Tessellating The Earth

Daniel Cher, Hamza Iqbal, Eric Xing, Brian Wei, and Nathan Jacobs

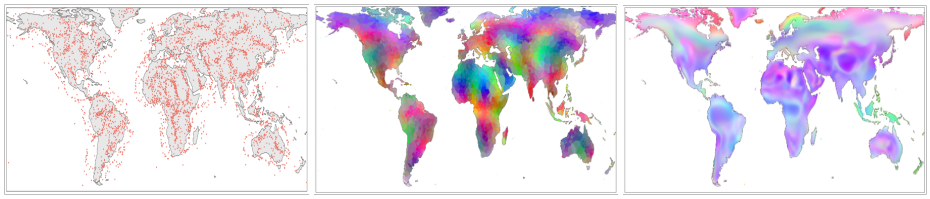
Washington University in St. Louis
{cher, h.iqbal, e.xing, b.j.wei, jacobsn}@wustl.edu

Abstract. Geolocation encoders, which map geographic coordinates to learned representations, are emerging as an effective means of capturing visual and non-visual characteristics from a latitude-longitude pair alone. However, existing approaches project coordinates onto fixed bases (e.g., spherical harmonics), allocating representational capacity uniformly and devoting equal resources to the open ocean and to a developing city. We introduce Tessellating the Earth (TTE), a location encoder built from learnable Spherical Voronoi partitions that concentrates representational capacity where it is needed in a fully differentiable, end-to-end manner. Each Voronoi site carries its own embedding and migrates during training toward discriminative areas. To bridge the gap between local spatial structure and global semantic understanding, we introduce *global semantic tokens*: a set of shared learnable concept tokens that distill semantic knowledge from the satellite imagery into a compact vocabulary the location encoder can reference at inference, enabling geographically distant sites covering similar environments to share semantics. TTE sets a new state of the art for location encoders across a suite of geospatial classification and regression tasks, and achieves the strongest results when used as a geographic prior for fine-grained species classification on iNaturalist-2018. Code, and weights are available at https://github.com/mvrl/TTE.

1 Introduction

Where something is on Earth says a great deal about what is there. Climate, ecology, land cover, and built infrastructure are all closely tied to location, and all are highly predictive for a wide range of geospatial tasks. Location encoders aim to capture these latent properties by learning mappings from latitude–longitude coordinates to high-dimensional representations [19]. These representations benefit applications including species distribution modeling [6,24], crop yield estimation [43], air quality forecasting [17], canopy height estimation [20], and satellite image synthesis [5, 39]. Once trained, they provide a compact representation of any location from coordinates alone, without requiring imagery at inference.

Recent work in building location encoders has focused on contrastive alignment between location and images. SatCLIP [18] aligns coordinates with satellite images, while GeoCLIP [45] uses ground-level photographs. In each case, a positional encoding projects the raw coordinates into a high-dimensional feature space, and a neural network maps these features into an embedding that is trained to match the co-located image embedding. The choice of positional



(a) Learned site positions (b) Site embeddings (ICA) (c) Location embeddings (ICA)

Fig. 1: Learned spatial decomposition. TTE partitions the sphere into learnable Voronoi regions that adapt to geographic complexity. (a) After training, sites concentrate on land masses, clustering densely along coastlines and ecological boundaries while leaving oceans sparsely covered. (b) Each site carries a learned embedding, with neighboring sites sharing similar embeddings (as indicated by color proximity), thereby forming a coherent spatial code. (c) The final location embeddings produced by TTE vary smoothly across space, reflecting underlying geographic and ecological structure.

encoding is critical. Without explicit high-frequency structure, neural networks default to smooth, low-frequency solutions [32], making raw coordinates insufficient for fine-grained spatial variation. Existing methods address this with fixed bases, whether spherical harmonics [18, 37], multi-scale Fourier features [26, 28], or random Fourier features [45]. These bases are mathematically convenient but share a fundamental limitation: their spatial structure is determined before training and cannot adapt to the data. Spherical harmonics, for instance, distribute capacity uniformly across the sphere, devoting as many basis functions to the open ocean as to a metropolitan city. As we show in Sec. 4.2, replacing a learned spatial decomposition with a fixed one significantly reduces performance, confirming that this limitation is not merely theoretical.

We propose to replace these fixed bases with a learnable spatial decomposition. A Voronoi tessellation divides a surface into regions based on proximity to a set of generating sites, a construction used widely across computational geometry [2, 29], climate modeling [35, 41], and neural scene representation [34]. *Tessellating the Earth* (TTE) encodes locations through Spherical Voronoi partitions [10] on the unit sphere using a set of learnable sites, each carrying its own embedding, whose soft assignment weights define how nearby coordinates are encoded (see Fig. 1 for visualization of the learned model). Both site positions and embeddings are learnable parameters, jointly optimized through contrastive pretraining with satellite imagery. The contrastive objective drives sites to arrange themselves such that the resulting location embeddings discriminate between visually distinct places (see Fig. 2 for an architecture overview).

Spatial partitioning alone does not address semantic understanding. Our Spherical Voronoi representation assigns each site a local embedding, but two sites covering tropical forest on different continents have no explicit pathway to share or align their representations. To address this, we observe that the pretrained image encoder contains rich semantic knowledge about land types, but this information is only available during training. We introduce *global semantic*

