tasks during the 2014 West Africa Ebola outbreak [10], the 2019 Cyclone Idai and Kenneth in Mozambique [22], and the 2023 Turkey Syria Earthquake [47]. Although the OSM data availability in the Global South has been greatly improved via recent humanitarian mapping campaigns, large rural areas still remain unmapped. Moreover, considering the time-crucial nature of disaster responses and humanitarian aid, existing OSM mapping workflows become less efficient and unsatisfactory in filling huge data gaps in OSM within a rather short time. Fortunately, the emergence of high-resolution satellite imagery allows for the augmentation and refinement of OSM data with GeoAI techniques [38, 46], thus providing a promising solution to address this challenge that humanitarian organizations currently encounter. Early works in this direction [19, 20] report an interesting finding on improving the speed and accuracy of humanitarian mappings via a machine-assisted manner. However, a majority of existing approaches rely on models trained in OSM data-rich areas, which can be only applied to limited nearby areas. Therefore, one key challenge for large-scale machine-assisted mapping in OSM is how the GeoAI models can be effectively replicated in remote areas with little or no OSM data.

Geospatial Object Detection - Recently, the ever-increasing availability of multimodal Earth Observation (EO) data, including Very High Resolution (VHR) images, Multispectral (MSI) imaging, and Hyperspectral (HSI) imaging, offers a promising data source for modern GeoAI models to automatic detection and map geographical objects, ranging from common objects like buildings to those special objects such as planes or ships. Successful examples of building detection include but are not limited to Global Urban Footprint (GUF) [13] from German Aerospace Center, High Resolution Settlement Layer (HRSL) [42] from the Connectivity Lab at Meta, and the Google Open Building Layer (GOB) [39]. As another example, researchers detected over 1.8 billion individual trees in the West African Sahara, Sahel, and sub-humid areas from VHR satellite imagery (i.e., sub-meter resolution) using a deep learning method [3]. Although such GeoAI models offer an unprecedented ability to monitor and map geospatial objects at scale, the lack of large-scale training data has become a major bottleneck for the advancement of geospatial object detection [9, 24, 48]. To mitigate this effect, considerable effort has been dedicated to creating benchmark datasets for multi-class geospatial object detection, including NWPU VHR-10 [6], DOTA [9], and FAIR1M [41]. Meanwhile, there is an increasing interest in how one can leverage the knowledge of existing GeoAI models pre-trained in different geographic regions to achieve consistent geospatial object detection performance without much additional training data.

Transfer Learning and Spatial Explicit AI - Modern advances in GeoAI models and algorithms often benefit from two streams of learning techniques. First, the concept of transfer learning has been well-integrated into almost all state-of-the-art AI architectures. Specifically, it is a common practice to pre-train vision models on large-scale image datasets, like Microsoft COCO dataset [26], ImageNet dataset [8], and PASCAL VOC dataset [14], which can be fine-tuned in downstream tasks to achieve better generalization performance. For example, Wang et al. [45] confirmed the effectiveness of fine-tuning-based approaches on few-shot object detection with a systematic analysis of the state-of-the-art benchmarks. Inspired by this, Li et al. [23] proposed a model-agnostic Few-Shot Transfer Learning (FSTL) method to improve the performance of OSM missing building detection in Sub-Saharan Africa. Besides transfer learning, a second stream of research goes to the representation learning direction, where a location encoder is learned to improve GeoAI model performance while preserving spatial information (e.g., distance and direction) after the encoding process. Existing works include approaches to learn a location representation from image-location pairs for geo-localization [53] and to apply a dual-encoder for the geo-aware image classification task [34, 35]. Herein, an interesting topic is the so-called spatial explicit AI [21, 27, 31], where the design of GeoAI models explicitly considers a range of spatial concepts, spatial principles, and spatial inductive biases. Previous works incorporate several important spatial principles, such as spatial dependency [25], spatial heterogeneity [16, 29], and temporal periodicity [4].

Inspired by existing research, we rethink the geographical generalizability of GeoAI (e.g., geospatial building detection) models by considering both transfer learning and representation learning with an unsupervised self-attention model ensemble method. Moreover, we use the task of OSM missing building detection in sub-Saharan Africa as a case study, where GeoAI methods have shown great potential in supporting humanitarian mapping activities at scale.

3 DEFINITION AND PRELIMINARIES

We assume an OSM-labeled training dataset for the base model (BM) to be a triplet $\mathbb{S}_{bm} = \{(\mathbf{x}_i, \mathbf{I}_i, \mathbf{y}_i)\}$ with $i = 1, \dots, N$ in OSM data-rich area A_{bm}. Here, $\mathbf{x}_i$ is a satellite image, $\mathbf{y}_i$ is a set of object bounding boxes (bbox) in this image, and $\mathbf{I}_i$ refers to the location (e.g., longitude and latitude) and optionally geographic distances to the hold-out test dataset $\mathbb{S}_{test} = \{(\mathbf{x}_i, \mathbf{I}_i)\}$ with $i = 1, \dots, M$. Then, the BM is supervised pre-trained by optimizing the loss function $\mathcal{L}$ via gradient descent, which is described as $\mathcal{F}(\mathbb{S}_{bm}) \to \mathbb{M}_{bm}$, and where $\mathbb{M}_{bm}$ represents the BM for the GeoAI task (e.g., a building detection model trained in Tanzania). From here, we identify a list of FSTL datasets $\mathbb{S}^j_{fs} = \{(\mathbf{x}^j_i, \mathbf{I}^j_i, \mathbf{y}^j_i)\}$ with $i = 1, \dots, n$ and $j = 1, \dots, T$ from T reference areas A_j ($j = 1, \dots, T$) in the proximity of the target test area A_{test}. The FSTL can be then formulated as a similar function $\mathcal{F}(\mathbb{S}^j_{fs}) \to \mathbb{M}^j_{fs}$, which gives us a set of FSTL models $\mathbb{M}^j_{fs}$ from distinct nearby areas A_j. Before we apply the GWME method, we first conduct model forward inference, represented by $\mathcal{P}$, in the test dataset $\mathbb{S}_{test}$ about the test area A_{test} with different FSTL models $\mathbb{M}^j_{fs}$ as follows:

$$\mathbb{FSP}^j = \bigcup_{(\mathbf{x}_i, \mathbf{I}_i) \in \mathbb{S}_{test}} \mathcal{P}(\mathbb{M}^j_{fs}, (\mathbf{x}_i, \mathbf{I}_i)). \tag{1}$$

Where $\{\mathbb{FSP}^j\}^T_{j=1}$ refers to the corresponding set of predictions of T models in the test area $\mathbb{S}_{test}$, where $\{\mathbb{M}^j_{fs}\}^T_{j=1}$ are a set of FSTL models. For each FSTL model, the GWME method then jointly considers their vision representation ($\mathbf{x}^j_i$) and geographic locations ($\mathbf{I}^j_i$) with **an explicit weighting function $\mathcal{W} \in \mathbb{R}^{T \times M}$ in a tile-based manner**. Following, this generates a list of corresponding weights $\mathbf{w}^j$ which can be used in the final step of GWME, resulting in predictions ($\mathbb{P}$) and denoted as:

$$\mathbb{P} = \sum_{j=0}^{T} Q(\mathbb{FSP}^j, \mathbf{w}^j, \mathbf{th}). \tag{2}$$

Here, Q is a **weighted boxes fusion** [40] function and **th** represents the corresponding hyperparameters, such as the threshold of

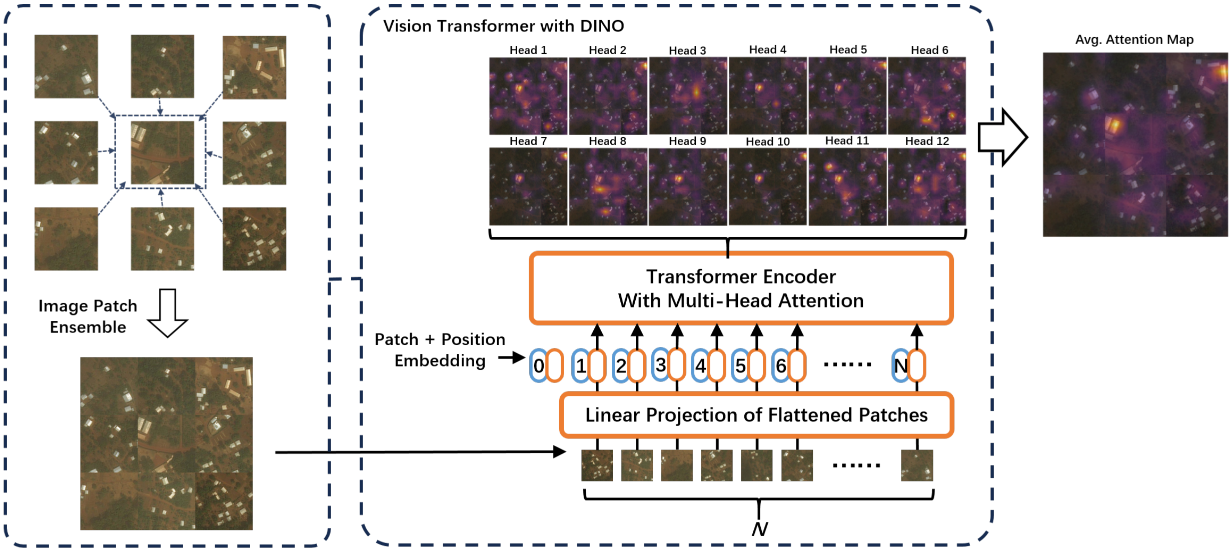
Image Patch Ensemble
Vision Transformer with DINO
Head 1
Head 2
Head 3
Head 4
Head 5
Head 6
Head 7
Head 8
Head 9
Head 10
Head 11
Head 12
Avg. Attention Map
Transformer Encoder With Multi-Head Attention
Patch + Position Embedding
0
1
2
3
4
5
6
......
N
Linear Projection of Flattened Patches
......
N

Figure 2: The extraction of self-attention-based weights for the GWME using a pre-trained ViT with DINO.

confidential scores. The general idea of the weighted boxes fusion is visualized in Figure 3.

Therefore, the objective of measuring the model's **geographical generalizability** is achieved by minimizing the discrepancy between the GWME predictions ($\mathbb{P}$) and the ground truth label ($\mathbb{Y}_{test}$) of the test dataset $\mathbb{S}_{test}$. A more detailed description of the GWME method and how it is implemented is given in Section 4.

4 METHOD

In this section, we present the proposed GWME method, which aims to tackle the problem of the model's **geographical generalizability**, specifically with the OSM missing building detection task.

