Geolocation-Aware Pretraining Strategies for Globally Applicable Remote Sensing Foundation Models

Mojgan Madadikhaljan, Jonathan Prexl, Michael Schmitt

Department of Aerospace Engineering, University of the Bundeswehr Munich, Neubiberg, Germany
(mojgan.madadikhaljan, jonathan.prexl, michael.schmitt)@unibw.de

Keywords: Remote Sensing, Global Foundation Models, Local Foundation Models, Region Awareness.

Abstract

Foundation models have achieved remarkable success across various domains due to their ability to learn generalizable representations from large-scale, unlabeled datasets. In the geospatial domain, several foundation models have been developed to leverage the abundance of unlabeled remote sensing data and support Earth observation tasks across diverse regions and sensor types. However, the geolocation-dependent characteristics of remote sensing data introduce unique challenges in adapting these models to region-focused applications. By conducting a comprehensive empirical analysis across diverse geographical regions and tasks, we explore whether incorporating regional information during pretraining or fine-tuning improves performance on region-specific downstream tasks. We show that regional representation learning, as well as regional adaptation of features extracted from a globally trained foundation model, is beneficial when the region-specific performance of the downstream tasks is of interest. To this end, we also propose a regional adaptation to the globally trained foundation models to balance global diversity with regional representation learning for improved performance.

1. Introduction

Foundation models (FMs) have emerged as a transformative paradigm in machine learning, enabling significant advances in different domains, including natural language processing and computer vision. Using large-scale pretraining on vast and diverse datasets, these models capture generalizable representations that can be adapted to a wide range of downstream tasks. The capabilities of FMs, combined with the abundance of globally available, unlabeled remote sensing (RS) data, have driven the development of several geospatial foundation models (GFMs) - such as TerraMind, Prithvi, ScaleMAE, and SatMAE (Jakubik et al., 2025, Jakubik et al., 2023, Reed et al., 2023, Cong et al., 2022) - aimed at improving scalability and transferability across geographic regions, sensor modalities, and Earth observation tasks.

GFMs are designed to be highly generalizable and are often trained on globally distributed RS data. However, the content and characteristics of such data are inherently geolocation-dependent, with geographic variability often leading to shifts in data distributions (Roscher et al., 2024, Ekim et al., 2025, Madadikhaljan and Schmitt, 2025). At the same time, geolocation information is consistently available as metadata in RS datasets. Given these factors, we believe that it is important to investigate how region-specific representations in the pretraining influence the region-specific performance of FMs after the fine-tuning. We refer to this perspective as geolocation awareness.

While several geospatial foundation models have recently been introduced, to the best of our knowledge, there are no widely adopted or publicly released **regional GFMs** (i.e. pre-trained only on a specific geographical region). Considering this, and to retain full control over pretraining setup, we opted to pretrain our own model. This allowed us to fix the number of training samples, explicitly define the geolocations involved, and ensure consistency and fairness across regions.

When pretraining a foundation model (FM) to be used for downstream applications on a specific region, there are two possible scenarios for the pretraining: a) global pretraining and b) regional pretraining, which only contains data from the region of interest where the fine-tuned model will be employed. During the second stage (i.e. fine-tuning for a certain downstream task), a similar choice can be made by utilizing data from all regions or tailoring it to a specific region of interest. Given the designs described above, there exist three meaningful combinations:

- Global pretraining, global fine-tuning. This setup is most common in current remote sensing foundation models (RSFM) research since both the pretraining and benchmark datasets are globally distributed.
- Global pretraining, regional fine-tuning. By doing so, there will be a greater focus on regional fine-tuning of the model, tailoring the model's performance to be more suited to the region of interest.
- Regional pretraining, regional fine-tuning. A strong regional focus during the representation learning and fine-tuning of the model. This setup has the strongest focus on regional characteristics of the data.

In this paper, we investigate the impact of different geolocation-aware pretraining and fine-tuning strategies on the regional performance of geospatial FMs. We systematically explore different combinations of global and regional pretraining and fine-tuning strategies to identify setups that best support region-focused downstream performance without compromising generalization. Our goal is to provide insights into how FMs can be evaluated and tailored for geospatial applications, balancing global diversity with regional specificity.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-849-2026 | © Author(s) 2026. CC BY 4.0 License.