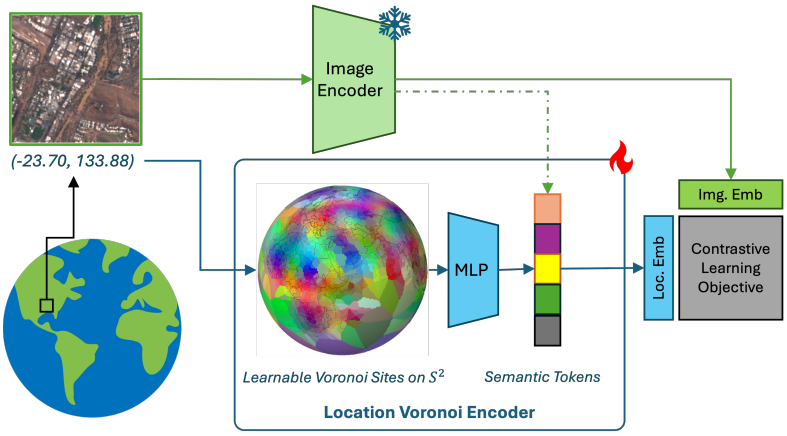


Fig. 2: Overview of TTE. The *location pathway* (bottom) maps a coordinate onto $\mathbb{S}^2$, where learnable Voronoi sites perform soft assignment; the resulting embedding attends over shared *semantic tokens* to produce the location embedding. The *image pathway* (top) encodes the co-located satellite image with a frozen ViT. The resulting image embedding enters the contrastive objective directly and supervises the tokens during training (dashed). At inference only the location pathway is required.

tokens: a set of shared learnable concept tokens that distill this visual semantic knowledge into a compact vocabulary the location encoder can reference at inference time, when no imagery is available (see token spatial coherence Fig. 3). This factors the problem into local spatial structure (site embeddings) and global semantic alignment (tokens), letting distant sites covering similar environments share capacity through common concepts.

Our contributions are:

- **Learnable Spherical Voronoi partitions for geographic encoding.** We introduce a location encoder whose spatial decomposition is learned end-to-end from visual supervision, replacing fixed bases with data-driven site placement. TTE sets a new state of the art among parametric location encoders across a suite of geospatial classification and regression benchmarks.
- **Global semantic tokens for cross-modal alignment.** We propose shared learnable concept tokens that distill semantic knowledge from the pretrained image encoder, bridging the representational gap between coordinates and imagery. We show that they learn coherent visual semantic themes, drive quantitative improvements, and achieve the strongest results as a geographic prior for fine-grained species classification on iNaturalist-2018.

2 Related Work

Location encoders and geo-visual pretraining. Geographic coordinate encoding has progressed from geometry-aware transforms (e.g., wrap encodings) [24] and

soft grid-based methods [49] through multi-scale positional encodings in $\mathbb{R}^n$ [26, 28] to bases defined directly on the sphere [25]. Rußwurm et al. couple spherical harmonics, an orthogonal basis on S^2, with sinusoidal representation networks (SIREN) [40], providing tunable resolution via the maximum harmonic degree [37]. Many of these encoders have been trained via contrastive alignment with geotagged imagery. SatCLIP applies CLIP-style pretraining on globally sampled Sentinel-2 imagery [18], GeoCLIP frames worldwide geo-localization as image–location retrieval using hierarchical random Fourier feature encodings [45], and CSP learns location representations by aligning a location encoder with a frozen image encoder on ground-level and overhead imagery [27]. These methods vary in imagery source and positional basis, but none adapt their spatial support to the underlying signal. A separate line of work addresses the limitations of contrastive alignment itself rather than the positional encoding. RANGE augments the location embeddings with image features retrieved from an external database at inference time [9]. While effective, this couples the encoder's performance to database coverage and introduces deployment constraints. Our goal is instead to improve the parametric location encoder itself, without external retrieval at test time.

Sphere-native representations and learnable partitions. Neural networks exhibit a well-documented spectral bias toward low frequencies [32], motivating Fourier feature mappings [42] and periodic activations [40] to lift coordinates into high-dimensional spaces before processing. On the sphere, this role is filled by spherical harmonics [37], multi-scale sinusoidal encodings [26,28], and random Fourier features [45]. Despite strong performance, these encoders fix their spatial support *a priori*, which introduces several limitations. Spherical harmonics distribute capacity uniformly across the sphere, including over oceans ($\sim$71% of the surface), and the number of coefficients grows quadratically with the maximum harmonic degree, making high-resolution representations expensive while still degrading near sharp spatial transitions [14, 37]. Cai and Balestriero [4] confirm that this uniformity leads to systematic performance disparities, with coastlines and small landmasses consistently underserved. Localized alternatives such as spherical Gaussians [46] offer compact support but can be sensitive to initialization and yield weak gradients when primitives are misaligned [10]. Concurrent work by Rao et al. [33] pursues localization differently, using Slepian functions as an analytically concentrated basis that achieves high resolution within a predefined region of interest. Both directions improve on uniform bases by focusing capacity locally, but each fixes that focus in advance: spherical Gaussians through their initial placement and Slepian functions through a region that must be specified before training, which limits the latter to localized prediction tasks. Di Sario et al. instead introduce differentiable Spherical Voronoi (SV) partitions with learnable sites and per-site temperatures, where softmax assignment provides non-zero gradients to every site and enables fully data-adaptive capacity allocation [10]. DeRF [34] earlier demonstrated the principle of learnable Voronoi decomposition for radiance fields, showing that learned partitions outperform fixed grids for heterogeneous scenes. TTE adapts SV from directional appear-

ance modeling to geographic representation learning and extends it with global semantic tokens for cross-modal pretraining.

Global semantic tokens. Several architectures use small sets of learnable tokens, not tied to any input, as shared communication channels for aggregation across local components. Many popular generalist vision architectures, from the original ViT [12] to modern VLM architectures [1, 3, 7, 16, 22] use attention mechanisms combined with learnable queries to distill sequence information. Darcet et al. show that adding explicit register tokens to Vision Transformers provides dedicated global memory and eliminates feature-map artifacts caused by repurposing input patches for implicit storage [8], while Set Transformer [21] uses inducing points that serve a similar role. MetaEmbed [48] extends learnable queries as a flexible way to scale late interaction for retrieval. TTE's global semantic tokens share the learnable-token mechanism but serve a distinct purpose. Rather than cross-attending within a single encoder, they act as shared latent anchors between two modalities, distilling semantic knowledge from the pretrained image encoder into a compact vocabulary that the location encoder can reference at inference time, enabling information exchange across distant Voronoi cells while preserving the locality benefits of a partition-based encoding.