4.1 Multiple Few Shot Transfer Learning

Given the base model of OSM missing building detection $\mathbb{M}_{bm}$, we first implement the FSTL method [23] in order to learn a set of similar models in the geographic proximity of the target test area A_{test} (as shown in Figure 1), where OSM building features are entirely missing. With the so-called reference areas A_j, the FSTL method results in a set of FSTL models $\{\mathbb{M}^j_{fs}\}^T_{j=1}$ as well as the corresponding predictions of OSM missing buildings $\{\mathbb{FSP}^j\}^T_{j=1}$. Herein, the assumption is that although the target area is completely missing in OSM, it is still possible to find some nearby areas with a small amount of "training shots". They can be used to improve the performance of the base model ($\mathbb{M}_{bm}$) trained in a geographically-remote area. Obviously, the performance of FSTL models is still limited by factors, such as the amount of few-shot samples and their distance to the target area, which are essential to **geographical generalizability**. In this context, we seek to take a step towards combining these FSTL predictions by establishing an effective weighting strategy to do model ensembles with these limitation factors in mind.

In the case study, we choose a commonly-used single-stage object detector, specifically the Single Shot Multibox Detector (SSD) [28], and deploy their pre-trained parameters from Microsoft COCO dataset [26] as the backbone of all our OSM missing building detection models. After the base model training, we follow the multiple FSTL and predictions algorithm (as shown in Algorithm 1) to extrapolate $\mathbb{M}_{bm}$ to many less-accurate models $\{\mathbb{M}^j_{fs}\}^T_{j=1}$. For more details about the FSTL approach, one can refer to [23].

Algorithm 1 Multiple Few-shot Transfer Learning and Predictions

```
Input:
$\mathbb{M}_{bm}$: the base model;
T: number of reference areas neighbouring the test area;
$\mathbb{S}^j_{fs} = \{(\mathbf{x}^j_i, \mathbf{I}^j_i, \mathbf{y}^j_i)\}$ with $i = 1, \ldots, n_j$ and $j = 1, \ldots, T$: FSTL samples from reference areas;
$\mathbb{M}^j_{fs} \leftarrow \{\}$: few-shot models fine-tuned on reference areas;
$\mathbb{S}_{test} = \{(\mathbf{x}_i, \mathbf{I}_i)\}$ with $i = 1, \ldots, M$: dataset from the test area;
$\mathbb{FSP}^j, j = 1, \ldots, T \leftarrow []$; predictions from single FSTL models
for dataset $\mathbb{S}^j_{fs}$ of each reference area $A_j$ in $\{\mathbb{S}^j_{fs}\}^T_{j=1}$ do
    few-shot model $\mathbb{M}^j_{fs} \leftarrow \mathcal{F}(\mathbb{S}^j_{fs}, \theta)$;
    for each $(\mathbf{x}_i, \mathbf{I}_i)$ in $\mathbb{S}_{test} = \{(\mathbf{x}_i, \mathbf{I}_i)\}^M_{i=1}$ do
        update $\mathbb{FSP}^j_i \leftarrow \mathcal{P}(\mathbb{M}^j_{fs}, (\mathbf{x}_i, \mathbf{I}_i))$
    end for
end for
Output:
$\mathbb{FSP}^j$: list of objects and scores predicted from reference few-shot models;
```

4.2 Model Ensemble and Weighting Strategy

Now, we want to use GWME to ensemble T few-shot learned models $\{\mathbb{M}^j_{fs}\}^T_{j=1}$ into the target area. The most important step is to decide the weights of each individual FSTL model ($\mathbb{M}_{fs}^{j}$) during model ensemble. Inspired by the lessons learned in spatial explicit AI research, we develop an unsupervised method to learn model ensemble weights by considering both image feature embedding and location embedding with a self-attention mechanism. We call it **self-attention-based weighting** (as shown in Figure2) and compare it with three other weighting strategies. We elaborate on these different weighting strategies as follows:

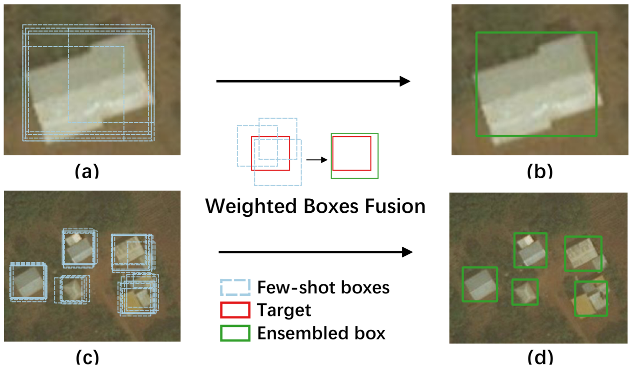

Figure 3: Examples of the weighted boxes fusion. (a) and (c) multiple predicted boxes from different FSTL models $\{\mathbb{M}_{fs}^{j}\}_{j=1}^{T}$; (b) and (d) the ensembled boxes.

Average Weighting (*average*)- In the simplest case, we consider equal weights when combining the FSTL models. After selecting the weights, we apply the weighted boxes fusion method [40] to conduct a prediction-level ensemble of all $\mathbb{FSP}_{fs}^{j}$ as a proof-of-concept. At the same time, future works are encouraged to explore model-level weighted model ensembles.

$$\{\mathbf{w}^{j} = 1\} \longrightarrow \mathbb{FSP}^{j} \tag{3}$$

Image Similarity Weighting (*similarity*) - As a second weighting strategy, it is intuitive to think about considering the similarity of satellite images among the FSTL areas and the test area. Therefore, we consider an average cosine similarity between the histograms of satellite image pairs, namely $\mathbf{x}_{i}^{j} \in \mathbb{S}_{fs}^{j}$ and $\mathbf{x}_{i} \in \mathbb{S}_{test}$, as a proxy of their image similarity weights (see Equation 4). Herein, $\cos(\cdot)$ indicates the cosine similarity function. $HIS(\cdot)$ indicates a function to compute the image histogram. n_j is the number of few-shot data samples we used in $\mathbb{S}_{fs}^{j}$ in the jth FSTL areas A_j.