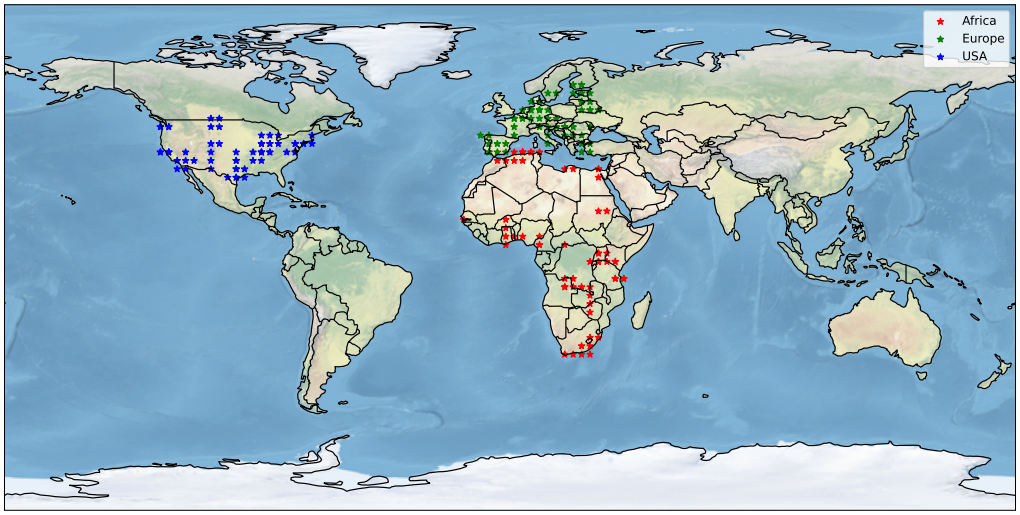

Figure 1. The location for the Sentinel-2 tiles used in this study. All locations are centered around major urban areas in order to ensure as diverse a land-cover distribution as possible.

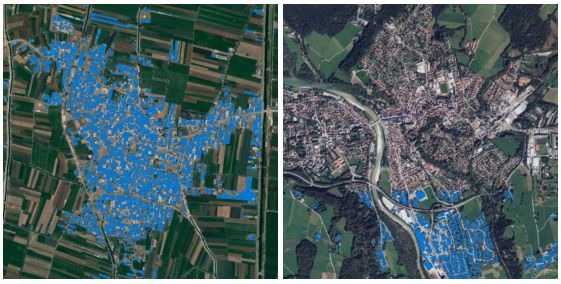

Figure 2. Examples of building footprint with missing labels and misaligned segments.

2. Related Work

Recent advancements in FMs have extended to the geospatial domain, enabling large-scale, transferable representations across various Earth observation tasks. Inspired by the success of FMs in natural language processing and computer vision, researchers have developed geospatial equivalents that pretrain on massive satellite imagery datasets such as EuroSAT, fMoW, DOTA, Five Billion Pixels, DeepGlobe (Helber et al., 2019, Christie et al., 2018, Ding et al., 2021, Tong et al., 2023, Demir et al., 2018) and fine-tune across diverse downstream applications.

There is a wide range of FMs focusing on many aspects of satellite imagery. To name a few, SatMAE (Cong et al., 2022) is trained with a focus on multi-temporal and multispectral aspects of the Sentinel-2 imagery. SatMAE++ (Noman et al., 2024) builds on this by incorporating cross-sensor pretraining (e.g., Sentinel-2 and NAIP), refined positional encoding, and improved generalization. Scale-MAE (Reed et al., 2023) is pretrained by masking the images at a known input scale; this way, the scale of the ViT positional encoding is determined by the area of the Earth covered by the image. Prithvi (Jakubik et al., 2023), developed by NASA and IBM, is a FM pretrained on Harmonized Landsat-Sentinel (HLS) (Claverie et al., 2018) data for climate and environmental monitoring. Writers in CROMA (Fuller et al., 2023) employ contrastive learning and reconstruction approaches for pretraining on unimodal and multimodal data. DOFA (Xiong et al., 2024) introduces a domain-oriented approach by integrating geospatial domain knowledge during pretraining to enhance task-specific adaptation. MMEarth (Nedungadi et al., 2024) is a multimodal model that combines satellite imagery with auxiliary data like elevation and climate to enable cross-modal reasoning. TerraMind (Jakubik et al., 2025), a ViT-based model from IBM and ESA, the first any-to-any generative, multimodal FM that is trained on a large amount of data from optical, radar, land use, elevation, and vegetation modalities for versatile downstream tasks such as building footprint detection and disaster monitoring.