3 Methodology

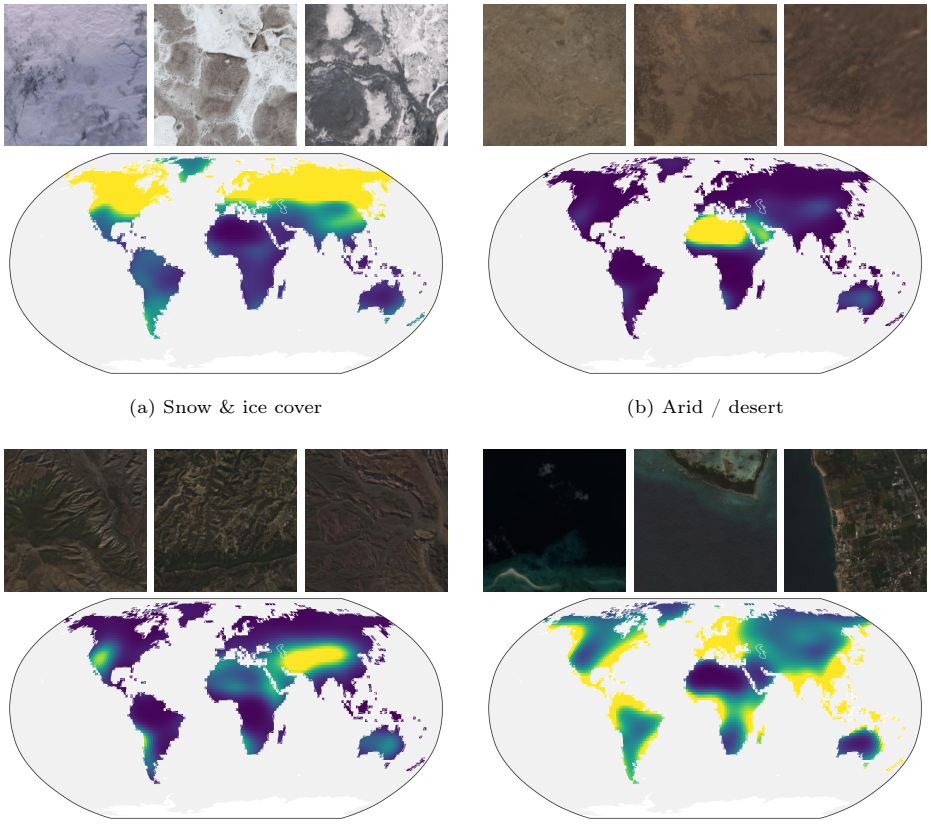
An overview of TTE is shown in Fig. 2, with additional details on the global semantic tokens provided in Fig. 4. The model has three components: a Spherical Voronoi encoder that maps coordinates to embeddings (Sec. 3.1), semantic tokens that enable global semantic sharing across localized sites (Sec. 3.2), and a set of training objectives that jointly optimize spatial and semantic representations (Sec. 3.3).

3.1 Preliminaries: Spherical Voronoi Partitions

We build on the differentiable Spherical Voronoi representation introduced by Di Sario *et al.* [10] for view-dependent appearance modeling. Given K sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_K\}$ on the unit sphere S^2, the soft assignment weight of a query point $\mathbf{x} \in S^2$ to site k is:

$$w_k(\mathbf{x}) = \frac{\exp(\tau_k \cdot (\mathbf{s}_k \cdot \mathbf{x}))}{\sum_{j=1}^{K} \exp(\tau_j \cdot (\mathbf{s}_j \cdot \mathbf{x}))} \tag{1}$$

where $\tau_k > 0$ is a per-site learnable temperature controlling partition sharpness. Since both $\mathbf{s}_k$ and $\mathbf{x}$ are unit vectors on S^2, their dot product equals the cosine of the angular distance between them, providing a natural proximity measure on the sphere. The softmax formulation ensures well-defined gradients for all sites and avoids competition between overlapping kernels that other spherical representations exhibit [10].



(a) Snow & ice cover

(b) Arid / desert

(c) Rocky terrain

(d) Coastal regions

Fig. 3: Learned global semantic tokens. Each sub-figure displays a specific semantic token with three top-attending satellite image patches (top) and the corresponding global spatial attention map (bottom). The visualization demonstrates that the model's internal tokens naturally specialize in distinct geographic and climatic features without explicit supervision.

We adapt this formulation for geographic encoding by replacing the scalar site values of [10] with learnable embedding vectors $\mathbf{e}_k \in \mathbb{R}^D$. The location embedding for a query $\mathbf{x}$ is the soft combination:

$$\mathbf{f}(\mathbf{x}) = \sum_{k=1}^{K} w_k(\mathbf{x}) \cdot \mathbf{e}_k \tag{2}$$

We use $K = 4{,}096$ sites, each producing a 384-dimensional embedding, which is then processed by a two-block residual MLP (linear projection to 512 dimensions, two residual blocks with ReLU and 0.5 dropout) to produce the final 512-dimensional location embedding $\mathbf{f}_{\text{out}}$. Sites are initialized on a Fibonacci lattice filtered to land areas (excluding Antarctica), with temperatures initialized to $\tau_k = 45$. All parameters, including site positions $\mathbf{s}_k$, temperatures τ_k, and embeddings $\mathbf{e}_k$, are jointly optimized [23] end-to-end. Positions are re-normalized to S^2 after each gradient step.

3.2 Global Semantic Tokens

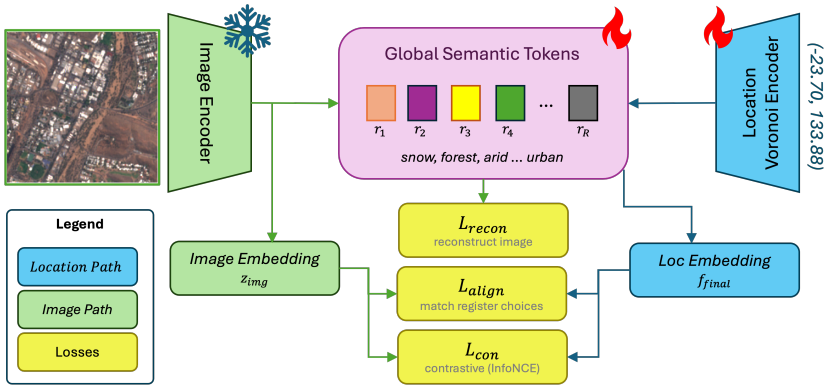


Fig. 4: Global semantic token mechanism. Both paths attend to shared learnable tokens. The image path (training only) produces peaked attention via a fixed low temperature; the location path learns to match this distribution. Three losses jointly optimize the system.