$$\{\mathbf{w}_{i}^{j} = \frac{1}{n_j} \sum_{(\mathbf{x}_{i}^{j}, \mathbf{I}_{i}^{j}, \mathbf{y}_{i}^{j}) \in \mathbb{S}_{fs}^{j}} \cos(HIS(\mathbf{x}_{i}^{j}), HIS(\mathbf{x}_{i})\} \longrightarrow \mathbb{FSP}^{j} \tag{4}$$

Geographical Distance Weighting (*distance*) - Given Tobler's First Law of Geography, we expect a similar object to be observed if they are close to each other. To this end, we consider a reverse distance weighting strategy for model ensemble, which is based on the prior knowledge of their geographical locations, specifically $\mathbf{I}_{i} \in \mathbb{S}_{test}$ and the center of each $\mathbb{S}_{fs}^{j}$. Equation 5 illustrates the general idea where $CEN(\cdot)$ indicates the geometric center of the study area and DIS indicates a distance function, e.g., Euclidean distance, great circle distance, geodesic distance, and so on.

$$\{\mathbf{w}_{i}^{j} = DIS(\mathbf{I}_{i}, CEN(\mathbb{S}_{fs}^{j}))\} \rightarrow \mathbb{FSP}^{j} \tag{5}$$

Algorithm 2 Geographical Weighted Model Ensemble (GWME)

```
Input:
$\mathbf{M}_{ViT}$: the pre-trained ViT model;
$\mathbb{S}_{fs}^{j} = \{(\mathbf{x}_{i}^{j}, \mathbf{I}_{i}^{j}, \mathbf{y}_{i}^{j})\}$ with $i = 1, \ldots, n_j$ and $j = 1, \ldots, T$: FSTL samples from reference areas;
$\mathbb{S}_{test} = \{(\mathbf{x}_{i}, \mathbf{I}_{i})\}$ with $i = 1, \ldots, M$: dataset from the test area;
$\mathbb{FSP}^{j}, j = 1, \ldots, T$: list of objects and scores predicted from different models;
th: threshold of prediction score
$\mathbb{P}$: ensembled objects and scores;
Mode: weighting mode;
$\mathbf{w}^{j}$: corresponding weights for $\mathbb{M}_{fs}^{j}$.
Weights $\mathcal{W} \leftarrow []$;
for each $(\mathbf{x}_{i}, \mathbf{I}_{i})$ in $\mathbb{S}_{test} = \{(\mathbf{x}_{i}, \mathbf{I}_{i})\}_{i=1}^{M}$ do
    for dataset $\mathbb{S}_{fs}^{j}$ of each reference area $A_j$ in $\{\mathbb{S}_{fs}^{j}\}_{j=1}^{T}$ do
        if Mode == "average" then
            average weights $\mathbf{w}_{i}^{j} = 1$;
        else if Mode == "similarity" then
            $\mathbf{w}_{i}^{j} = \frac{1}{n_j} \sum_{(\mathbf{x}_{i}^{j}, \mathbf{I}_{i}^{j}, \mathbf{y}_{i}^{j}) \in \mathbb{S}_{fs}^{j}} \cos(HIS(\mathbf{x}_{i}), HIS(\mathbf{x}^{j}))$;
        else if Mode == "distance" then
            $\mathbf{w}_{i}^{j} = DIS(\mathbf{I}_{i}, CEN(\mathbb{S}_{fs}^{j}))$;
        else if Mode == "attention" then
            image patches patch_list[] $\leftarrow \mathbf{x}_{i}$;
            patch_list.$append\_patch(\mathbf{x}_{i}^{j}, \mathbf{I}_{i}^{j})$;
            multi_heads_attentions = $\mathbf{M}_{ViT}$(patch_list);
            attention_map = $attention$(multi_heads_attentions);
            $\mathbf{w}_{i}^{j} = subset$(attention_map);
        end if
        $\mathbf{w}_{i} \leftarrow \mathbf{w}_{i}^{j}$
        prediction candidates $\mathbb{FSP}_{i} \leftarrow \mathbb{FSP}_{i}^{j}$;
    end for
    $\mathcal{W} \leftarrow normalize(\mathbf{w}_{i})$;
    update $\mathbb{P}_{i} = Q(\mathbb{FSP}_{i}, \mathbf{w}_{i},)$;
end for
Output:
$\mathbb{P}$: ensembled results and scores;
```

Self-Attention-Based Weighting (*attention*) - As the most interesting part, we develop an unsupervised method to learn self-attention-based weight from a pre-trained ViT model – the Self-Supervised ViT with DINO [5]. This model is pre-trained on ImageNet. To adopt DINO into GWME, we design an image patch ensemble approach according to their relative positions, where the central image patch is taken from the test area and the context image patches are taken from T few-shot reference areas A_j ($j = 1, \ldots, T$) as shown in Figure 1. In other words, unlike the original ViT which splits one single image into different patches, we pick different image patches from different reference or target areas to form a big image (see Figure 2) and use relative position embedding to capture their relative spatial relations.

$$\{\mathbf{w}_i^j = subset(attention\_map)\} \rightarrow \mathbb{FSP}^j \quad (6)$$

This approach is inspired by the relative position embedding idea in the origin design of ViT. The advantage of this approach is twofold: first, the self-attention-based weighting can simultaneously consider the location (via position embedding) and image feature embedding (via patch image embeddings) for weighting; second, the extraction of self-attention relies only on pre-trained ViT and satellite image patches without any prior knowledge (e.g., geographical location, image source). In order to leverage the pre-trained ViT weighted, we adopt the relative position encoding method used by ViT and DINO [5]. Future work can extend this method by explicitly encoding the patch's geographic locations as the patch position embeddings and fine-tuning the ViT model.

