Spatioformer: A Geo-Encoded Transformer for Large-Scale Plant Species Richness Prediction

Yiqing Guo, *Member, IEEE*, Karel Mokany, Shaun R. Levick, Jinyan Yang,
and Peyman Moghadam, *Senior Member, IEEE*

Abstract— Earth observation (EO) data have shown promise in predicting species richness of vascular plants (α-diversity), but extending this approach to large spatial scales is challenging because geographically distant regions may exhibit different compositions of plant species (β-diversity), resulting in a location-dependent relationship between richness and spectral measurements. In order to handle such geolocation dependence, we propose *Spatioformer*, where a novel geolocation encoder is coupled with the transformer model to encode geolocation context into remote sensing imagery. The Spatioformer model compares favorably to state-of-the-art models in richness predictions on a large-scale ground-truth richness dataset harmonized Australian vegetation plot (HAVPlot) that consists of 68 170 in situ richness samples covering diverse landscapes across Australia. The results demonstrate that geolocational information is advantageous in predicting species richness from satellite observations over large spatial scales. With Spatioformer, plant species richness maps over Australia are compiled from the Landsat archive for the years from 2015 to 2023. The richness maps produced in this study reveal the spatiotemporal dynamics of plant species richness in Australia, providing supporting evidence to inform effective planning and policy development for plant diversity conservation. Regions of high richness prediction uncertainties are identified, highlighting the need for future in situ surveys to be conducted in these areas to enhance the prediction accuracy.

Index Terms— Biodiversity, geolocation encoder, mapping, species richness, transformer, vascular plant.

I. INTRODUCTION

AUSTRALIA is home to a large and diverse range of plant species, with over 21 000 known native species of vascular plants and 93% of these being endemic [1], [2]. The richness of plant species, also known as α-diversity, is highly important in maintaining the functioning of ecosystems, such as habitat provision, carbon sequestration, and

Yiqing Guo, Karel Mokany, and Jinyan Yang are with CSIRO Environment, Acton, ACT 2601, Australia (e-mail: yiqing.guo@csiro.au; karel.mokany@csiro.au; jinyan.yang@csiro.au).

Shaun R. Levick is with CSIRO Environment, Winnellie, NT 0822, Australia (e-mail: shaun.levick@csiro.au).

Peyman Moghadam is with CSIRO Robotics, Data61, Pullenvale, QLD 4069, Australia, and also with the School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, QLD 4000, Australia (e-mail: peyman.moghadam@csiro.au).

water cycling [3], [4], [5], [6]. However, anthropogenic interference, such as deforestation, overgrazing, and urbanization, has resulted in a decline in plant species richness [7], [8]. In response, conservation activities have been initialized and conducted across the country aiming to preserve plant diversity [9], [10], [11]. Accurate and up-to-date maps of plant species richness will strongly support effective planning and policy-making for these activities [12], [13].

Earth observation (EO) data provide rapid and near-real-time estimates of changes in land surface conditions across large regions [14], [15], [16], [17]. This makes remote sensing imagery a favorable data source for plant species richness modeling compared with another widely adopted approach where environmental variables, such as temperature, precipitation, soil texture, and topographic heterogeneity, are used as richness predictors [18]. The reason is that environmental variables drive mainly the environmental potential of plant habitats (i.e., the capacity to sustain a certain level of richness), rather than represent the actual conditions on the ground like those observed by remote sensing satellites. For example, deforestation, floods, and bushfires could cause a reduction in richness [19], but such reduction might not be reflected by environmental variables. Therefore, environmental variables are often aimed at predicting the natural patterns in diversity in a preintensification reference state, while remote sensing data are more valuable for monitoring actual changes in those patterns.

Australia covers an area of over seven million square kilometers. As a result of the relatively large geographical extents, versatile types of plant habitats are found across the country, differing in their inventories of plant species present, which have been shaped by a variety of factors such as biogeographic history, climate, and geography [20]. To understand the spatiotemporal distribution of plant species richness, perseverant in situ field surveys have been conducted over the past several decades. Via various survey campaigns, a wealth of 219 552 richness samples have been gathered across the country as of the year 2022 [18]. These samples represent a broad range of landscapes across the continent and, therefore, present a unique opportunity to unravel the potentially intricate relationship between richness measurements and satellite observations. Nevertheless, geographically distant regions may exhibit distinct assemblages of plant species with differed compositional properties (i.e., β-diversity [21]), making it challenging to model richness over large spatial scales. Due to spatial variations in plant species composition, a location

with a set of plant species would be expected to display quite different spectral features in remote sensing imagery from another location with a dissimilar plant composition, even if the two locations sustain the same richness of species [22], [23]. Through statistical regression analysis for two regions in southeast Australia, previous studies [13], [24] suggested that the relationship between plant species richness and hyper/multispectral satellite observations is region-specific. To account for the location dependence, we need a model that is capable of taking in geolocation context when mapping plant species richness over large spatial scales.

The transformer model, first introduced in [25], is built upon the self-attention mechanism [26]. The model attends effectively to information of high importance in the input data, as the self-attention module is capable of capturing intricate data structures and dependencies [26], [27]. Initially proposed for language tasks, the transformer model has shown promise for image understanding due to its superior ability over convolutional neural networks (CNNs) in capturing global dependencies between different regions of an image [28]. As a seminal work on applying transformer to image data, the vision transformer (ViT) model [28] first divides an image into nonoverlapping patches, followed by projecting each patch into a feature vector, which is then fed into the self-attention module, with state-of-the-art performance being achieved on benchmark datasets. Given remote sensing imagery captures rich features in the spectral dimension, the SpectralFormer model [29] was developed to effectively embed the spectral information. Recently, FactoFormer [30] developed a self-supervised factorized spectral–spatial transformer, where attention is computed by individually focusing on spectral and spatial dimensions in each transformer. These studies have demonstrated the effectiveness of transformer in processing remote sensing images (e.g., [29], [30]), but, to advance the model's application to remote sensing images recorded over large spatial scales, geolocational information could be leveraged. Unlike many types of imagery whose semantics are independent of the location where they are recorded, remote sensing images are intrinsically associated with geolocations [31], [32]. Considering that the composition of plant species is location-specific, incorporating geolocation context could be helpful in modeling the location-dependent relationship between richness and remote sensing imagery.

Geo-coordinates provide geographical priors that supplement geolocation context to the image data [31], [32], [33], [34], [35], [36], [37]. While a straightforward way to utilize geolocational features is to concatenate the original longitude and latitude coordinates into the model, this approach has been shown to yield almost no gain in performance [34]. To deal with this problem, a geo-feature extraction approach was proposed in [34] for CNN models, where the geo-coordinates were projected into a higher dimensional feature space with a geolocation encoder, whose outputs were then merged into those of a CNN-based image network. It was observed that by leveraging geolocation context, a 7% increase in accuracy was achieved for an image dataset spanning over the continental United States [34]. This geo-encoded CNN model was later applied to global-scale vegetation canopy height mapping

with satellite imagery [38], where the geolocations served as a prior. Other state-of-the-art geolocation encoders include Space2Vec [39], Sphere2Vec [40], PE-GNN [41], and a more recent algorithm that is based on the spherical harmonic basis functions [42]. Multiscale sinusoidal functions are favored in building these encoders (e.g., [39], [42]), thanks to their merits of being bounded in value, infinitely extended in space, and possessing multiresolution scalability. Geolocation encoding is demonstrated to be effective in many large-scale geospatial problems, such as animal species categorization [38], [40], water quality prediction [43], event/activity recognition [44], and remote sensing scene classification [31], [40].

In this study, we aim to predict the spatiotemporal distribution of plant species richness in Australia from EO imagery with geolocation context being taken into account. Considering that the relationship between plant species richness and remote sensing imagery varies from one location to another due to differences in vegetation composition, we propose *Spatioformer*, where a novel geolocation encoder is coupled with the transformer model in order to incorporate the geolocation context. The performance of Spatioformer in richness mapping is compared with a CNN model, a ViT model, and the FactoFormer model where the geolocational information is not encoded. Through quantitative analyses, we seek to address primarily the following important questions.

1) Does Spatioformer perform better than state-of-the-art algorithms in predicting plant species richness over large spatial scales?
2) What are the spatial patterns of plant species richness in Australia inferred from remote sensing evidence?
3) Where should future in situ surveys be conducted as suggested by the mapping results?

The rest of this article is organized as follows. Section II describes the study area and the datasets used for modeling, including the ground-truth samples of plant species richness and satellite imagery. Section III introduces the methods with a focus on the proposed Spatioformer model. This model is developed with the aim of capturing the location-dependent relationships between plant species richness and remote sensing imagery over large spatial scales. Section IV describes the experimental settings for training and validation of the Spatioformer model. Section V presents the results of applying Spatioformer to plant richness mapping across Australia and discusses the implications of these findings for biodiversity conservation and future research directions. Finally, Section VI concludes this article.

II. STUDY AREA AND DATASETS

A. Study Area

This study was focused on natural and near-natural functioning terrestrial ecosystems within the Australian continent and nearby islands, as shown in Fig. 1. The ecosystems considered include natural lands (e.g., conservation reserves, managed resource protected areas, and lands of minimal use) and lands utilized for production purposes from near-natural environments (e.g., native grasslands and forests for production purposes) (see Fig. 1). We excluded heavily modified
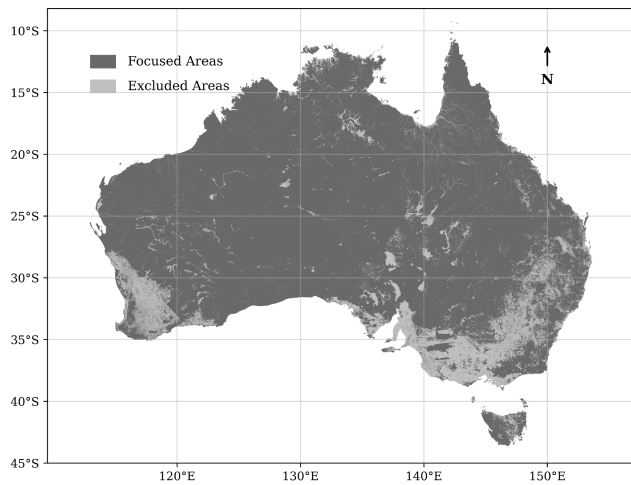
Fig. 1. Map of the study area. This work was focused on natural and near-natural functioning terrestrial ecosystems within the Australian continent and nearby islands, as colored in dark gray in the figure, while heavily modified landscapes and water bodies colored in light gray were excluded from our analysis.

landscapes from our analysis, including agricultural lands (e.g., croplands, horticultural lands, and plantation forests), urban regions (e.g., industrial and residential lands), and water bodies (e.g., lakes, rivers, and reservoirs) (see Fig. 1). Those areas were excluded because vegetation surveys, which primarily aim to gather information for the preservation of native plant species, have been rarely conducted in heavily modified landscapes. The focused and excluded areas were identified with the Catchment Scale Land Use of Australia dataset in 50-m spatial resolution (updated in December 2018) [45].