A visualization of the token alignment is shown in Fig. 4. Each Voronoi site carries an embedding that captures the representation of its local region, but sites are spatially isolated. A site covering tropical forest in South America has no mechanism to share meaning with a site covering tropical forest in Central Africa. The pretrained image encoder contains rich semantic knowledge about land types, but this information is only available during training. Global semantic tokens address both limitations by distilling visual semantics into a compact set

of learnable concept tokens that the location encoder can reference at inference time, providing a global layer of shared concepts that all sites can access.

We define R learnable concept tokens $\{\mathbf{r}_1, \ldots, \mathbf{r}_R\} \subset \mathbb{R}^{512}$, initialized as orthogonal unit vectors. Both the location encoder and image encoder compute attention distributions over these tokens.

Token attention. Both the location encoder and image encoder attend to the tokens through the same mechanism: a linear projection to R-dimensional logits followed by temperature-scaled softmax. For the location path:

$$\mathbf{a}_{\text{loc}} = \text{softmax}((\mathbf{W}_{\text{loc}}\,\mathbf{f}(\mathbf{x}) + \mathbf{b}_{\text{loc}})\,/\,T_{\text{loc}})\,, \quad \mathbf{z}_{\text{loc}} = \sum_{r=1}^{R} a_{\text{loc},r} \cdot \mathbf{r}_r \tag{3}$$

The image path is analogous, projecting image features $\mathbf{v}$ through a separate projection $\mathbf{W}_{\text{img}}$. The critical difference is temperature. The location path uses a learnable T_{loc}, while the image path uses a fixed low temperature $T_{\text{img}} = 0.05$ that produces peaked, nearly one-hot attention. Each image activates one or two tokens strongly, and the location path must learn to predict this peaked distribution from coordinates alone. This asymmetry provides a sharper alignment target than symmetric soft attention, where both paths tend toward diffuse distributions that weaken token specialization (Fig. 6).

The final location embedding is an equal-weight combination of the token-attended representation and a linear projection of the Voronoi embedding:

$$\mathbf{f}_{\text{final}} = \text{LayerNorm}((1-\alpha) \cdot \mathbf{W}\,\mathbf{f}(\mathbf{x}) + \alpha \cdot \mathbf{z}_{\text{loc}}) \tag{4}$$

where $\mathbf{W}$ is a linear projection that maps the Voronoi embedding to the output dimensionality and $\alpha = 0.5$.

3.3 Training Objective

The model is trained with losses that jointly optimize spatial encoding and semantic alignment.

Contrastive loss. The primary objective is a symmetric contrastive loss [18, 31] over batches of N co-located image-location pairs. Let $\mathbf{V}, \mathbf{F} \in \mathbb{R}^{N \times 512}$ denote the L2-normalized image and location embeddings, and τ_{logit} a learnable temperature. The logit matrix is $\mathbf{L} = \tau_{\text{logit}} \cdot \mathbf{V}\mathbf{F}^\top$, and the loss is:

$$\mathcal{L}_{\text{con}} = \tfrac{1}{2}\left[\text{CE}(\mathbf{L}, \mathbf{I}) + \text{CE}(\mathbf{L}^\top, \mathbf{I})\right] \tag{5}$$

where $\mathbf{I}$ is the identity matrix.

Reconstruction loss. The reconstruction loss requires that the tokens can reconstruct image features, preventing token collapse and ensuring the concept tokens form a useful basis covering the image feature space:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{z}_{\text{img},i} - \mathbf{v}_i\|^2 \tag{6}$$

where $\mathbf{z}_{\text{img},i} = \sum_r a^{(i)}_{\text{img},r} \cdot \mathbf{r}_r$ is the token-attended image representation.

Alignment loss. The alignment loss encourages the location attention distribution over tokens to match the image attention distribution via KL divergence, with stop-gradient on the image side:

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} \text{KL}(\mathbf{a}_{\text{loc},i} \,\|\, \text{sg}(\mathbf{a}_{\text{img},i})) \tag{7}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. Gradients flow only to the location path with image attention serving as a fixed target derived from visual content. The location encoder learns to predict which semantic concepts are present at a coordinate from the coordinate alone.

The total loss objective is:

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda_{\text{recon}} \cdot \mathcal{L}_{\text{recon}} + \lambda_{\text{align}} \cdot \mathcal{L}_{\text{align}} \tag{8}$$

with $\lambda_{\text{recon}} = 100$ and $\lambda_{\text{align}} = 0.1$. The high reconstruction weight ensures strong semantic association, while the lower alignment weight allows location attention to deviate when geographically appropriate.

4 Experiments

We evaluate TTE on an extensive geospatial benchmark suite covering classification and regression tasks, as well as fine-grained species classification on iNaturalist-2018 [44]. Classification tasks include Biome and EcoRegion classification using the WWF ecoregion delineations [11], and Country classification. Regression tasks include gridded air temperature [15], elevation and population density [36], and California Housing prices derived from 1990 census data [30]. Following prior work, we train a linear probe on top of inferenced coordinate embeddings for each downstream task and model. For contrastive pretraining, we follow the SatCLIP [18] protocol using globally sampled Sentinel-2 imagery, and use the same MoCo-pretrained image encoder [47] across all methods for fair comparison. In addition, we analyze inference-time augmentation strategies in the appendix.