4.3 Put it All Together

To put everything together, the pseudocode of the complete GWME is presented in Algorithm 2, where we start from multiple FSTL model predictions as well as those FSTL datasets $\mathbb{S}_{fs}^j = \{(\mathbf{x}_i^j, \mathbf{I}_i^j, \mathbf{y}_i^j)\}$ with $i = 1, \ldots, n$, and end with the ensemble predictions together with their confidential scores for OSM missing building detection in the target test area.

To evaluate the proposed GWME method, we conduct intensive experiments with our cast study of cross-country OSM missing building detection in Sub-Saharan Africa by comparing the performance of 1) the base model $\mathbb{M}_{bm}$ and eight distinct FSTL models $\{\mathbb{M}_{fs}^j\}_{j=1}^T$ and 2) four different weighting strategies with GWME.

In the following Section 5, we present the experimental results together with the findings and insights in the direction of rethinking the geographical generalizability of GeoAI models in a broader scope.

5 EXPERIMENT RESULT

In this section, we examine the effectiveness of GWME in detecting OSM missing buildings across different countries in sub-Saharan Africa. In this context, we are particularly interested in how different weighting strategies perform during the model ensemble.

5.1 Dataset and Evaluation Metrics

For our cast study, we take the dataset collected in [23], where a well-mapped area in Tanzania is used to train the base model ($\mathbb{M}_{bm}$) and a geographically remote area in Cameroon is selected as the test area A_{test} who does not have any training samples. We identify eight reference areas A_j with few-shot training samples for the FSTL purpose around our test area A_{test} (as shown in Figure 1). More specifically, OSM buildings within the training area A_{bm} in Tanzania were fully mapped during a humanitarian mapping campaign organized by Humanitarian OpenStreetMap Team (HOT). For the test area A_{test}, since it is completely missing in OSM, we have organized an expert mapping campaign and digitized in total **1,811 buildings within an 8.57**km^2 **area** in Cameroon as the reference data. Table 1 gives the statistic of all datasets (i.e., train in Tanzania, test in Cameroon, and FSTL areas in Cameroon). Moreover, We provide the data and code used in this paper openly available in https://github.com/tum-bgd/GWME.

Table 1: Summary statistic of the datasets.

Counts	A_{bm}	A_{test}	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
Buildings	6,272	1,811	66	45	116	46	71	61	40	79
Areas (km^2)	232.50	8.57	0.35	0.16	0.35	0.22	0.34	0.44	0.25	0.20
Tiles n_j	1,744	343	5	5	9	7	9	9	7	7

To generate the training data, we use the ohsome2label package [49] to combine OSM building geometries with Bing satellite imagery at a zoom level of 18 (i.e., a spatial resolution of 0.6m), then convert them to training datasets for the TensorFlow Object Detection API [1]. For the SSD [28], the pre-trained parameters are downloaded from the TensorFlow Detection Model Zoo. The training process for the base model in Tanzania was run for 50,000 epochs, with an initial learning rate of 0.0004. The FSTL fine-tuning epochs were then set to 10,000 for all reference areas. The algorithms were implemented using Python 3.10, TensorFlow 2.2, and TensorFlow object detection API on a Linux server with a GeForce RTX 3080Ti graphical processing unit (GPU) of 12 GB memory.

For evaluation, we consider common metrics for a single-class object detection task, such as Precision, Recall, Accuracy, and F1-score. Specifically, we use a default IoU threshold of 0.5 as the criteria to decide whether a prediction bounding box refers to a building bounding box in the reference data, which then distinguishes all predictions into False Negatives (FN), False Positives (FP), and True Positives (TP). There is no True Negative (TN), as we are not interested in detecting non-building objects.

Table 2: Evaluation metrics of predictions from the base model and single FSTL models on the test dataset. BM and $\mathbb{FSP}^j$ indicate the model predictions of the base model $\mathbb{M}_{bm}$ as well as different FSTL models $\{\mathbb{M}_{fs}^j\}_{j=1}^T$.

Predictions	Precision (%)	Accuracy (%)	Recall (%)	F1
BM	97.66	13.71	13.75	0.2411
$\mathbb{FSP}^1$	99.00	60.90	61.27	0.7570
$\mathbb{FSP}^2$	96.94	**68.93**	**70.46**	**0.8160**
$\mathbb{FSP}^3$	98.18	53.06	53.58	0.6933
$\mathbb{FSP}^4$	98.22	49.06	49.50	0.6582
$\mathbb{FSP}^5$	98.44	61.27	61.87	0.7598
$\mathbb{FSP}^6$	84.65	40.90	44.18	0.5806
$\mathbb{FSP}^7$	**99.12**	52.66	52.91	0.6899
$\mathbb{FSP}^8$	98.73	52.60	52.96	0.6894
Mean($\mathbb{FSP}^j$)	96.66	54.92	55.84	0.7055

5.2 Geographic Weighted Model Ensemble

In Table 2, we evaluate to which extent our FSTL models $\{\mathbb{M}_{fs}^j\}_{j=1}^T$ can improve the performance of the base model $\mathbb{M}_{bm}$ on OSM missing building detection task in the test dataset $\mathbb{S}_{test}$ with a

[1] https://github.com/tensorflow/models/tree/master/research/object_detection

limited amount of **geographically nearby training samples** ("few-shot samples").

Herein, we found an overall significant accuracy improvement of FSTL models over the base model, with so call $Mean(\mathbb{FSP}^j)$, over the base model. However, an interesting observation is that the individual model performance varies a lot, where $\mathbb{FSP}^4$ leads to the biggest improvement and $\mathbb{FSP}^6$ the lowest. Such a distinct behavior implies the different levels of geographical generalizability among a set of FSTL models.