B. Plant Species Richness Samples

The ground-truth samples of plant species richness were obtained from the harmonized Australian vegetation plot (HAVPlot) dataset [18], [46], where each sample represented a plot area of 400 m^2 (20 × 20 m). Out of the 219 552 samples in the HAVPlot dataset, 68 170 samples were selected for modeling in this study. These samples were collected via various field campaigns as a perseverant effort spanning the years from 1986 to 2020 (please refer to the acknowledgment section for details on the custodians of these samples). As detailed in [18], efforts have been made to harmonize the samples from different field campaigns, with the aim of minimizing the temporal and spatial variations among samples arising from discrepancies in experimental design. The rest samples in the HAVPlot dataset were removed from our analysis because they consisted of less than 70% native Australian species or were collected before the year 1986 and were unable to be matched with a Landsat observation (as described later in Section II-C). Fig. 2 shows the spatial locations of the selected species richness samples, colored by the richness values (in the unit of the number of species per 400 m^2) [see Fig. 2(a)] and by the years of survey [see Fig. 2(b)]. These samples cover diverse landscapes across Australia. Scattered pockets of natural or near-natural areas among heavily modified lands are also represented by samples (see the insets of Fig. 2). Regions of easier human access, such as the southeast and southwest
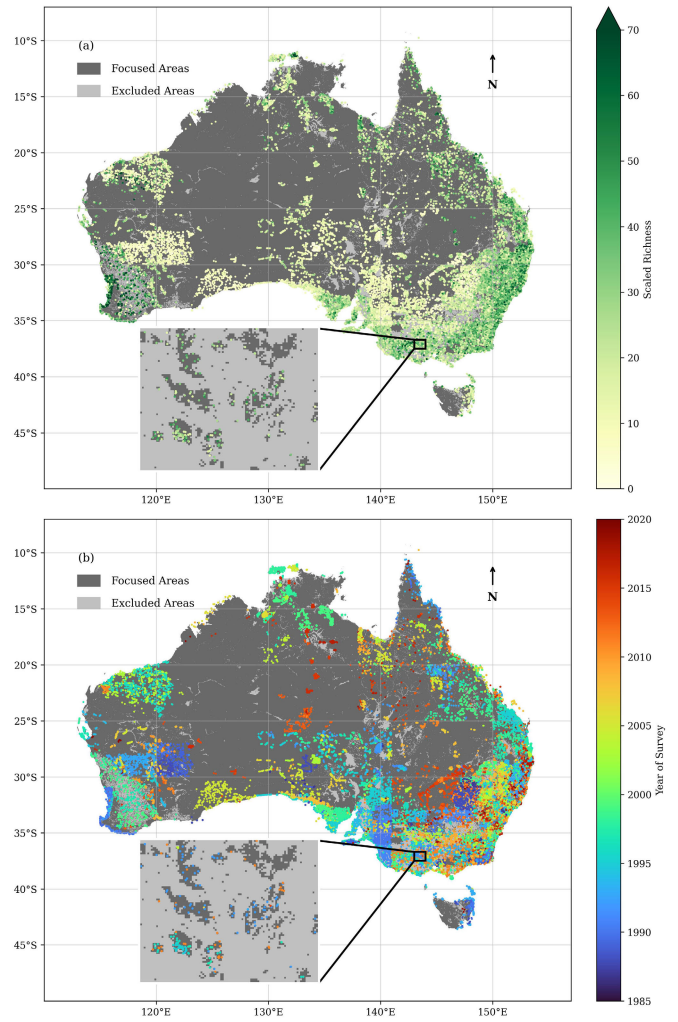

Fig. 2. Locations of ground survey samples colored by (a) species richness values (in the unit of the number of species per 400 m^2) and (b) years of the survey. The insets provide zoomed-in views of a region in southeast Australia. A total of 68 170 samples from the HAVPlot dataset [18], [46] were used for modeling in this study. These samples were collected via various field campaigns as a perseverant effort spanning the years from 1986 to 2020 (please refer to the acknowledgments section for details on the custodians of these samples).

coastal areas, generally show a higher sample representativeness than the less habitable inland of arid or semiarid climate. In particular, the arid/semiarid western interior has low or no coverage of plots.

C. Satellite Imagery

The satellite data were queried within CSIRO's implementation of the Open Data Cube [47] known as the Earth analytics science and innovation (EASI) platform. Landsat images were extracted from the Geoscience Australia Landsat Geometric Median (Geomedian) and median absolute deviation (MAE) collection 3 product (version 3.10) [48]. The Geomedian Landsat-5 and Landsat-8 images with a spatial resolution of 30 × 30 m were downloaded for years from 1986 to 2011 and 2013 to 2023, respectively. Each pixel in the Geomedian imagery represented the statistical median of all observations for that pixel from the given calendar

year [49], [50]. The Geomedian observations provided a summarized representation of annual habitat conditions for Australia's landscapes, considering that vegetation in the vast arid and semiarid regions revives after each rainfall and exhibits no definite phenological cycles. These Geomedian images were analysis-ready surface reflectance data, with a series of preprocessing steps having been applied, including atmospheric correction, terrain correction, and bidirectional reflectance distribution function (BRDF) adjustment. The six spectral bands shared by Landsat-5 and Landsat-8 were selected for analysis, namely, blue, green, red, near-infrared, and short-wave-infrared 1 and 2. For each richness sample, we matched it with a Geomedian Landsat image from the same year in which the sample was collected. The image consisted of 9×9 pixels with the center pixel aligned to the location of the survey plot. We supplied spatially adjacent pixels neighboring the plot for modeling, along with the center pixel itself, because plants often live as a community where each plant interacts with its spatial neighbors in a close relationship. For example, the variation in spectral patterns among neighboring pixels, known as the spectral diversity or optical diversity [51], is related to plant species diversity via the spectral variation hypothesis [52]. The spatially neighboring pixels, thus, may provide supplementary information to assist in the richness prediction.

III. METHODS

A. Overview

To encode geospatial information into remote sensing imagery for large-scale plant species richness mapping, we propose the Spatioformer model that integrates the transformer with a geolocation encoder. In the following, we first introduce our approach to geolocation encoding in Section III-B, followed by detailing the Spatioformer structure in Section III-C. Theoretical analysis of Spatioformer is given in Section III-D. Finally, plant species richness mapping with Spatioformer is described in Section III-E.

B. Geolocation Encoder

We choose multiscale sinusoidal functions, i.e., sine and cosine with different frequencies, for geolocation encoding. This choice is supported by previous studies (e.g., [39], [42]) that demonstrated the advantages of employing multiscale sinusoidal functions for geolocation encodings, such as their bounded values, infinite spatial extension, and multiresolution scalability. For a given geolocation (x_i, y_i) with x_i and y_i being the longitude and latitude coordinates, its geolocation token is encoded as

$$\boldsymbol{g}^{(x_i,y_i)} = \left[g_1^{(x_i,y_i)}, g_2^{(x_i,y_i)}, \ldots, g_j^{(x_i,y_i)}, \ldots, g_d^{(x_i,y_i)} \right]^{\mathrm{T}} \quad (1)$$

where d is the dimension of geolocation token and T denotes the transpose. The jth element in $\boldsymbol{g}^{(x_i,y_i)}$, $g_j^{(x_i,y_i)}$, is calculated as follows:

$$g_j^{(x_i,y_i)} = \begin{cases} \sin\left(\dfrac{x_i}{w_j}\right) + \sin\left(\dfrac{y_i}{v_j}\right), & \text{if } j \text{ is even} \\ \cos\left(\dfrac{x_i}{w_j}\right) + \cos\left(\dfrac{y_i}{v_j}\right), & \text{if } j \text{ is odd} \end{cases} \quad (2)$$
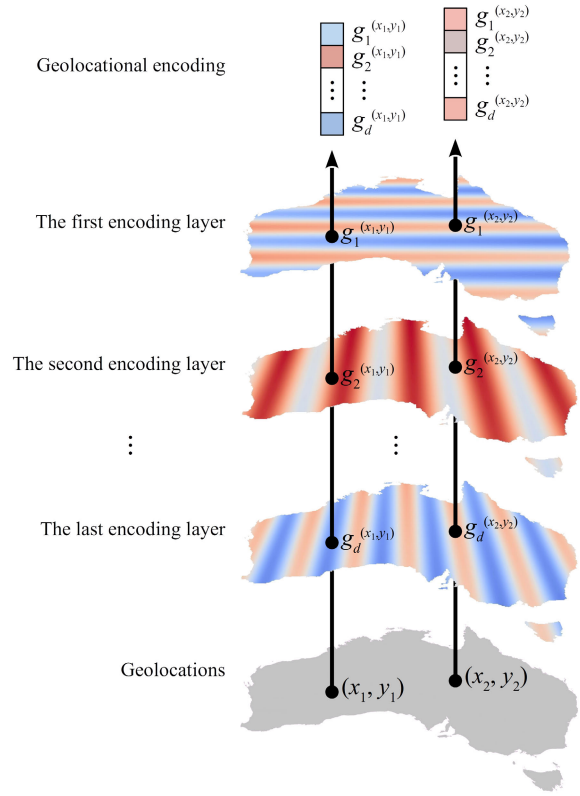

Fig. 3. Graphic illustration of geolocation encoding for two example locations (x_1, y_1) and (x_2, y_2). The geolocation encoding vectors for these two locations were constructed with values referenced from the corresponding positions on the encoding layers.

where $w_j = a \cdot c^{(j/d)}$ and $v_j = a \cdot c^{((d-j)/d)}$, with a and c being preset constants, determine the spatial frequencies of encoding layers. The constants a and c need to be set such that each pixel within the mapping area is provided with a sufficiently distinctive geolocation token by the encoding layers.

A graphic illustration of the proposed geolocation encoder is shown in Fig. 3 for two example locations. The geolocation encoding vectors for these two locations are constructed with values referenced from the corresponding positions in a total of d encoding layers. This geolocation encoder could provide geo-gradient information not only in the cardinal directions but also in various diagonal directions (as shown in Section V-A). The d, a, and c values configured in this study for plant species richness mapping are described in Section IV-B.