4.1 Downstream Task Performance

Tab. 1 presents results across the geospatial benchmark suite. TTE leads on a majority of individual tasks, with the largest margins on Biome (+5.5 over TaxaBind), Country (+6.5 over SINR), and Elevation (+0.173 over SatCLIP). Fig. 5 visualizes the spatial biome predictions from a linear probe, where TTE produces more spatially coherent regions compared to other approaches. The two tasks where TTE does not lead reflect known data alignment advantages of other methods. On EcoRegion, TaxaBind performs best, with its location encoder aligned to

Table 1: Geospatial Benchmarks. Comparison across geospatial classification (accuracy) and regression (R^2) tasks. Classification tasks include Biome, EcoRegion and Country classification. Regression tasks include Temperature, Elevation, World Population and Cali-Housing estimation. Averages are computed within task type. Across N=5 training runs, TTE's results vary by at most ±0.3 (classification accuracy) and ±0.009 (R^2), well below the margins over competing methods.

	Classification				**Regression**				
Method	Bio	Eco	Ctry	Cls ↑	Temp	Elev	Pop	Cali	Reg ↑
Direct	29.1	0.6	66.9	32.2	0.381	0.025	0.053	0.238	0.174
Cartesian 3D	30.2	1.8	66.9	33.0	0.362	0.030	0.162	0.240	0.199
Wrap [24]	34.4	1.1	69.7	35.1	0.861	0.085	0.328	0.239	0.378
Theory [13]	33.5	1.0	72.5	35.7	0.849	0.093	0.330	0.254	0.382
SphereM [28]	36.4	27.3	72.7	45.5	0.629	0.139	0.302	0.423	0.373
SphereM+ [28]	58.7	50.1	76.1	61.6	0.886	0.294	0.421	0.543	0.536
SphereC [28]	36.3	52.9	72.9	54.0	0.461	0.185	0.335	0.496	0.369
SphereC+ [28]	53.2	61.6	73.6	62.8	0.842	0.260	0.392	0.544	0.510
CSP-INat [27]	61.1	57.1	75.9	64.7	0.717	0.388	0.554	0.462	0.530
CSP-FMoW [27]	61.4	58.0	81.3	66.9	0.865	0.399	0.580	0.541	0.596
SINR [6]	67.9	54.9	*88.3*	70.4	*0.942*	0.644	*0.726*	0.420	0.683
GeoCLIP [45]	70.2	*71.6*	81.3	74.4	0.916	0.604	0.698	**0.708**	*0.732*
SatCLIP [18]	68.9	69.3	82.8	73.7	0.825	*0.666*	0.684	0.400	0.644
TaxaBind [38]	*72.3*	**72.9**	86.4	*77.2*	0.897	0.581	0.712	*0.684*	0.719
TTE	**77.8**	67.4	**94.8**	**80.0**	**0.946**	**0.839**	**0.790**	0.532	**0.777**

iNaturalist species observations and WorldClim bioclimatic variables that provide training signal directly matched to ecological boundaries. The California Housing uses 1990 census data, which creates a temporal mismatch between target labels and contemporary satellite imagery disproportionately negatively affecting encoders trained on Sentinel-2 [9], while methods trained on ground-level photographs with dense coverage of Western countries encode street-level cues more aligned with housing value. Despite these task-specific gaps, TTE achieves the highest average classification accuracy (80.0%) and regression R^2 (0.777) among all parametric methods, improving over the previous best by 2.8% and 0.045 respectively.

Tab. 2 evaluates whether self-supervised location representations generalize as a geographic prior for fine-grained species classification on iNaturalist-2018 [44]. Following [24], let y denote the species label, I the image, and G the geographic location. We train a linear model on pre-trained geo-embeddings to produce a location-conditioned species distribution $P(y \mid G)$, and combine this with a pretrained image classifier's predictions $P(y \mid I)$ via $P(y \mid I, G) \propto P(y \mid I) \cdot P(y \mid G)$, under the conditional independence assumption of Mac Aodha *et al.* [24]. TTE achieves the best Top-k accuracy across all k, improving Top-1 from 66.1% (image only) to 76.2% and outperforming SatCLIP (75.1) and GeoCLIP (72.9). Notably, it also outperforms TaxaBind, whose location encoder is aligned with species data and trained on iNaturalist imagery. iNaturalist con-

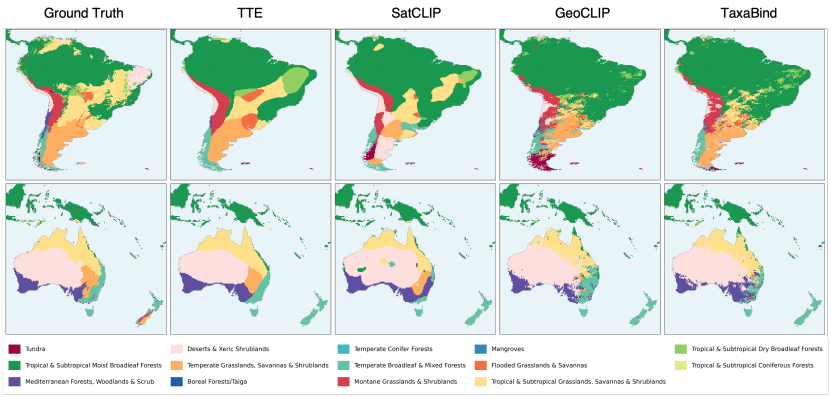


Fig. 5: Biome classification predictions. Ground truth WWF biome labels (left) with predictions from a host of location encoders. We used a linear probe on each encoder's embeddings. TTE produces more spatially coherent predictions with smoother biome boundaries, while other models show more fragmentation within homogeneous regions.

Table 2: Geo-prior Benchmarking. Top-k classification accuracy (%) on the iNaturalist-2018 test split. Incorporating geographic priors consistently improves fine-grained species classification over the image-only baseline. Our method (TTE) achieves the best performance across all k.