Table 3 compares the performance of four different weighting strategies with the proposed GWME method, where a threshold of confidential scores (**th**) is set to ones with $precision \geq 95\%$ as shown in Figure 4. Several key findings can be observed. First, even with average weighting, the model ensemble leads to a significant improvement over all performances of single FSTL models, which proves the effectiveness of GWME compared to the baseline model. Second, although we assume that image similarities play a role, image similarity weighting *similarity* ends up with the least improvement in the model ensemble, while the simple reverse distance weighting gives a surprisingly nice result (72,98% Accuracy and 0.84 F1). Last but foremost, we observe the biggest performance improvement via the GWME method with self-attention-based weighting *attention*, which leads to more than 6% improvement in overall accuracy and the highest Recall of 78.99% in the test area A_{test}. In addition, the GWME method outperforms the FSTL results presented in [23] based on a similar reference dataset.

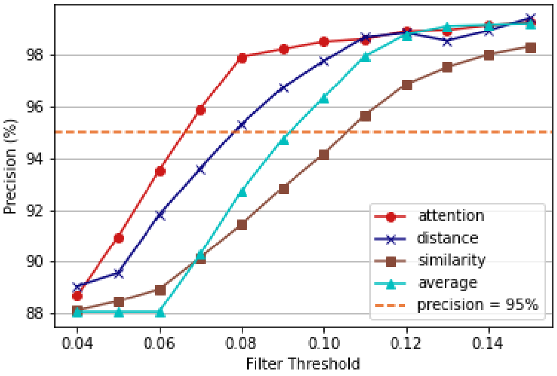

Figure 4: The curve of Precision and the threshold of confidential scores in the GWME using different weighting strategies.

In this context, we summarize the advantages of self-attention-based weighting as twofold: 1) the self-attention map generated from the pre-trained ViT with DINO can capture the general relative importance of model predictions across different image patches; 2) the position embedding in ViT can be equivalent (or even better than) to the geographical reverse distance weighting while requiring no prior knowledge (e.g., latitude and longitude) but only relative positions. The unsupervised nature of this self-attention-based weighting *attention* makes it a promising solution to the problem of geographical generalizability, even beyond this case study of building an ensemble of multiple object detection models, which in principle, can be replaced with other GeoAI models.

Table 3: Evaluation metrics of predictions from ensembled results by different weighting modes.

GWME Weightings	Precision (%)	Accuracy (%)	Recall (%)	F1
average	96.35	71.70	73.70	0.8352
similarity	95.68	71.16	73.52	0.8315
distance	**97.76**	72.98	74.22	0.8438
attention	96.95	**77.07**	**78.99**	**0.8705**

5.3 Visual Interpretation

Our GWME method can effectively improve the geographical generalizability of GeoAI models in an unsupervised manner. In Figure 5, we compare evaluation metrics between the baseline method (e.g., the base model $\mathbb{M}_{bm}$ and the best single FSTL model $\mathbb{M}^4_{fs}$) and GWME results with different weighting strategies.

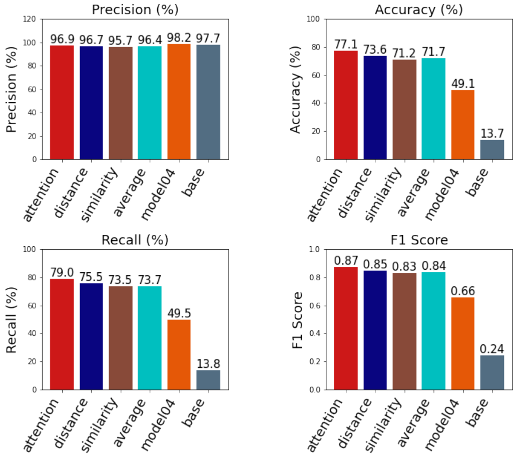

Figure 5: Performance of GWME predictions (precision > 95%) using different weighting strategies.

To visually interpret the advantages of our GWME, Figure 6 compares the OSM missing building detection results of three different models: the base model $\mathbb{M}_{bm}$, the single FSTL model $\mathbb{M}^4_{fs}$, and the results from our GWME with self-attention-based weights *attention*. Comparing Figure 6 (b) and (c) with Figure 6 (a), we can see a significant decrease in FN. Such a decrease in FN, originating from valid buildings that are overlooked by a model trained in geographically remote areas, confirms our assumption that few-shot learning is very effective to improve model performance in geographically remote area. Comparing Figure 6 (c) with (b), we see that our GWME with *attention* further reduces the FN and FP. This confirms the effectiveness of GWME in achieving better geographical generalizability for GeoAI models.

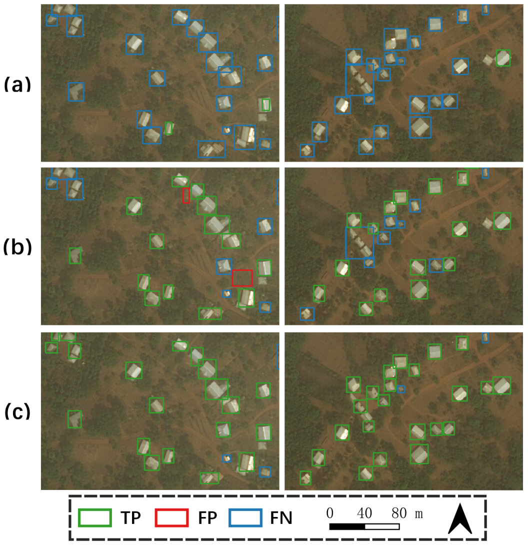

Figure 6: The comparison map of prediction results. (a) the base model ($\mathbb{M}_{bm}$); (b) the single FSTL model (i.e., $\mathbb{M}^4_{fs}$); (c) the GWME result with self-attention-based weights $attention$.