C. Spatioformer Structure

The Spatioformer structure is shown in Fig. 4. The input of the model is an image in two spatial dimensions (i.e., longitude and latitude) and the spectral dimension. The image is first spatially divided into separate pixels or image patches and then flattened, before being fed into a linear forward layer that projects the pixels/patches into the embedding space. For each pixel/patch, its embedding $\boldsymbol{e}_i$ is added by its geolocation token $\boldsymbol{g}^{(x_i,y_i)}$, which is calculated from its geo-coordinates (x_i, y_i) (as described in Section III-B)

$$\boldsymbol{e}_i' = \boldsymbol{e}_i + \lambda \boldsymbol{g}^{(x_i,y_i)} \quad (3)$$
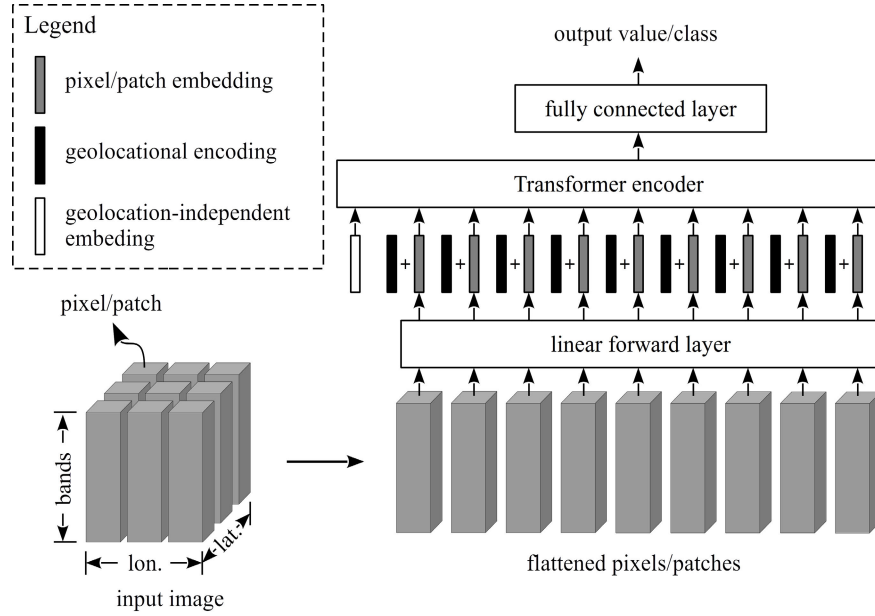
Fig. 4. Graphic illustration of the Spatioformer structure. An image is first spatially divided into separate pixels or image patches and then flattened, before being fed into a linear forward layer which projects the pixels/patches into the embedding space. For each pixel/patch, its embedding is added by its geolocation token. The geolocation-encoded embeddings are then fed into the transformer encoder, together with a geolocation-independent token to account for geolocation-independent components in the input–output relationship. A fully connected layer is set as the output layer to produce the predicted value or class.

where λ is a learnable parameter balancing the relative contribution between pixel values and geolocational information, given that these two different types of data may differ in their magnitudes. The geolocation-encoded embeddings $\boldsymbol{e}'_i$ are then fed into the transformer encoder, together with a learnable geolocation-independent token at the pixel/patch level to account for geolocation-independent components in the input–output relationship. A fully connected layer is set as the output layer to produce the predicted value or class, depending on whether the problem to be solved is a regression or classification problem. The configuration of Spatioformer for plant species richness mapping is described in Section IV-B. The geolocation encoder in Spatioformer integrates geolocation context into the images, enabling the model to leverage this information in modeling plant species richness distribution in this study.

D. Theoretical Analysis

At the core of standard transformer models lies the self-attention module. For a single-head self-attention module [53], with an input length of n tokens, its output for the ith input token $\boldsymbol{x}_i$ can be expressed as

$$z_i = \sum_{j=1}^{n} \frac{\exp(\alpha_{ij})}{\sum_{j'=1}^{n} \exp(\alpha_{ij'})} (\boldsymbol{x}_j \boldsymbol{W}^V)$$

$$\alpha_{ij} = \frac{1}{\sqrt{d}} (\boldsymbol{x}_i \boldsymbol{W}^Q)(\boldsymbol{x}_j \boldsymbol{W}^K)^{\mathrm{T}} \qquad (4)$$

where $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, and $\boldsymbol{W}^V$ are trainable weights for the query, key, and value matrices, respectively.

In Spatioformer, the pixel embeddings are added together with their respective geolocation encoding tokens, before being fed into the self-attention module. Compared with concatenating longitude and latitude coordinates directly to pixel embeddings, the geolocation encoding is able to project geocoordinates into a higher dimension (i.e., the pixel embedding space) as a better feature descriptor. As a result, the calculation of α_{ij} is different from that shown in (4)

$$\alpha_{ij} = \frac{1}{\sqrt{d}} \big[(\boldsymbol{x}_i + \lambda \boldsymbol{g}_i)\boldsymbol{W}^Q\big]\big[(\boldsymbol{x}_j + \lambda \boldsymbol{g}_j)\boldsymbol{W}^K\big]^{\mathrm{T}}. \qquad (5)$$

By expanding (5), we get

$$\alpha_{ij} = \underbrace{\frac{1}{\sqrt{d}}(\boldsymbol{x}_i \boldsymbol{W}^Q)(\boldsymbol{x}_j \boldsymbol{W}^K)^{\mathrm{T}}}_{(1)} + \underbrace{\frac{1}{\sqrt{d}}(\boldsymbol{x}_i \boldsymbol{W}^Q)(\lambda \boldsymbol{g}_j \boldsymbol{W}^K)^{\mathrm{T}}}_{(2)}$$
$$+ \underbrace{\frac{1}{\sqrt{d}}(\lambda \boldsymbol{g}_i \boldsymbol{W}^Q)(\boldsymbol{x}_j \boldsymbol{W}^K)^{\mathrm{T}}}_{(3)} + \underbrace{\frac{1}{\sqrt{d}}(\lambda \boldsymbol{g}_i \boldsymbol{W}^Q)(\lambda \boldsymbol{g}_j \boldsymbol{W}^K)^{\mathrm{T}}}_{(4)}.$$

$$(6)$$

The four terms in the right-hand side of (6) correspond to the pixel-to-pixel (Term 1), pixel-to-geolocation (Term 2), geolocation-to-pixel (Term 3), and geolocation-to-geolocation attention (Term 4), respectively. Therefore, training Spatioformer is equivalent to joint optimization of pixel embedding and geolocation encoding by taking into account their respective self-correlations and the correlations between each other.

E. Plant Species Richness Mapping

In this study, the Spatioformer model was trained with ground-truth richness samples and their corresponding Landsat images. With the trained model, plant species richness maps were compiled by inferring richness values from Landsat images. Mapping results were demonstrated for the years

from 2015 to 2023. Compared with older years, maps for these recent years can better inform current and future conservation activities and policy making. We then aggregated these annual maps to compile the mean richness map across the nine years in order to study plant species richness with enhanced spatial patterns. We also calculated the standard deviation map of richness over the nine years to identify places where predicted plant species richness showed high temporal variations. The size of input images ranged from 1×1 to 9×9 pixels in order to examine how mapping accuracy varied when different spatial scales of plant community were taken into account.

The uncertainty map for richness predictions was produced with the Monte Carlo dropout approach [54]. For each predicted richness value, its uncertainty was calculated as the coefficient of variation over multiple additional model predictions under a dropout rate of 0.5 that randomly drops 50% of model parameters as follows:

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \Big/ \bar{y}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad (7)$$

where ε is the uncertainty metric, y_i is the predicted species richness under the dropout mode, and n is the number of additional model predictions. In this study, n was set to 100.

IV. EXPERIMENTAL SETUP

A. Dataset Partition

In this study, the in situ samples across all years were used to build a model under a training/validation/test data splitting scheme, followed by applying the model to make mapped predictions for each year. A random data splitting scheme for accuracy assessment in large-scale mapping was cautioned by [55], [56], and [57], as randomly sampled hold-outs of validation and test data may spatially autocorrelated with the training samples, leading to an overestimation of accuracy. Data splitting schemes based on spatially blocked hold-outs were, therefore, recommended to mitigate this problem [56], [57].

In our work, a block-based training/validation/test scheme was adopted. We first divided the Australian territory into 958 tiles of 100×100 km based on a geographical zoning scheme provided by Geoscience Australia, as shown in Fig. 5. Among them, we randomly selected 766 training tiles (approximately 80%), 96 validation tiles (approximately 10%), and 96 test tiles (approximately 10%) (see Fig. 5). Ground-truth richness samples located within the training, validation, and test tiles were assigned into the respective sets (see Fig. 5). As a result, the training, validation, and test sets consisted of 55 159, 6753, and 6258 samples, respectively. In addition to mitigating the spatial autocorrelation problem, having training, validation, and test sets spatially separate from each other could also help the Spatioformer avoid overfitting to local data anomalies, which would otherwise generate hotspot artifacts (i.e., extreme values at local scales due to overfitting) in the predicted richness maps.
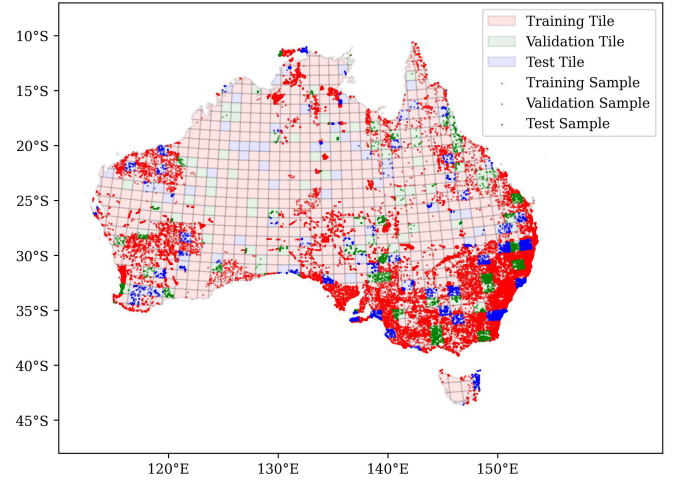

Fig. 5. Partition of ground samples into training, validation, and test sets based on geographical tiles. The Australian territory was divided into 958 tiles of 100×100 km, with 766 tiles (approximately 80%), 96 tiles (approximately 10%), and 96 tiles (approximately 10%) being randomly selected as the training, validation, and test tiles. Samples located within the training, validation, and test tiles were assigned to the respective sets.

B. Spatioformer Model Setup and Hyperparameter Selection

The input images of 9×9 pixels, with each pixel being of 30×30 m spatial resolution, were embedded into a 16-D feature space using a linear feedforward layer. In cases where input images have irregular edges, a patching strategy can be implemented by applying zero-padding along the edges to standardize the input dimensions. A geolocation token of 16-D was added to the embedding of each pixel, where the embedding feature dimension of 16 was determined based on multiple test runs. The geolocation-encoded pixel embeddings were fed into three layers of eight-head self-attention module with each layer followed by a 64-D feedforward network. A fully connected network with a hidden layer of 1024 nodes was set as the output layer. The mean square error (MSE) between model-predicted and ground-truth richness values was employed as the loss function. The Adam optimizer was employed for network optimization.

Model hyperparameters were determined with multiple test runs. The learning rate and weight decay were set to 1×10^{-3} and 1×10^{-4}, respectively. A cosine annealing scheme with a warm start was scheduled for the learning rate. A dropout rate of 0.1 was applied. The constants a and c in the geolocation encoder were set to 1 and 100, respectively. The initial value for the learnable parameter λ balancing the relative contribution between pixel values and geolocational information was set to 1×10^{4}.