	Top-1	Top-3	Top-5	Top-10
Img	66.1	83.3	88.0	92.2
Img + CSP	72.9	87.9	91.6	94.8
Img + GeoCLIP	72.9	88.2	91.9	95.2
Img + CSP INat*	74.4	88.8	92.2	94.9
Img + SatCLIP	75.1	88.7	91.9	94.5
Img + TaxaBind	75.1	*89.7*	*93.0*	*95.6*
Img + TTE	**76.2**	**90.0**	**93.2**	**95.7**

tains over 8,000 species classes whose distributions are governed by fine-grained environmental variation. With this many classes, even modest improvements in the geographic prior translate to meaningful gains in species-level predictions, as the prior effectively reweights the classifier's output across thousands of candidate species. The strong performance confirms that TTE's embeddings capture semantically rich geographic structure that transfers beyond the tasks seen during pretraining.

4.2 Ablation Study

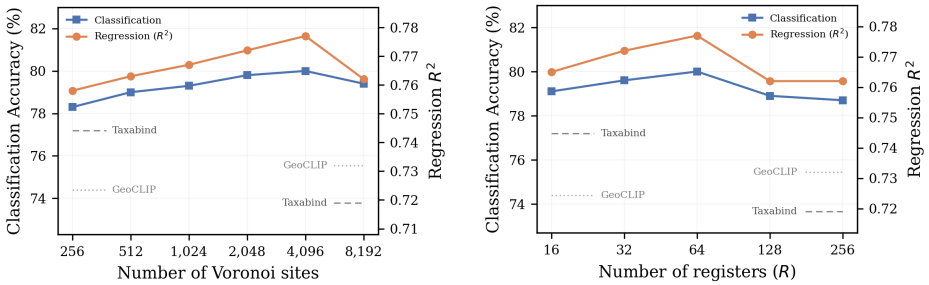
Component ablations. Fig. 6 isolates the contribution of each component. Fixing sites at their initial locations yields the largest decline among ablations (-12.9% classification, -0.125 regression), falling below SatCLIP on several tasks. This

suggests that the ability of sites to migrate during training is critical, reinforcing that uniform spatial allocation is a practical bottleneck for location encoders, not merely a theoretical concern.

Among token ablations, removing the entire semantic branch (*w/o tokens*) has the largest effect (−3.7% classification, −0.038 regression), with the steepest losses on Biome, EcoRegion, and Elevation. Removing individual losses produces smaller drops. *w/o* $\mathcal{L}_{align}$ allows the location path to leverage semantic content but selects tokens less reliably without explicit alignment supervision, with classification affected more than regression. *w/o* $\mathcal{L}_{recon}$ shows the smallest degradation overall, mainly on EcoRegion (−1.7) and Housing (−0.033), suggesting that the reconstruction loss contributes most when fine-grained token specialization is required. Symmetric soft attention (*sym. soft att.*) degrades across all tasks, indicating that peaked image-side attention provides a more effective alignment target than soft distributions on both paths.

Site and token count. Fig. 6 (right) shows sensitivity to site and token counts. Classification and regression both improve with site count and saturate near 4,096 sites. Doubling to 8,192 provides no further gain and slightly degrades regression. Token count peaks at $R = 64$ and declines for larger values. We hypothesize that too few tokens under-specify the semantic vocabulary, while too many dilute the alignment signal. We use 4,096 sites and 64 tokens in all subsequent experiments. We also ablate different site initialization strategies, and find that across our suite of geospatial tasks there is minimal difference (results in appendix).

Fig. 6: Ablation study. *Top:* performance vs. Voronoi site count (left) and token count R (right). Performance peaks at 4,096 sites and $R = 64$ tokens. *Bottom:* component ablations. Each row removes or modifies one component from the full model.



	Bio	Eco	Ctry	Cls ↑	Temp	Elev	Pop	Cali	Reg ↑
TTE (full)	**77.8**	**67.4**	**94.8**	**80.0**	**0.946**	**0.839**	**0.790**	**0.532**	**0.777**
Site ablations									
fixed sites	66.6	45.4	89.1	67.1	0.936	0.644	0.727	0.299	0.652
Token ablations									
w/o tokens	74.2	62.2	92.6	76.3	0.950	0.781	0.773	0.448	0.739
w/o $\mathcal{L}_{\text{align}}$	75.5	65.3	94.0	78.3	0.950	0.822	0.784	0.496	0.763
w/o $\mathcal{L}_{\text{recon}}$	76.3	65.7	94.5	78.8	0.952	0.828	0.784	0.499	0.766
sym. soft att. (T=0.2)	76.0	64.2	93.2	77.8	0.952	0.794	0.776	0.484	0.752

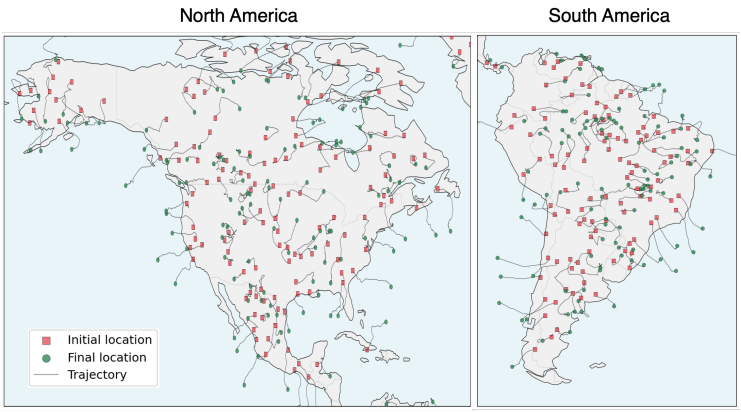


Fig. 7: Dynamics of Site Migration. To maintain visual clarity, we visualize the training trajectories of a random 30% subset of Voronoi sites. Markers indicate initial positions (■) and converged positions (●). Sites dynamically move toward geographic boundaries. Notably, many migrate offshore to specialize on narrow coastal segments, providing higher-resolution representational capacity where geographic features are densest.