In future work, we aim to extend this method to multiple countries across Africa, even other continents, to support better and faster humanitarian mapping with OSM by establishing a robust and effective indicator of global OSM building completeness.

6 CONCLUSION

In this work, we proposed Geographical Weighted Model Ensemble (GWME), an unsupervised model ensemble method to improve the geographical generalizability of GeoAI models, with a case study of cross-country OpenStreetMap (OSM) missing building detection in sub-Saharan Africa. Based on existing methods of transferring GeoAI models across geographical space, such as FSTL [23], we develop a self-attention-base weighting for the GWME by simultaneously considering the location and image feature embedding for weighting different FSTL models. More importantly, self-attention can be intuitively learned from a pre-trained ViT model with DINO without prior knowledge (e.g., geographical locations) in a fully unsupervised manner. For comparison, we consider three other weighting strategies: 1) average weighting, 2) image similarity weighting, and 3) geographical distance weighting. To evaluate the effectiveness of GWME, we conduct intensive experiments with a cast study of OSM missing building detection, where the base model is trained in Tanzania, and the test area is in Cameroon. Experimental results confirmed the capability of GWME with the self-attention-based weighting which can outperform both the base model and single FSTL model with more than 6% accuracy improvement over the best single FSTL model.

Despite the promising results, we identify several limitations and future work directions with this case study. For instance, we use a prediction-level model ensemble as a proof-of-concept, while a parameter-level model ensemble can be preferred by considering the computational efficiency. Also, our reference areas are still in proximity to the test area. Thus, our future work will extend this into a larger scale study area for more aggressive improvement of geographical generalizability. Moreover, we now use the default position embedding from the pre-trained ViT model, but it would be interesting to integrate spatially explicit location embedding into the proposed method [34].

In short, the proposed GWME sheds inspiring light on the general topic of rethinking the geographical generalizability of GeoAI models, with an unsupervised self-attention model ensemble method. To this end, this work is a step towards developing replicable and spatially explicit models for geospatial artificial intelligence.

REFERENCES

[1] Luc Anselin. 1989. What is special about spatial data? Alternative perspectives on spatial data analysis (89-4). (1989).

[2] Olivier Bousquet and André Elisseeff. 2002. Stability and generalization. *The Journal of Machine Learning Research* 2 (2002), 499–526.

[3] Martin Brandt, Compton J Tucker, Ankit Kariryaa, Kjeld Rasmussen, Christin Abel, Jennifer Small, Jerome Chave, Laura Vang Rasmussen, Pierre Hiernaux, Abdoul Aziz Diouf, et al. 2020. An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature* 587, 7832 (2020), 78–82.

[4] Ling Cai, Krzysztof Janowicz, Gengchen Mai, Bo Yan, and Rui Zhu. 2020. Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting. *Transactions in GIS* 24, 3 (2020), 736–755.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[6] Gong Cheng, Peicheng Zhou, and Junwei Han. 2016. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 54, 12 (2016), 7405–7415. https://doi.org/10.1109/TGRS.2016.2601622

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[9] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Micheal Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. 2021. Object detection in aerial images: A large-scale benchmark and challenges. *arXiv preprint arXiv:2102.12219* (2021).

[10] Martin Dittus, Giovanni Quattrone, and Licia Capra. 2016. Analysing volunteer engagement in humanitarian mapping: building contributor communities at large scale. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 108–118.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[12] Amna Elmustafa, Erik Rozi, Yutong He, Gengchen Mai, Stefano Ermon, Marshall Burke, and David Lobell. 2022. Understanding economic development in rural Africa using satellite imagery, building footprints and deep models. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–4.

[13] T. Esch, M. Marconcini, A. Felbier, A. Roth, W. Heldens, M. Huber, M. Schwinger, H. Taubenböck, A. Müller, and S. Dech. 2013. Urban Footprint Processor—Fully Automated Processing Chain Generating Settlement Masks From Global Data of the TanDEM-X Mission. *IEEE Geoscience and Remote Sensing Letters* 10, 6 (Nov 2013), 1617–1621.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (jun 2010), 303–338.

[15] Michael F Goodchild. 2004. The validity and usefulness of laws in geographic information science and geography. *Annals of the Association of American Geographers* 94, 2 (2004), 300–303.

[16] Michael F Goodchild and Wenwen Li. 2021. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences* 118, 35 (2021), e2015759118.

[17] Moritz Hardt, Ben Recht, and Yoram Singer. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*. PMLR, 1225–1234.

[18] Benjamin Herfort, Sven Lautenbach, João Porto de Albuquerque, Jennings Anderson, and Alexander Zipf. 2021. The evolution of humanitarian mapping within

the OpenStreetMap community. *Scientific reports* 11, 1 (2021), 1–15.

[19] Benjamin Herfort, Hao Li, Sascha Fendrich, Sven Lautenbach, and Alexander Zipf. 2019. Mapping Human Settlements with Higher Accuracy and Less Volunteer Efforts by Combining Crowdsourcing and Deep Learning. *Remote Sensing* 11, 15 (2019).

[20] Jonathan J Huck, Chris Perkins, Billy T Haworth, Emmanuel B Moro, and Mahesh Nirmalan. 2021. Centaur VGI: A Hybrid Human–Machine Approach to Address Global Inequalities in Map Coverage. *Annals of the American Association of Geographers* 111, 1 (2021), 231–251.