C. Benchmarking With State-of-the-Art Models

The performance of Spatioformer was compared with a CNN model [58], a ViT model [25], and the FactoFormer model [30]. The CNN model consisted of three convolutional layers, with each layer followed by an ReLU activation layer and a batch normalization layer. Each convolutional layer consisted of eight filters with a kernel size of 3×3. Following the convolutional layers were a flattened layer and a fully
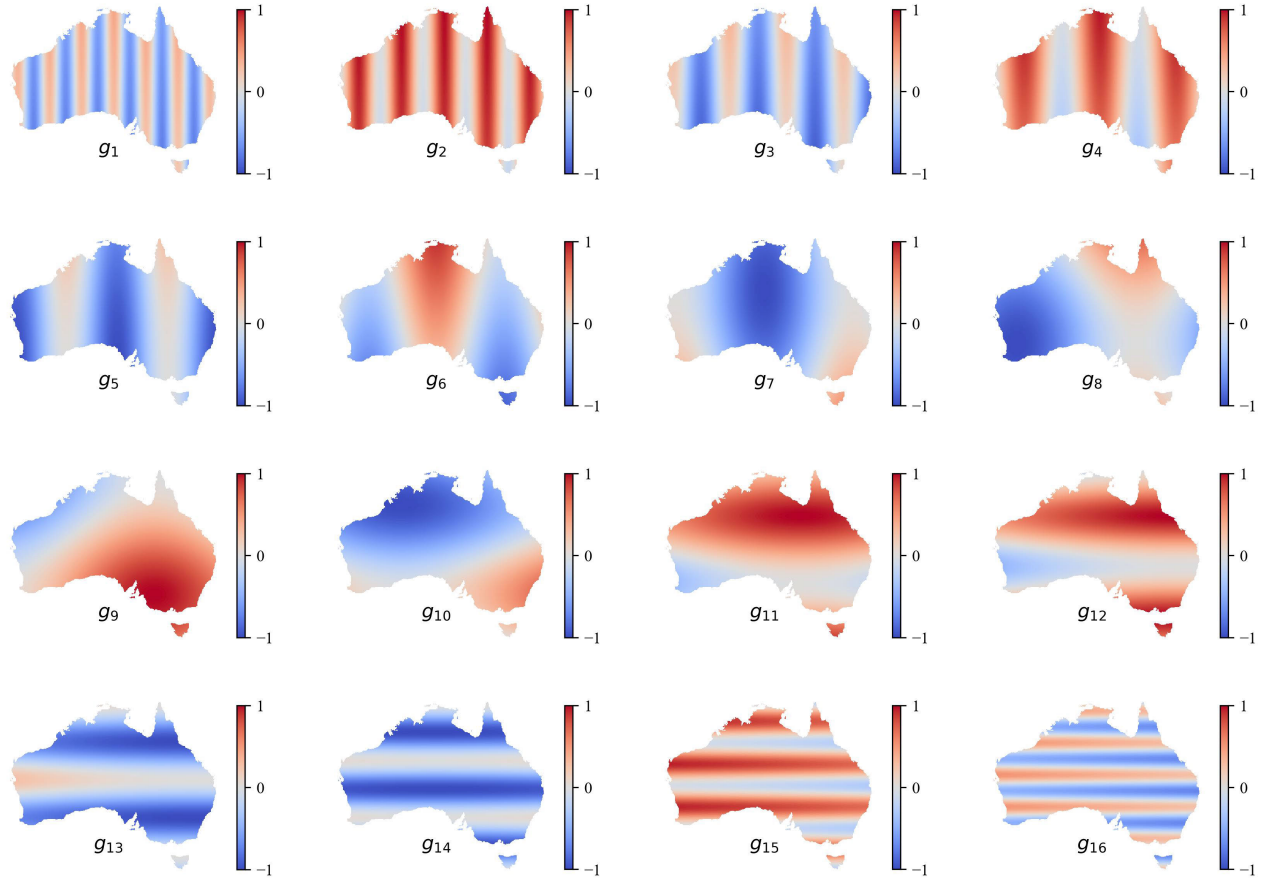
Fig. 6. Geolocation encoding layers for plant species richness mapping in our study. These layers were generated with the proposed geolocation encoder. The first few layers (from g_1 to g_4) highlight the longitudinal gradients, the last few layers (from g_{13} to g_{16}) highlight the latitudinal gradients, and the middle layers (from g_5 to g_{12}) show smoother spatial pattern in diagonal directions.

connected network. For the ViT model, each pixel was first embedded into a 16-D feature space via a linear feedforward layer, which was then fed into three layers of an eight-head self-attention module. Each self-attention layer was followed by a feedforward network of 64-D. A flattened layer and a fully connected network were set as the output layers. For the FactoFormer model, each input image cube was first split into nonoverlapping tokenized patches along spectral and spatial dimensions, followed by processing them with two transformers simultaneously. Then, the outputs of each transformer were flattened, concatenated, and passed to a fully connected network.

These benchmark models were optimized on their performance in the same way as the optimization of Spatioformer structure, i.e., grid search of the optimal model hyperparameters via multiple test runs. The same as the setting for Spatioformer, 9×9 image pixels over each richness sample were supplied to CNN, ViT, and FactoFormer as input. Seven evaluation metrics were calculated and compared for these models, including coefficient of correlation (r), coefficient of determination (r^2), mean absolute error (MAE), relative absolute error (RAE), MSE, relative square error (RSE), and root MSE (RMSE) between model-predicted and ground-truth species richness (in the unit of the number of species per 400 m^2).

V. RESULTS AND DISCUSSION

A. Geolocation Encoding

The 16 geolocation encoding layers adopted in our study for plant species richness prediction are shown in Fig. 6. By comparing the spatial patterns of these layers, it was observed that the first few layers (from g_1 to g_4) highlighted the longitudinal gradients of different spatial frequencies, while the last few layers (from g_{13} to g_{16}) highlighted the latitudinal gradients. The middle layers (from g_5 to g_{12}) showed smoother spatial patterns that presented diagonal gradients in various directions. The differed spatial patterns rendered by these layers were capable of providing adequate discriminating power to identify the encoding of one location from that of another.

In richness prediction, each geo-location was encoded as a 16-D token whose elements were referenced from the corresponding positions in the encoding layers shown in Fig. 6. These tokens could project the geo-coordinates into a higher dimensional feature space, given that utilizing the original longitude and latitude coordinates as features was shown to yield almost no gain in performance [34]. As observed from the figure, the high-dimensional encoding could provide geo-gradient information not only in the cardinal directions but also in various diagonal directions. The encoding layers with information of varied spatial frequencies also made it
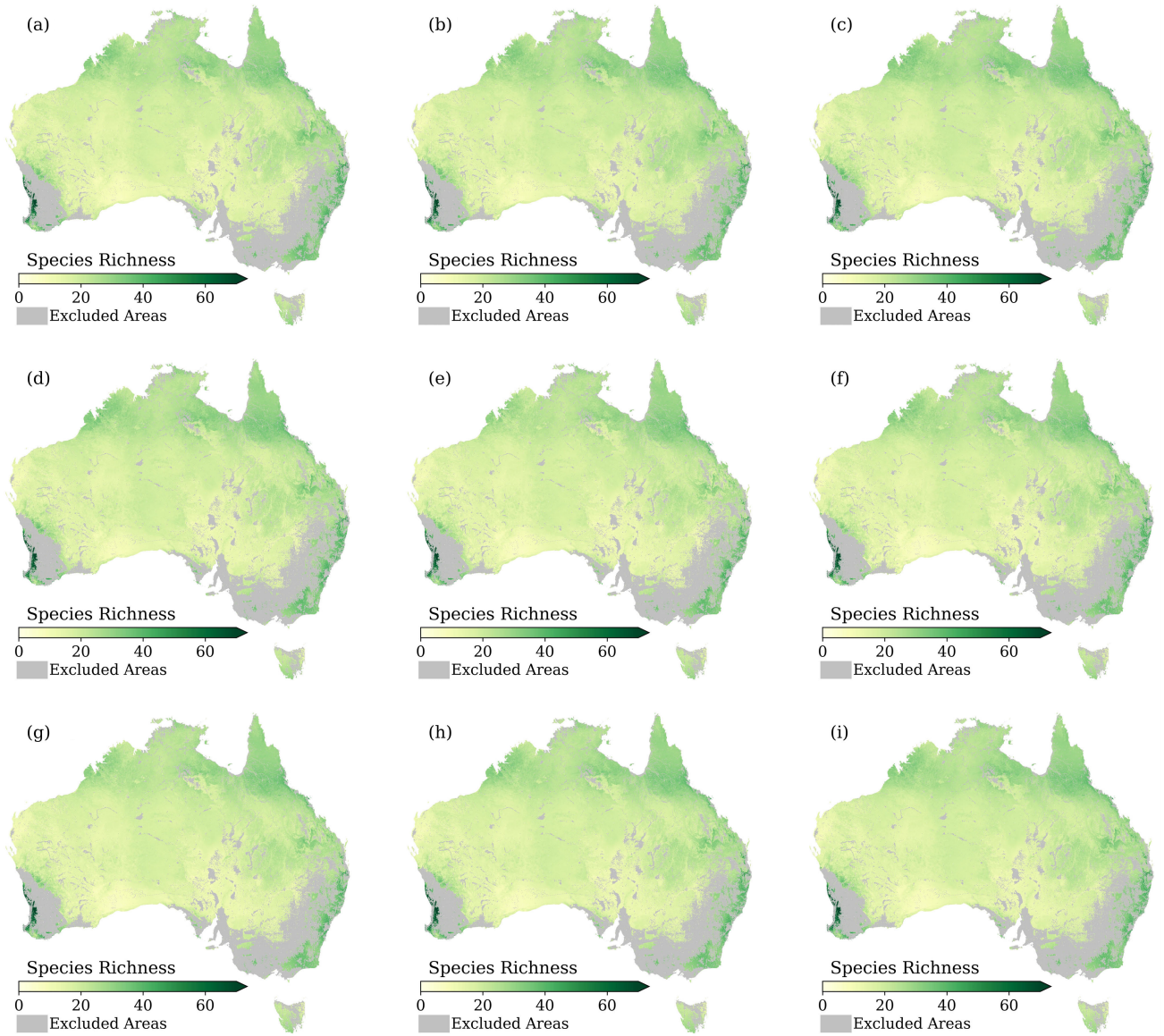
Fig. 7. Predicted annual maps of plant species richness distribution in Australia from 2015 to 2023 (a)–(i). These maps were compiled by applying the proposed Spatioformer model to Landsat observations from the respective years over the Australian territory. Richness values are in the unit of the number of species per 400 m^2.

possible to distinguish geo-information at different spatial scales.

The geolocation tokens in richness prediction aimed at providing sufficient geolocation context such that the location dependence in the relationship between plant species richness and remote sensing imagery could be accounted for. Considering the continental scale of plant species richness mapping that this study aimed to handle, the geolocation context might help the model distinguish locations with dissimilar plant compositions, with the potential to increase the richness prediction accuracy across large spatial scales.

B. Annual Richness Distributions

Fig. 7 shows the annual maps of plant species richness distribution in Australia from 2015 to 2023, which were compiled by applying the trained Spatioformer model to Landsat observations from the respective years. It was observed that the

spatial patterns of Landsat-derived richness as displayed in these maps were consistent with the ground-sampled richness values displayed in Fig. 2(a).

From the annual maps shown in Fig. 7, it was seen that the Jarrah Forest bioregion near the southwestern coast of Western Australia showed the highest plant species richness within Australia, with values generally higher than 60 species per 400 m^2. The high species richness in this region could be evidenced by the high richness samples recorded in field surveys [see Fig. 2(a)]. With a Mediterranean climate, this region harbors a biodiverse ecosystem with a distinctive composition of native vascular plants including the endemic Eucalyptus species *Eucalyptus laeliae*, making it a region of high conservation value.