4.3 Embedding Analysis

Learned token semantics. Fig. 3 shows the token attention visualization for a few example tokens. Each panel displays the spatial attention map (which locations activate this token most strongly) alongside the satellite images with the highest attention scores. Individual tokens learn to specialize in semantically coherent environmental concepts. We observe tokens corresponding to snow and ice cover, arid desert, rocky terrain, and other visually distinct land types. The full set of token visualizations appears in the appendix.

Learned tessellation. Fig. 1 visualizes the learned spatial decomposition at three stages of the encoding pipeline. Panel (a) plots the positions of all 4,096 sites after training where we observe sites concentrating on land masses, clustering near coastlines and areas of discrimination while leaving oceans sparsely covered. Panel (b) assigns each land pixel to its nearest site and colors it by the ICA projection of that site's embedding vector, revealing that neighboring sites learn similar representations and form a coherent spatial code. Panel (c) shows the final location embeddings produced by the full model, including soft assignment and semantic token fusion, visualized via ICA. We see that embeddings vary smoothly across space. The non-uniform allocation in all three views emerges entirely from the contrastive training signal, without explicit supervision on site placement.

Site migration. Fig. 7 visualizes site displacement over training for two regions. Sites are initialized on a Fibonacci lattice (red squares) and migrate during

contrastive training to their converged positions (green circles), with trajectories shown as connecting lines. In North America, sites move from interior positions toward the coasts, the Appalachians, and the US-Mexico border region. In South America, sites converge along the boundaries of the Amazon Basin, the Andes, and the coastline. In both regions, sites in homogeneous interiors undergo little displacement, whereas those near geographic or ecological transitions migrate substantially, often moving offshore to specialize on narrow coastal segments rather than splitting capacity between the coast and the interior. We present additional site maps for different regions of the world in the appendix.

5 Conclusion

We introduced TTE, a location encoder that replaces fixed positional bases with learnable Spherical Voronoi partitions augmented by global semantic tokens. The Voronoi backbone provides an adaptive spatial decomposition with sites migrating during training to concentrate representational capacity in discriminative areas, while leaving homogeneous areas sparsely covered. Global semantic tokens complement this local spatial structure with a global semantic layer, distilling visual knowledge from the co-aligned satellite imagery into a compact vocabulary that enables geographically distant sites covering similar environments to share meaning through a common set of learned concept tokens. TTE sets a new state of the art among parametric location encoders, improving average classification accuracy by 2.8% and average regression R^2 by 0.045 over the previous best on a suite of geospatial benchmarks, while achieving the strongest results as a geographic prior for fine-grained species classification on iNaturalist-2018. Beyond the quantitative results, the qualitative analysis reveals that TTE learns intuitive global semantic tokens specializing in coherent environmental concepts such as snow cover, arid desert, rocky terrain, and coastal areas. Voronoi sites discover a non-obvious geometric strategy, migrating offshore to dedicate their receptive fields to narrow coastal segments rather than splitting capacity between coast and interior. These behaviors emerge entirely from the contrastive training signal, suggesting that the combination of adaptive spatial partitioning and shared semantic anchors provides a rich inductive bias for geographic representation learning.

Several directions remain open. TTE currently trains on Sentinel-2 imagery without conditioning on time. Integrating multi-temporal observations would enable temporal knowledge accumulation, addressing limitations on temporally sensitive tasks such as housing price prediction. Incorporating additional modalities, such as ground-level photographs or environmental attributes could further improve performance on tasks that would benefit from them, such as fine-grained ecological classification. Finally, the Voronoi-plus-tokens architecture is not specific to geographic encoding. Any domain in which a function on the sphere exhibits non-uniform complexity, from planetary science to molecular surface modeling, could benefit from the same principle: learnable, adaptive partitioning combined with a shared semantic structure.

Acknowledgments

This research used the TGI RAILs advanced compute and data resource, which is supported by the National Science Foundation (award OAC-2232860) and the Taylor Geospatial Institute.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022) 5
2. Aurenhammer, F.: Voronoi diagrams—a survey of a fundamental geometric data structure. ACM computing surveys (CSUR) **23**(3), 345–405 (1991) 2
3. Bai, S., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 5
4. Cai, D., Balestriero, R.: No location left behind: Measuring and improving the fairness of implicit representations for earth data. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025) 4
5. Cher, D., Wei, B., Sastry, S., Jacobs, N.: Vectorsynth: Fine-grained satellite image synthesis with structured semantics. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7019–7029 (March 2026) 1
6. Cole, E., Van Horn, G., Lange, C., Shepard, A., Leary, P., Perona, P., Loarie, S., Mac Aodha, O.: Spatial implicit neural representations for global-scale species mapping. In: International conference on machine learning. pp. 6320–6342. PMLR (2023) 1, 10
7. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023) 5
8. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: International Conference on Learning Representations (ICLR) (2024) 5
9. Dhakal, A., Sastry, S., Khanal, S., Ahmad, A., Xing, E., Jacobs, N.: Range: Retrieval augmented neural fields for multi-resolution geo-embeddings. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24680–24689 (2025) 4, 10
10. Di Sario, F., Rebain, D., Verbin, D., Grangetto, M., Tagliasacchi, A.: Spherical voronoi: Directional appearance as a differentiable partition of the sphere. arXiv preprint arXiv:2512.14180 (2025) 2, 4, 5, 7
11. Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., et al.: An ecoregion-based approach to protecting half the terrestrial realm. BioScience **67**(6), 534–545 (2017) 9
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 5