[21] Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. 2020. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science* 34, 4 (2020), 625–636.

[22] Hao Li, Benjamin Herfort, Wei Huang, Mohammed Zia, and Alexander Zipf. 2020. Exploration of OpenStreetMap missing built-up areas using twitter hierarchical clustering and deep learning in Mozambique. *ISPRS Journal of Photogrammetry and Remote Sensing* 166 (2020), 41–51. https://doi.org/10.1016/j.isprsjprs.2020.05.007

[23] Hao Li, Benjamin Herfort, Sven Lautenbach, Jiaoyan Chen, and Alexander Zipf. 2022. Improving OpenStreetMap missing building detection using few-shot transfer learning in sub-Saharan Africa. *Transactions in GIS* 26, 8 (2022), 3125–3146.

[24] Hao Li, Johannes Zech, Danfeng Hong, Pedram Ghamisi, Michael Schultz, and Alexander Zipf. 2022. Leveraging openstreetmap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection. *International Journal of Applied Earth Observation and Geoinformation* 110 (2022), 102804.

[25] Wenwen Li, Chia-Yu Hsu, and Maosheng Hu. 2021. Tobler's First Law in GeoAI: A spatially explicit deep learning model for terrain feature detection under weak supervision. *Annals of the American Association of Geographers* 111, 7 (2021), 1887–1905.

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014).

[27] Pengyuan Liu and Filip Biljecki. 2022. A review of spatially-explicit GeoAI applications in Urban Geography. *International Journal of Applied Earth Observation and Geoinformation* 112 (2022), 102936.

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[29] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. 2019. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9596–9606.

[30] Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. 2022. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–4.

[31] Gengchen Mai, Yingjie Hu, Song Gao, Ling Cai, Bruno Martins, Johannes Scholz, Jing Gao, and Krzysztof Janowicz. 2022. Symbolic and subsymbolic GeoAI: Geospatial knowledge graphs and spatially explicit machine learning. *Trans GIS* 26, 8 (2022), 3118–3124.

[32] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. 2023. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798* (2023).

[33] Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, Song Gao, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. 2022. A review of location encoding for GeoAI: methods and applications. *International Journal of Geographical Information Science* 36, 4 (2022), 639–673.

[34] Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. 2020. Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells. In *The Eighth International Conference on Learning Representations*. openreview.

[35] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. 2023. CSP: Self-Supervised Contrastive Spatial Pre-Training for Geospatial-Visual Representations. In *International Conference on Machine Learning*. PMLR.

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* 30 (2017).

[37] Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. 2020. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 200–201.

[38] Sancho Salcedo-Sanz, Pedram Ghamisi, María Piles, Martin Werner, Lucas Cuadra, A Moreno-Martínez, Emma Izquierdo-Verdiguier, Jordi Muñoz-Marí, Amirhosein Mosavi, and Gustau Camps-Valls. 2020. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion* 63 (2020), 256–272.

[39] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Edine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. 2021. Continental-Scale Building Detection from High Resolution Satellite Imagery. *arXiv preprint arXiv:2107.12283* (2021).

[40] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing* (2021), 1–6.

[41] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 184 (2022), 116–130. https://doi.org/10.1016/j.isprsjprs.2021.12.004

[42] Tobias G. Tiecke, Xianming Liu, Amy Zhang, Andreas Gros, Nan Li, Gregory Yetman, Talip Kilic, Siobhan Murray, Brian Blankespoor, Espen B. Prydz, and Hai-Anh H. Dang. 2017. Mapping the world population one building at a time. *CoRR* abs/1712.05839 (2017). arXiv:1712.05839

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[44] Leye Wang, Xu Geng, Xiaojuan Ma, Feng Liu, and Qiang Yang. 2018. Cross-City Transfer Learning for Deep Spatio-Temporal Prediction. arXiv:1802.00386 [cs.AI]

[45] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. 2020. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957* (2020).

[46] Martin Werner and Hao Li. 2022. AtlasHDF: an efficient big data framework for GeoAI. In *Proceedings of the 10th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. 1–7.

[47] Wikipedia. 2023. the 2023 Turkey and Syria Earthquakes. https://wiki.openstreetmap.org/wiki/2023_Turkey_Earthquakes

[48] Xin Wu, Danfeng Hong, Jiaojiao Tian, Jocelyn Chanussot, Wei Li, and Ran Tao. 2019. ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features. *IEEE Transactions on Geoscience and Remote Sensing* 57, 7 (2019), 5146–5158.

[49] Z Wu, H Li, and A Zipf. 2020. From Historical OpenStreetMap data to customized training samples for geospatial machine learning. In *Proceedings of the Academic Track at the State of the Map 2020 Online Conference*.

[50] Yiqun Xie, Erhu He, Xiaowei Jia, Han Bao, Xun Zhou, Rahul Ghosh, and Praveen Ravirathinam. 2021. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 767–776.

[51] Yanyu Xu, Zhixin Piao, and Shenghua Gao. 2018. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5275–5284.

[52] Yifang Yin, Zhenguang Liu, Ying Zhang, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. 2019. Gps2vec: Towards generating worldwide gps embeddings. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 416–419.

[53] Menghua Zhai, Tawfiq Salem, Connor Greenwell, Scott Workman, Robert Pless, and Nathan Jacobs. 2019. Learning geo-temporal image features. *arXiv preprint arXiv:1909.07499* (2019).

[54] Yating Zhang, Adam Jatowt, and Katsumi Tanaka. 2017. Is Tofu the Cheese of Asia? Searching for Corresponding Objects across Geographical Areas. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 1033–1042.