The eastern and southeastern coastal areas of Australia showed a medium-high richness of plant species with values ranging from 30 to 60 species per 400 m^2 (see Fig. 7). These

Metric	CNN [58]	ViT [25]	FactoFormer [30]	Spatioformer (Ours)
r	0.52	0.55	0.60	0.77
r^2	0.27	0.29	0.35	0.59
MAE	10.41	10.37	9.81	7.83
RAE	0.37	0.37	0.35	0.29
MSE	188.69	182.81	168.03	105.85
RSE	0.18	0.17	0.16	0.11
RMSE	13.74	13.52	12.96	10.29

areas cover mostly the pristine ecosystems within the Great Dividing Range, where the climate varies from temperate to subtropical. Many vascular plant species of high conservation importance have been identified in these areas. For example, the Snowy Mountains region, located to the southwest of Canberra, is home to 212 species of vascular plants, of which 21 are endemic [59]. Another example is the Southern Table-lands region to the southwest of Sydney, where out of the approximately 1200 species of vascular plants, 30 have been listed as threatened due to a high degree of human interference and habitat alteration [60].

It was observed that the Australian Savanna in the north of the continent (e.g., the northern parts of Western Australia, Northern Territory, and Queensland) showed a range of richness from 20 to 50 species per 400 m^2 (see Fig. 7). The tropical/subtropical savanna climate in this region provides a copious amount of rainfall in the wet season, followed by a long dry season during the rest of the year. This unique climatic pattern results in a landscape that is characterized by broad grassland interspersed with small trees and shrubs. It was identified from Fig. 7 that the Kimberley Tropical Savanna (along the northwest coast of Kimberley), the Carpentaria Tropical Savanna (along the south coast of the Gulf of Carpentaria), and the Einasleigh Uplands Savanna (in mid-north Queensland) showed a relatively higher richness of plant species than other areas within the Australian Savanna.

The inland areas showed a low richness of plant species with the number of species per 400 m^2 lower than 20 (see Fig. 7). This observation is in line with the range of richness values sampled in interior Australia [see Fig. 2(a)]. The low richness could be mainly attributed to the arid/semiarid climate that prevents the growth and sustenance of vascular plants, apart from the xeromorphic species. Although most coastal regions showed high to medium richness, the Nullarbor Plain (along the northwest coast of the Great Australian Bight) and the westernmost coastal region in Western Australia showed a very low richness of less than 10 species per 400 m^2. These two coastal regions are of limited plant diversity mainly because of the arid climate.

Compared with measuring richness via in situ ground sampling, the EO-derived richness maps in Fig. 7 provided a more complete spatiotemporal coverage. This spatially continuous and up-to-date knowledge of richness distribution could serve as a valuable asset to inform effective conservation strategies and activities. It is worth noting that heavily modified land-scapes (e.g., agricultural and urban regions) were masked out in the annual richness maps in Fig. 7. It was because the in situ dataset, on which our model was trained, consisted mainly of samples from natural and near-natural landscapes; thus, our model was not representative of heavily modified landscapes. Water bodies were also excluded from these richness maps as this work focused on terrestrial plants only.

C. Model Performance

Table I compares the performance of CNN [58], ViT [25], FactoFormer [30], and the proposed Spatioformer model in predicting ground-truth species richness from the test dataset. It was seen from the figure that the CNN model achieved 0.52, 0.27, 10.41, 0.37, 188.69, 0.18, and 13.74 on r, r^2, RAE, MSE, RSE, and RMSE, respectively, between model-predicted and ground-truth species richness. The ViT model achieved 0.55, 0.29, 10.37, 0.37, 182.81, 0.17, and 13.52, and the FactoFormer model achieved 0.60, 0.35, 9.81, 0.35, 168.03, 0.16, and 12.96 on these metrics. The Spatioformer model compared favorably to the aforementioned benchmark models, showing 0.77, 0.59, 7.83, 0.29, 105.85, 0.11, and 10.29 on r, r^2, RAE, MSE, RSE, and RMSE, respectively. As the location-dependent relationship between satellite observations and on-ground richness values was accounted for in our modeling with Spatioformer, the results shown in Table I suggested that the geolocation context could help improve prediction accuracy in plant species richness prediction.

Fig. 8 shows the predicted distribution of plant species richness across Australia in the year 2020, generated by the three benchmark models: CNN [58] [see Fig. 8(a)], ViT [25] [see Fig. 8(b)], and FactoFormer [30] [see Fig. 8(c)]. The relative difference of these maps to the map produced by the proposed Spatioformer model [see Fig. 7(f)] reveals notable discrepancies in certain regions, as shown in Fig. 8(d)–(f). For instance, the Jarrah Forest bioregion near the southwestern coast of Western Australia is a biodiverse region with richness generally higher than 60 species per 400 m^2, as evidenced by the ground-truth samples [see Fig. 2(a)]. While the high richness in this region was successfully mapped with the proposed Spatioformer model [see Fig. 7(f)], the three bench-mark models predicted this region to have 30∼40 species per
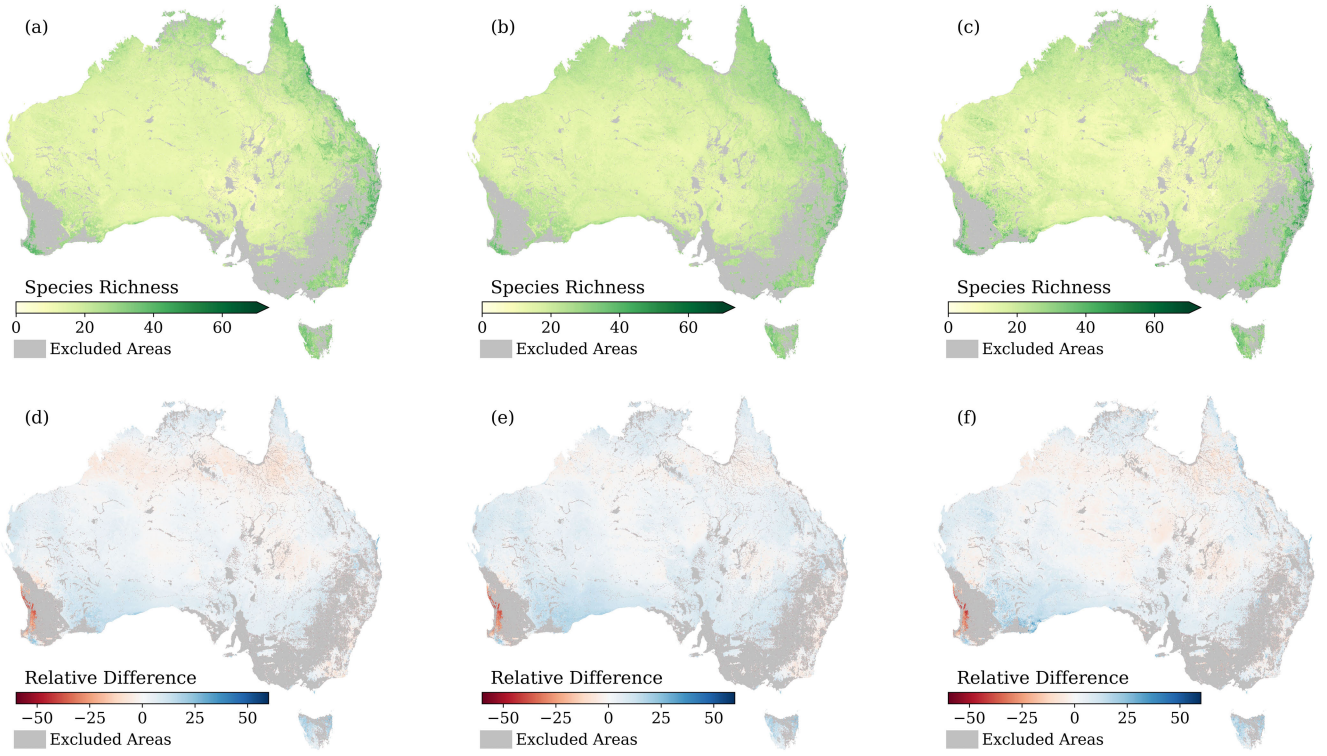
Fig. 8. Predicted plant species richness distribution in Australia with (a) CNN [58], (b) ViT [25], and (c) FactoFormer [30] for the year 2020 and (d)–(f) their relative difference to the result obtained with the proposed Spatioformer for the same year [see Fig. 7(f)]. Richness values are in the unit of the number of species per 400 m^2.

400 m^2 [see Fig. 8(a)–(c)], roughly 20∼40 lower relative to the Spatioformer predictions [see Fig. 8(d)–(f)]. The difference in richness prediction results between Spatioformer and the benchmark models is shown in more detail in Fig. 9 with enlarged maps of the predicted plant species richness distribution in the Jarrah Forest bioregion. As observed from Fig. 9, the proposed Spatioformer model predicted the richness to be generally higher than 60 species per 400 m^2 [see Fig. 9(d)], matching better with the ground-truth evidence [see Fig. 2(a)] than the benchmark models, which estimated roughly 30–40 species per 400 m^2 [see Fig. 9(a)–(c)]. Another example is the Nullarbor Plain along the northwest coast of the Great Australian Bight. The extremely dry climate in this region has resulted in poor plant diversity of fewer than 10 species per 400 m^2, as supported by field surveys shown in Fig. 2(a). While the Spatioformer model accurately predicted the very low richness values observed in the Nullarbor Plain [see Fig. 7(f)], the predicted richness values by three benchmark models [see Fig. 8(a)–(c)] were not as low as those indicated by the ground-truth samples, showing about 10∼20 higher relative to the Spatioformer map [see Fig. 8(d)–(f)]. These results suggested that the proposed Spatioformer model is capable of effectively adapting to local richness characteristics, thanks to the incorporation of geolocation context in its modeling.

As regions with a similar level of richness may present quite different spectral features in remote sensing imagery, predicting plant species richness over large spatial scales

is a challenging task. Previous studies suggested that the relationship between plant species richness and remote sensing observations is location-specific, primarily due to differences in the assemblages of plant species among regions, which results from their unique biogeographic histories, climate conditions, and geographical features [20]. For example, statistical analysis showed that the Southern Tablelands and Snowy Mountains regions differ in their relationships between spectral data and plant species richness [13]. Therefore, geolocation context might be useful for mapping richness across regions. The location-dependent modeling with Spatioformer encoded such geolocational information into the input spectral data, which could help accommodate the location dependence in mapping richness over large spatial scales. The results shown in Table I, the visual comparison between Figs. 7(f) and 8, and the enlarged comparison maps shown in Fig. 9 indicated that it is beneficial to utilize geolocation context in large-scale richness mapping, with a higher overall accuracy being observed.