13. Gao, R., Xie, J., Zhu, S.C., Wu, Y.N.: Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. In: International Conference on Learning Representations (ICLR) (2019) 10
14. Gelb, A.: The resolution of the gibbs phenomenon for spherical harmonics. Mathematics of Computation **66**(218), 699–717 (1997). https://doi.org/10.1090/S0025-5718-97-00828-4 4
15. Hooker, J., Duveiller, G., Cescatti, A.: A global dataset of air temperature derived from satellite remote sensing and weather stations. Scientific Data **5**(1) (2018) 9
16. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021) 5
17. Karimzadeh, M., Wang, Z., Crooks, J.L.: Performance and generalizability impacts of incorporating location encoders into deep learning for dynamic pm2.5 estimation. arXiv preprint arXiv:2505.18461 (May 2025), https://arxiv.org/abs/2505.18461 1
18. Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rußwurm, M.: Satclip: Global, general-purpose location embeddings with satellite imagery. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4347–4355 (2025) 1, 2, 4, 8, 9, 10
19. Klemmer, K., Rolf, E., Rußwurm, M., Camps-Valls, G., et al.: Earth embeddings: Towards ai-centric representations of our planet. EarthArXiv (2025) 1
20. Lang, N., Jetz, W., Schindler, K., Wegner, J.D.: A high-resolution canopy height model of the earth. Nature Ecology & Evolution **7**(11), 1778–1789 (2023) 1
21. Lee, J., Lee, Y., Kim, J., Kosiorek, A.R., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 97, pp. 3744–3753. PMLR (2019) 5
22. Li, J., Feng, D., Li, D., Fan, D., Chang, X., Tan, X., Chao, W., Lu, Y., Zhou, J., Batra, D., Parikh, D., Girshick, R.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20076–20086 (2023). https://doi.org/10.1109/CVPR56525.2023.01974 5
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 7
24. Mac Aodha, O., Cole, E., Perona, P.: Presence-only geographical priors for fine-grained image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 1, 3, 10
25. Mai, G., Janowicz, K., Hu, Y., Gao, S., Yan, B., Zhu, R., Cai, L., Lao, N.: A review of location encoding for geoai: Methods and applications. International Journal of Geographical Information Science **36**(4), 639–673 (2022) 4
26. Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., Lao, N.: Multi-scale representation learning for spatial feature distributions using grid cells. In: The Eighth International Conference on Learning Representations (ICLR) (2020) 2, 4
27. Mai, G., Lao, N., He, Y., Song, J., Ermon, S.: Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In: Proceedings of the International Conference on Machine Learning (ICML) (2023) 4, 10
28. Mai, G., Xuan, Y., Zuo, W., He, Y., Song, J., Ermon, S., Janowicz, K., Lao, N.: Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. ISPRS Journal of Photogrammetry and Remote Sensing **202**, 439–462 (2023) 2, 4, 10

29. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. Wiley Series in Probability and Statistics, John Wiley and Sons, 2nd edn. (2000) 2
30. Pace, R.K., Barry, R.: Sparse spatial autoregressions. Statistics & Probability Letters **33**(3), 291–297 (1997) 9
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021) 8
32. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F.A., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: ICML. pp. 5301–5310 (2019) 2, 4
33. Rao, A., Crasto, R., Ooms, T., Rolnick, D., Klemmer, K., Rußwurm, M.: Localized, high-resolution geographic representations with slepian functions (2026) 4
34. Rebain, D., Jiang, W., Yazdani, S., Li, K., Yi, K.M., Tagliasacchi, A.: DeRF: Decomposed radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14153–14161 (2021) 2, 4
35. Ringler, T., Ju, L., Gunzburger, M.: A multiresolution method for climate system modeling: application of spherical centroidal Voronoi tessellations. Ocean Dynamics **58**(5–6), 475–498 (2008). https://doi.org/10.1007/s10236-008-0157-2 2
36. Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., Hsiang, S.: A generalizable and accessible approach to machine learning with global satellite imagery. Nature Communications **12**(1), 4392 (2021) 9
37. Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., Tuia, D.: Geographic location encoding with spherical harmonics and sinusoidal representation networks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024) 2, 4
38. Sastry, S., Khanal, S., Dhakal, A., Ahmad, A., Jacobs, N.: Taxabind: A unified embedding space for ecological applications. In: Winter Conference on Applications of Computer Vision. IEEE/CVF (2025) 10
39. Sastry, S., Khanal, S., Dhakal, A., Jacobs, N.: Geosynth: Contextually-aware high-resolution satellite image synthesis. In: IEEE/ISPRS Workshop: Large Scale Computer Vision for Remote Sensing (EARTHVISION) (2024) 1
40. Sitzmann, V., Martel, J.N., Bergman, A.W., Lindell, D.B., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: NeurIPS. vol. 33, pp. 7462–7473 (2020) 4
41. Skamarock, W.C., Klemp, J.B., Duda, M.G., Fowler, L.D., Park, S.H., Ringler, T.D.: A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tesselations and C-grid staggering. Monthly Weather Review **140**(9), 3090–3105 (2012). https://doi.org/10.1175/MWR-D-11-00215.1 2
42. Tancik, M., Srinivasan, P.P., Mildenhall, B., et al.: Fourier features let networks learn high frequency functions in low dimensional domains. In: NeurIPS. vol. 33, pp. 7537–7547 (2020) 4
43. Tseng, G., Fuller, A., Reil, M., Herzog, H., Beukema, P., Bastani, F., Green, J.R., Shelhamer, E., Kerner, H., Rolnick, D.: Galileo: Learning global and local features in pretrained remote sensing models. arXiv e-prints pp. arXiv–2502 (2025) 1
44. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: CVPR. pp. 8769–8778 (2018) 9, 10

45. Vivanco Cepeda, V., Nayak, G.K., Shah, M.: Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. Advances in Neural Information Processing Systems **36**, 8690–8701 (2023) 1, 2, 4, 10
46. Wang, J., Ren, P., Gong, M., Snyder, J., Guo, B.: All-frequency rendering of dynamic, spatially-varying reflectance. In: Proceedings of the ACM SIGGRAPH Conference on Computer Graphics. pp. 133:1–133:10 (2009). https://doi.org/10.1145/1576246.1576363 4
47. Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X.: Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. arXiv preprint arXiv:2211.07044 (2022) 9
48. Xiao, Z., Ma, Q., Gu, M., Chen, C.c.J., Chen, X., Ordonez, V., Mohan, V.: Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. arXiv preprint arXiv:2509.18095 (2025) 5
49. Yin, Y., Liu, Z., Zhang, Y., Wang, S., Shah, R.R., Zimmermann, R.: Gps2vec: Towards generating worldwide gps embeddings. In: ACM SIGSPATIAL (2019) 4