D. Cross-Year Average and Standard Deviation of Richness

The maps in Fig. 10 show the cross-year average and standard deviation of plant species richness in Australia, with the average map [see Fig. 10(a)] showing the spatially enhanced long-term distribution of richness in Australia and the standard deviation map [see Fig. 10(b)] indicating the cross-year stability of richness across the country.
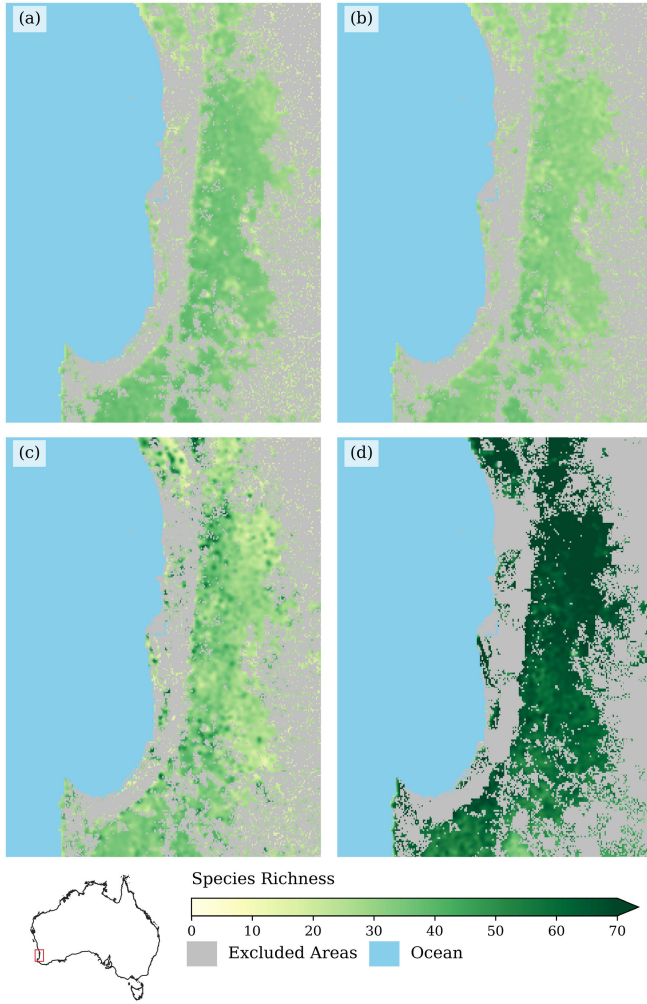
Fig. 9. Predicted plant species richness distribution in the Jarrah Forest bioregion near the southwestern coast of Western Australia with (a) CNN [58], (b) ViT [25], (c) FactoFormer [30], and (d) the proposed Spatioformer for the year 2020. Richness values are in the unit of the number of species per 400 m^2.
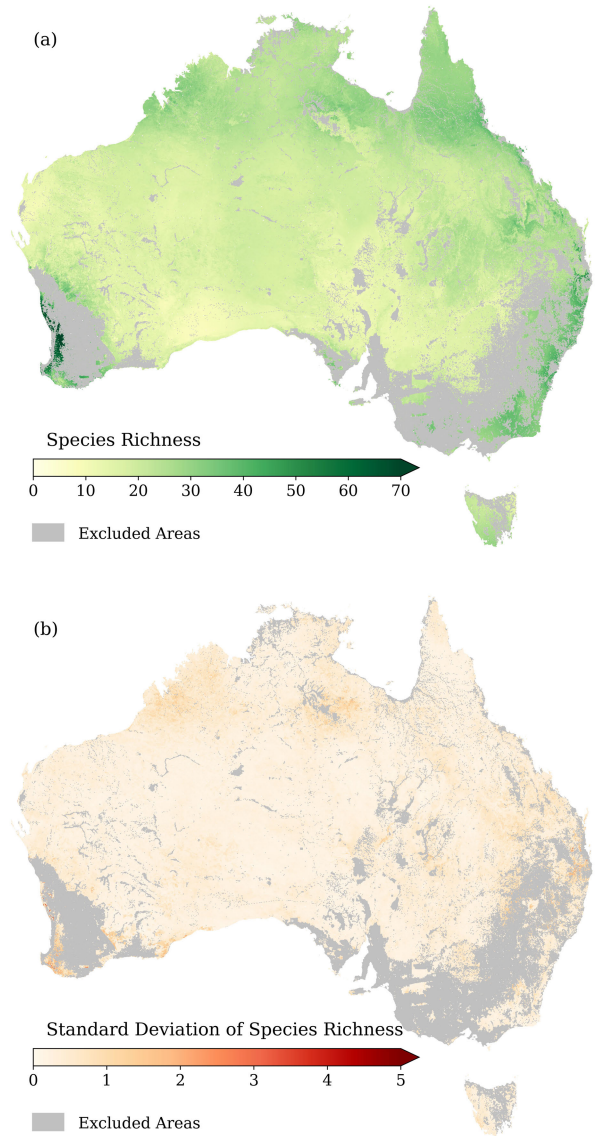

Fig. 10. (a) Average plant species richness map in Australia from 2015 to 2023. (b) Standard deviation map of richness across these years derived from annual richness maps compiled with the proposed Spatioformer model.

It was observed from Fig. 10(b) that the richness values were temporally stable with a standard deviation lower than one for most parts of the country, while higher cross-year variations were identified for several regions. Part of the southwestern, eastern, and southeastern coastal regions, as well as some areas in the northern savanna, showed a cross-year standard deviation in richness higher than one [see Fig. 10(b)]. Due to the relatively high biomass, these regions are susceptible to natural bushfires, such as the 2019–2020 Black Summer Bushfires. In addition to natural bushfires, controlled prescribed burning has been practiced in these regions aiming to reduce the risk of natural bushfires by intentionally burning excess flammable materials (e.g., dead wood). Floods, droughts, and human interference are also among the reasons that may contribute to cross-year variations in the richness levels.

With abrupt temporal variations and random noise being reduced, the average richness map in Fig. 10(a) provided a spatially enhanced reference for applications where the long-term distribution of richness needs to be taken into

consideration. The standard deviation of richness in Fig. 10(b) provided information on cross-year richness stability for different parts of the country, which could serve as a reference in conservation planning and practice.

E. Uncertainty in Richness Prediction

Fig. 11 shows the uncertainty maps for plant species richness predictions from 2015 to 2023 [see Fig. 11(a)–(f)] and the average uncertainty map across these years [see Fig. 11(g)]. By comparing the uncertainty maps with the distribution of in situ samples shown in Fig. 2, it was found that regions of denser sample coverage, such as the coastal regions, tended to have a lower prediction uncertainty than the interior of Australia with sparse or no coverage of in situ samples.
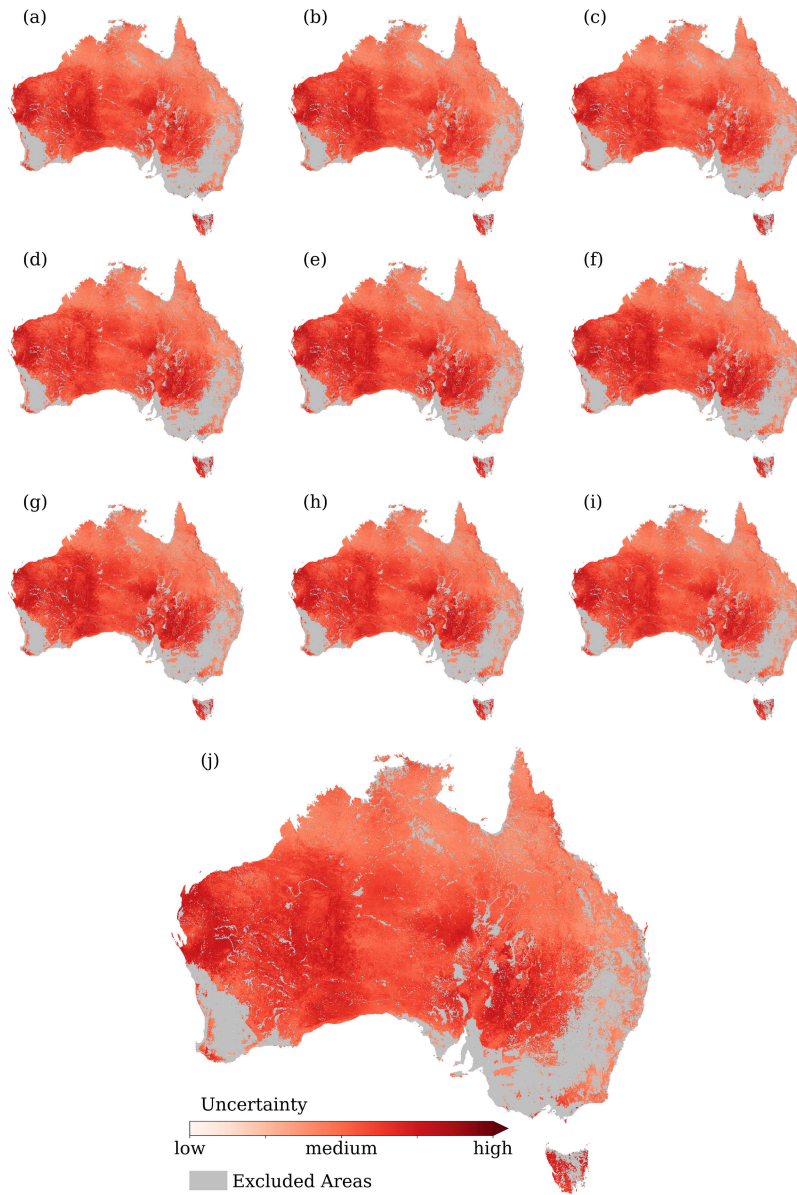
Fig. 11. Uncertainty maps for plant species richness predictions from 2015 to 2023 (a)–(i) and the average uncertainty map across these years (j). The uncertainty values were calculated with the Monte Carlo dropout approach [54] under a dropout rate of 0.5 for the proposed Spatioformer model.

It was seen from Fig. 11 that the eastern, southeastern, and northern coastal regions showed a lower uncertainty than the middle and mid-west interior of Australia, as well as the island of Tasmania. Compared with Fig. 2 where the locations of ground richness samples are displayed, the high-uncertainty regions shown in Fig. 11 mostly corresponded to places where fewer survey samples had been collected. Therefore, this uncertainty map could serve as a guidance as to where future field surveys should be focusing on. While conducting field experiments in remote locations could be considerably arduous, samples to be collected in high-uncertainty areas would be of high value in helping improve the overall accuracy of plant species richness maps throughout the country.

It is worth noting that several factors could contribute to the uncertainty in richness predictions. Areas with sparse ground sampling, such as the interior of Australia, exhibited higher prediction uncertainties, highlighting the model's sensitivity to the availability of ground-truth data. Moreover, the quality of in situ samples used for model training may affect the accuracy of the predictions. The model's performance is another critical factor, as its ability to generalize across diverse landscapes would also influence the reliability of the predicted richness values.

F. Impact of Input Image Size

The accuracies of plant species richness prediction achieved with different input image sizes from 1×1 to 9×9 pixels are shown in Table II. It was observed from the table that when the input image size increased from 1×1 to 9×9 pixels, improved mapping accuracies were observed, while the improvement became marginal with more pixels being

TABLE II
ACCURACIES OF THE PROPOSED SPATIOFORMER MODEL IN PLANT SPECIES RICHNESS MAPPING WITH INPUT IMAGE SIZE VARYING FROM 1 × 1 TO 9 × 9 PIXELS. EVALUATION METRICS INCLUDE COEFFICIENT OF CORRELATION (r), COEFFICIENT OF DETERMINATION (r^2), MAE, RAE, MSE, RSE, AND RMSE BETWEEN MODEL-PREDICTED AND GROUND-TRUTH SPECIES RICHNESS

Metric	Input Image Size				
	1 × 1 Pixel	3 × 3 Pixels	5 × 5 Pixels	7 × 7 Pixels	9 × 9 Pixels
r	0.66	0.71	0.74	0.76	0.77
r^2	0.44	0.50	0.55	0.58	0.59
MAE	8.62	8.27	8.04	7.93	7.83
RAE	0.31	0.30	0.29	0.29	0.29
MSE	165.38	123.65	111.88	107.74	105.86
RSE	0.13	0.12	0.12	0.11	0.11
RMSE	12.86	11.12	10.53	10.38	10.29

added especially when the input image size was larger than 5 × 5 pixels.

The results in Table II indicated that the spatial information, in combination with the spectral features provided by the imagery, could be beneficial in modeling the relationship between satellite observations and on-ground richness values. Considering that plants usually live as a community where each plant interacts with its spatial neighbors and surrounding habitat conditions in a close relationship, additional spatial information provided by the input image could better inform the model of landscape features. Supplying more pixels as input allowed the model to explore the spectral heterogeneity or variability among these pixels (i.e., the spatial variability in remotely sensed signals), considering that spectral diversity relates closely to biodiversity according to the spectral variation hypothesis proposed by [52]. A wider spatial range of the input image also allowed information on spatial heterogeneity and texture to be incorporated into the model.

G. Limitations and Future Work

In this study, we used Landsat Geomedian products to produce annual maps of plant species richness in Australia. A novel transformer architecture, Spatioformer, which is capable of taking in geo-coordinates, was proposed for the modeling in order to accommodate the location-dependent relationship between satellite observations and richness measurements. The Spatioformer model outperformed state-of-the-art models in richness prediction accuracy, as demonstrated in Section V-B. Nevertheless, there remains potential for enhancing the accuracy further, as discussed in the following.

This work used remote sensing imagery as the data source for plant species richness mapping. As demonstrated in [18] and [61], many environmental variables, such as temperature, precipitation, soil texture, and topographic heterogeneity, are closely related to the spatial pattern of richness though they are less effective than remote sensing data in capturing temporal changes such as those caused by deforestation, floods, bushfires, and land use intensification. Hence, in future studies, it is worth exploring the combination of environmental variables and remote sensing data to enhance the accuracy of plant species richness mapping.

Landsat's unique long-term record allowed us to cover a large number of historical survey samples, resulting in a

decent-sized dataset of richness versus image pairs for modeling. As shown in [13], both multispectral and hyperspectral satellite data demonstrated a reasonably strong correlation with on-ground plant species richness, with hyperspectral data showing better performance than multispectral. The deployment of a new generation of hyperspectral satellites, such as DESIS [62], PRISMA [63], and EnMAP [64], provides an opportunity to map plant species richness from hyperspectral data. However, it is important to note that the limited temporal and spatial coverage of these hyperspectral data may pose a major challenge in composing training datasets, as many historical survey samples would not be able to pair with a satellite observation.

In this study, plant species richness was predicted with spaceborne remote sensing data, which can be acquired at a lower cost than in situ data. However, we recognize that this method cannot fully replace in situ sampling. Field surveying is the gold-standard approach to richness measurement and provides ground-truth data for model validation. With remote sensing data as a proxy, our study aimed to extend the value of in situ samples. We first modeled the relationship between richness samples and satellite observations and then extrapolated the modeled relationship to locations where in situ samples are absent. In this way, knowledge of the survey samples is expanded to a wider spatial and temporal range, with satellite observations supplementing in situ measurements to achieve the spatiotemporal resolutions required for monitoring the richness dynamics.

While the Spatioformer model demonstrates good predictive capabilities, understanding the geographical and ecological mechanisms behind its predictions remains an important area for further investigation. In subsequent research, efforts could be made to improve the interpretability of the model by integrating domain knowledge from ecology and geography. Improved model interpretability would not only help explain the spatial and temporal patterns of plant species richness, which are related to environmental gradients, habitat fragmentation, and species aggregation, but also provide insights into the ecological processes that drive those patterns.

It is important to note that the relationship between plant species richness and remote sensing imagery could evolve with time due to the emergence of new influencing factors, such as those introduced by human activities and bushfires. Such factors may not be well represented in historical data

that our model has been trained on. In order to generalize the ability of our model under future conditions, it is essential to continuously update it by incorporating more recent data reflecting emerging influences. Potential techniques for model updating with new data, such as incremental learning, transfer learning, and model fine-tuning, can be explored in future studies to incorporate the influence of emerging factors.

As our study focuses on the Australian continent, the findings may be limited when generalizing to regions with different climatic, ecological, or biogeographical characteristics. A recent study [65] revealed that Australia is distinctive in a range of plant traits though its biotic and abiotic characteristics still share some degree of similarity with those in other regions of the world. Therefore, future research should explore our model's performance across different regions and biomes around the world by incorporating additional data sources that capture local ecological and environmental dynamics. Such efforts would enhance the generalizability of the model and its applicability for global biodiversity monitoring and conservation efforts.

Future work could also focus on extending the Spatioformer model for a broader range of applications. One potential direction is to adapt the model for other biodiversity metrics beyond plant species richness. For instance, the Spatioformer architecture could be modified to test its capability in predicting plant functional diversity [66] or spectral diversity [67], leveraging its ability to incorporate geolocation context into remote sensing data. Moreover, enhancing the model to integrate more complex environmental variables and considering temporal dynamics from long-term satellite observations could improve its applicability to diverse ecological and conservation challenges. Exploring these potential applications of Spatioformer would further assist biodiversity monitoring and management with large-scale remote sensing data.

VI. CONCLUSION

We presented a novel transformer architecture, *Spatioformer*, to map the spatial distribution of plant species richness in Australia from Landsat observations. The results demonstrated the feasibility of applying Spatioformer to a continental-scale in situ richness dataset (HAVPlot) for compiling large-scale richness maps. These maps derived from spaceborne remote sensing data offered a more comprehensive spatiotemporal representation of richness distribution compared with the traditional approach of measuring richness via ground sampling.

The Spatioformer model differs from standard transformer architectures in that it allows the encoding of geolocations into remote sensing images whose pixel values are intrinsically associated with geolocation coordinates. With Spatioformer, the location-dependent relationship between plant species richness and spectral features in satellite observations was accommodated in the modeling, with enhanced performance being observed relative to state-of-the-art models.

With Spatioformer, plant species richness maps over Australia were compiled from the Landsat archive for the years from 2015 to 2023. The richness maps produced in this study revealed the spatiotemporal dynamics of plant species richness in Australia. The resultant richness maps provided useful guidance for developing future conservation strategies and decision-making processes with the aim of preserving plant diversity in Australia. Through quantitative analyses, we identified regions where richness predictions are of high uncertainty. Future in situ surveys may focus on these areas in order to improve the overall accuracy of richness mapping. We also analyzed the impact of input image size on the accuracy of richness prediction, showing that incorporating spatially adjacent pixels into modeling is beneficial in richness modeling with satellite data.

This study provided a basis for further improving the accuracy of plant species mapping. In future studies, it is worthwhile to explore the integration of remote sensing data with environmental variables, the incorporation of seasonal information provided by satellite image time series, and the adoption of spaceborne hyperspectral data for richness modeling.

REFERENCES

[1] A. D. Chapman, "Numbers of living species in Australia and the world," Austral. Government Dept. Environ., Water, Heritage Arts Canberra, Canberra, ACT, Australia, Tech. Rep. 9780642568618, 2009.

[2] CHAH. (2022). *Australian Plant Census, Centre of Australian National Biodiversity Research. Council of Heads of Australasian Herbaria (CHAH)*. [Online]. Available: https://id.biodiversity.org.au/tree/51354547

[3] J.-B. Feret and G. P. Asner, "Tree species discrimination in tropical forests using airborne imaging spectroscopy," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 73–84, Jan. 2013.

[4] J. Yang et al., "Trajectories of plant nitrogen availability globally during 1984–2022 uncovered by satellite-derived nitrogen stable isotope ratio," Apr. 2023, doi: 10.21203/rs.3.rs-2843834/v1. [Online]. Available: https://www.researchsquare.com/article/rs-2843834/v1

[5] N. Li, S. Jiang, J. Xue, S. Ye, and S. Jia, "Texture-aware self-attention model for hyperspectral tree species classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502215.

[6] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501212.

[7] G. Ceballos, P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle, and T. M. Palmer, "Accelerated modern human–induced species losses: Entering the sixth mass extinction," *Sci. Adv.*, vol. 1, no. 5, Jun. 2015, Art. no. 1400253.

[8] F. Isbell et al., "Expert perspectives on global biodiversity loss and its drivers and impacts on people," *Frontiers Ecology Environ.*, vol. 21, no. 2, pp. 94–103, Jul. 2022.

[9] K. Mokany et al., "Reconciling global priorities for conserving biodiversity habitat," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 18, pp. 9906–9911, May 2020.

[10] R. E. Shaw et al., "Linking life history to landscape for threatened species conservation in a multiuse region," *Conservation Biol.*, vol. 37, no. 1, p. 13989, Feb. 2023.

[11] O. St-Laurent, K. Mokany, and L. J. Pollock, "Safeguarding eucalypt diversity through conservation-focused tree planting," *J. Appl. Ecology*, vol. 60, no. 3, pp. 519–529, Mar. 2023.

[12] W. Gould, "Remote sensing of vegetation, plant species richness, and regional biodiversity hotspots," *Ecological Appl.*, vol. 10, no. 6, pp. 1861–1870, Dec. 2000.

[13] Y. Guo, K. Mokany, C. Ong, P. Moghadam, S. Ferrier, and S. R. Levick, "Plant species richness prediction from DESIS hyperspectral data: A comparison study on feature extraction procedures and regression models," *ISPRS J. Photogramm. Remote Sens.*, vol. 196, pp. 120–133, Feb. 2023.

[14] E. Madigan, Y. Guo, M. Pickering, A. Held, and X. Jia, "Quantitative monitoring of complete rice growing seasons using sentinel 2 time series images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 7699–7702.

[15] Y. Guo, X. Jia, and D. Paull, "Effective sequential classifier training for SVM-based multitemporal remote sensing image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3036–3048, Jun. 2018.

[16] F. Zhao, W. Ma, J. Zhao, Y. Guo, M. Tariq, and J. Li, "Global retrieval of the spectrum of terrestrial chlorophyll fluorescence: First results with TROPOMI," *Remote Sens. Environ.*, vol. 300, Jan. 2024, Art. no. 113903.

[17] Y. Guo, X. Jia, D. Paull, and J. A. Benediktsson, "Nomination-favoured opinion pool for optical-SAR-synergistic rice mapping in face of weakened flooding signals," *ISPRS J. Photogramm. Remote Sens.*, vol. 155, pp. 187–205, Sep. 2019.

[18] K. Mokany et al., "Patterns and drivers of plant diversity across Australia," *Ecography*, vol. 2022, no. 11, p. 06426, Nov. 2022.

[19] A. M. Gill, "Biodiversity and bushfires: An Australia-wide perspective on plant-species changes after a fire event," *Australia's Biodiversity-Responses Fire, Plants, Birds Invertebrates. Environ. Aust. Biodiversity Tech. Paper*, vol. 1, pp. 9–53, Jul. 1999.

[20] M. L. Rosenzweig, *Species Diversity in Space and Time*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[21] S. Ferrier, G. Manion, J. Elith, and K. Richardson, "Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment," *Diversity Distrib.*, vol. 13, no. 3, pp. 252–264, May 2007.

[22] R. Wang et al., "Influence of species richness, evenness, and composition on optical diversity: A simulation study," *Remote Sens. Environ.*, vol. 211, pp. 218–228, Jun. 2018.

[23] K. Mokany, C. Ware, S. N. C. Woolley, S. Ferrier, and M. C. Fitzpatrick, "A working guide to harnessing generalized dissimilarity modelling for biodiversity analysis and conservation assessment," *Global Ecology Biogeography*, vol. 31, no. 4, pp. 802–821, Apr. 2022.

[24] Y. Guo, K. Mokany, C. Ong, P. Moghadam, S. Ferrier, and S. Levick, "Quantitative assessment of DESIS hyperspectral data for plant biodiversity estimation in Australia," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 1744–1747.

[25] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[26] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 551–561.

[27] M. Xu, F. Deng, S. Jia, X. Jia, and A. J. Plaza, "Attention mechanism-based generative adversarial networks for cloud removal in Landsat images," *Remote Sens. Environ.*, vol. 271, Mar. 2022, Art. no. 112902.

[28] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 2021 Int. Conf. Learn. Represent.*, 2021, pp. 1–7. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[29] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[30] S. Mohamed, M. Haghighat, T. Fernando, S. Sridharan, C. Fookes, and P. Moghadam, "FactoFormer: Factorized hyperspectral transformers with self-supervised pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5501614.

[31] K. Ayush et al., "Geography-aware self-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10181–10190.

[32] Y. Ge, X. Zhang, P. M. Atkinson, A. Stein, and L. Li, "Geoscience-aware deep learning: A new paradigm for remote sensing," *Sci. Remote Sens.*, vol. 5, Jun. 2022, Art. no. 100047.

[33] G. Mai et al., "A review of location encoding for GeoAI: Methods and applications," *Int. J. Geographical Inf. Sci.*, vol. 36, no. 4, pp. 639–673, Apr. 2022.

[34] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev, "Improving image classification with location context," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1008–1016.

[35] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2019–2026.

[36] S. Ardeshir, A. Zamir, A. Torroella, and M. Shah, "GIS-assisted object detection and geospatial localization," in *Proc. IEEE Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2014, pp. 602–617.

[37] K. Amlacher, G. Fritz, P. Luley, A. Almer, and L. Paletta, "Geo-contextual priors for attentive urban object recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 1214–1219.

[38] N. Lang, W. Jetz, K. Schindler, and J. D. Wegner, "A high-resolution canopy height model of the Earth," *Nature Ecology Evol.*, vol. 7, no. 11, pp. 1778–1789, Sep. 2023.

[39] G. Mai, K. Janowicz, B. Yan, R. Zhu, L. Cai, and N. Lao, "Multi-scale representation learning for spatial feature distributions using grid cells," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2020, pp. 1–9.

[40] G. Mai et al., "Sphere2Vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 439–462, Aug. 2023.

[41] K. Klemmer, N. Safir, and D. B. Neill, "Positional encoder graph neural networks for geographic data," in *Proc. Int. Conf. Artif. Intell. Statist.*, Jan. 2021, pp. 1379–1389.

[42] M. Rußwurm, K. Klemmer, E. Rolf, R. Zbinden, and D. Tuia, "Geographic location encoding with spherical harmonics and sinusoidal representation networks," in *Proc. 12th Int. Conf. Learn. Represent.*, Jan. 2023, pp. 1–11.

[43] K. Banerjee et al., "A machine-learning approach for prediction of water contamination using latitude, longitude, and elevation," *Water*, vol. 14, no. 5, p. 728, Feb. 2022.

[44] D. Joshi and J. Luo, "Inferring generic activities and events from image content and bags of geo-tags," in *Proc. Int. Conf. Content-Based Image Video Retr.*, Jul. 2008, pp. 37–46.

[45] *Catchment Scale Land Use of Australia—Update December 2020*, Austral. Bur. Agricult. Resource Econ., Canberra, ACT, Australia, 2021, doi: 10.25814/aqjw-rq15.

[46] K. Mokany et al., 2022, "Harmonised Australian vegetation plot dataset (HAVPlot). v4. CSIRO. data collection," doi: 10.25919/5cex-4s70.

[47] A. Lewis et al., "The Australian geoscience data cube—Foundations and lessons learned," *Remote Sens. Environ.*, vol. 202, pp. 276–292, Dec. 2017.

[48] N. Mueller. (2021). *DEA Geometric Median and Median Absolute Deviation (Landsat). Commonwealth of Australia (Geoscience Australia).* [Online]. Available: https://cmi.ga.gov.au/data-products/dea/645/dea-geometric-median-and-median-absolute-deviation-landsat

[49] D. Roberts, N. Mueller, and A. Mcintyre, "High-dimensional pixel composites from Earth observation time series," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6254–6264, Nov. 2017.

[50] D. Roberts, B. Dunn, and N. Mueller, "Open data cube products using high-dimensional statistics of time series," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 8647–8650.

[51] S. L. Ustin and J. A. Gamon, "Remote sensing of plant functional types," *New Phytologist*, vol. 186, no. 4, pp. 795–816, Jun. 2010.

[52] M. W. Palmer, P. G. Earls, B. W. Hoagland, P. S. White, and T. Wohlgemuth, "Quantitative tools for perfecting species lists," *Environmetrics*, vol. 13, no. 2, pp. 121–137, Mar. 2002.

[53] G. Ke, D. He, and T.-Y. Liu, "Rethinking positional encoding in language pre-training," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=09-528y2Fgf

[54] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[55] N. Karasiak, J.-F. Dejoux, C. Monteil, and D. Sheeren, "Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing," *Mach. Learn.*, vol. 111, no. 7, pp. 2715–2740, Jul. 2022.

[56] T. Kattenborn, F. Schiefer, J. Frey, H. Feilhauer, M. D. Mahecha, and C. F. Dormann, "Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks," *ISPRS Open J. Photogramm. Remote Sens.*, vol. 5, Aug. 2022, Art. no. 100018.

[57] P. Ploton et al., "Spatial validation reveals poor predictive performance of large-scale ecological mapping models," *Nature Commun.*, vol. 11, no. 1, p. 4540, Sep. 2020.

[58] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[59] C. Pickering, W. Hill, and K. Green, "Vascular plant diversity and climate change in the Alpine zone of the Snowy Mountains, Australia," *Biodiversity Conservation*, vol. 17, no. 7, pp. 1627–1644, Jun. 2008.

[60] M. Fallding, "A planning framework for natural ecosystems of the ACT and NSW southern tablelands," Natural Heritage Trust, NSW Nat. Parks Wildlife Service, Sydney, NSW, Australia, Tech. Rep. 0646419307, 2002.

[61] L. Cai et al., "Global models and predictions of plant diversity based on advanced machine learning techniques," *New Phytologist*, vol. 237, no. 4, pp. 1432–1445, Feb. 2023.

[62] A. Eckardt et al., "DESIS (DLR Earth sensing imaging spectrometer for the ISS-MUSES platform)," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1457–1459.

[63] S. Pignatti et al., "The PRISMA hyperspectral mission: Science activities and opportunities for agriculture and land monitoring," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2013, pp. 4558–4561.

[64] L. Guanter et al., "The EnMAP spaceborne imaging spectroscopy mission for Earth observation," *Remote Sens.*, vol. 7, no. 7, pp. 8830–8857, Jul. 2015.

[65] H. Flores-Moreno et al., "Is Australia weird? A cross-continental comparison of biological, geological and climatological features," *Frontiers Ecology Evol.*, vol. 11, May 2023, Art. no. 1073842.

[66] X. Ma et al., "Inferring plant functional diversity from space: The potential of Sentinel-2," *Remote Sens. Environ.*, vol. 233, Nov. 2019, Art. no. 111368.

[67] L. J. Williams et al., "Remote spectral detection of biodiversity effects on forest biomass," *Nature Ecology Evol.*, vol. 5, no. 1, pp. 46–54, Nov. 2020.

Yiqing Guo (Member, IEEE) received the Ph.D. degree from The University of New South Wales (UNSW), Canberra, ACT, Australia, in 2019.

He is currently a Research Scientist with the Commonwealth Scientific and Industrial Research Organization (CSIRO), Canberra. He served as the Inaugural Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) UNSW Canberra Student Chapter from 2018 to 2019. Since 2024, he has been serving as the Chair of the IEEE GRSS Australian Capital Territory and New South Wales Joint Chapter. His research interests include remote sensing and machine learning, and their applications to environmental problems.

Karel Mokany received the Ph.D. degree from The Australian National University, Canberra, ACT, Australia, in 2008.

He is currently a Principle Research Scientist with CSIRO, Acton, ACT, Australia. His research interests include biodiversity modeling and biodiversity assessment and ecology.


Shaun R. Levick received the Ph.D. degree in landscape ecology from the University of the Witwatersrand, Johannesburg, South Africa, in 2008.

He is currently a Principal Research Scientist with the Commonwealth Scientific and Industrial Research Organization (CSIRO), Darwin, NT, Australia, Australia's national science agency. He specializes in the development and application of remote sensing technologies to address environmental challenges. His research interests include savanna landscapes, particularly the improvement of sustainable land management and biodiversity conservation.

Jinyan Yang received B.Sc. degree from East China Normal University, Shanghai, China, in 2010, and the Ph.D. degree from Western Sydney University, Penrith, NSW, Australia, in 2019.

He is currently an Early Research Career Fellow with the Commonwealth Scientific and Industrial Research Organisation, Acton, ACT, Australia. His research interests include trait-based vegetation modeling, remote sensing, and model data synthesis.
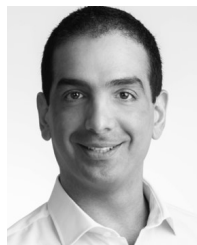
Peyman Moghadam (Senior Member, IEEE) is a Principal Research Scientist with CSIRO, DATA61, Brisbane, QLD, Australia, and a Professor (Adjunct) with the Queensland University of Technology (QUT), Brisbane. He leads the Embodied AI Research Cluster, CSIRO, DATA61, working at the intersection of robotics and machine learning. He is also the Spatiotemporal AI Portfolio Leader at CSIRO's Machine Learning and Artificial Intelligence (MLAI) Future Science Platform and oversees research and development of MLAI methods for scientific discovery in spatiotemporal data streams. In 2022, he was a Visiting Professor at ETH Zürich, Zürich, Switzerland. His research interests include embodied AI, robotics, and machine learning.