Along with the rise of GFMs, different evaluation benchmarks have been proposed. Writers in PhilEO (Fibaek et al., 2024) propose an evaluation benchmark on a large amount of Sentinel-2 data for the tasks of building density estimation, road segmentation, and land cover classification. Also, GEO-Bench (Lacoste et al., 2023) introduces a benchmark comprising six classification and six segmentation tasks. FoMo (Bountos et al., 2023) presents a unified forest monitoring benchmark that includes aerial and inventory data addition to satellite data, for multiple types of downstream applications such as forest-monitoring tasks, spanning classification, segmentation, and object detection. Last but not least, PANGAEA (Marsocci et al., 2024) presents a comprehensive and standardized evaluation protocol covering a diverse set of datasets, tasks, resolutions, sensor modalities, and temporalities.

Although different benchmarks considered several aspects of RS data, the geolocation-dependent characteristics of RS data and Earth observation tasks are neglected during the pretraining and fine-tuning of FMs. On the other hand, the regional biases in the pretraining impact the performance of the downstream applications when tested on both underrepresented and overrepresented data (Marsocci et al., 2024). This inspired us to introduce geospatial splitting and evaluate the performance of FMs from a perspective of regional representation learning and fine-tuning.

3. Dataset

For this paper, we created a dataset that includes 90 *Sentinel-2* tiles (10980×10980 pixels) from 3 geolocations, namely USA, Europe, and Africa (30 tiles per region). The tiles include 12

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-849-2026 | © Author(s) 2026. CC BY 4.0 License.

Sentinel-2 bands with their native resolution. Additionally, each tile has its corresponding *Microsoft building footprint* masks (Microsoft, 2022) and *Meta canopy height* maps (Meta, 2024). The distribution of data points over the globe can be seen in Fig. 1. In line with (Prexl and Schmitt, 2023, Prexl et al., 2024), the labels for the building footprint segmentation task will be sampled on a $2.5\,\mathrm{m}$ pixel spacing, where the network output will be accordingly designed.

For the purpose of this study, we focus on two representative downstream tasks in Earth observation:

- **Canopy Height Estimation**: This task provides critical insights into forest structure, biomass distribution, and carbon storage, serving as a key component in ecosystem monitoring, natural resource management, and climate impact assessment.
- **Building Footprint Detection**: Accurate delineation of built-up areas supports applications in urban planning, disaster management, and population mapping.

These two tasks were selected to cover both regression and segmentation problem types, offering methodological diversity while remaining sufficiently tractable. This allows us to maintain a clear focus on the core objective of this work—analyzing the effect of region-specific representations—without introducing confounding complexity from highly specialized task formulations.

The selection of the three regions is guided by both task relevance and the need to capture meaningful structural diversity. Since one of the primary downstream tasks is building footprint detection, we focused on tiles centered around urban and peri-urban areas to ensure sufficient presence of built-up structures. At the same time, these regions exhibit clear differences in urban morphology, building density, and spatial organization. For instance, cities in the USA typically feature regular grid-like layouts with larger and well-separated buildings, European cities often show more compact and heterogeneous structures with historical patterns, while many regions in Africa include smaller, less regular, and more fragmented settlements. These variations provide a suitable testbed for analyzing how region-specific characteristics influence representation learning. We intentionally limited the study to these three regions to maintain a controlled experimental setup with balanced data availability and annotation quality, while still covering a broad spectrum of structural variability relevant to the selected tasks.

Quality Considerations Acquiring accurate building footprint labels is yet a challenging task. The reason is that the quality of the building segments varies based on the location. For instance, when observing Microsoft's building footprint from Africa, the building segments appear as scattered, non-square-shaped segments, whereas the building segments of the USA are sharp-edged and larger on average. Additionally, there are either missing buildings or misalignments of the labels with the satellite data. Images in Fig. 2 show an example with small, scattered, and misaligned buildings of a village in Africa (Left), and the lack of a great proportion of labels in a city in Europe (Right). Considering the inaccurate building footprint labels, we visually inspected images to select the highest quality ones for validation and fine-tuning (6 tiles per region with human-inspected high-quality labels, whereas three will be used for fine-tuning and three for validation, as will be discussed later).

4. Experimental Setup

4.1 Pretraining

The experiments were structured to assess the impact of alternative geolocation-aware pretraining and fine-tuning configurations on downstream task performance in the region of interest. We consider four sets of configurations for pretraining and fine-tuning, and two baselines as below:

- **Global pretraining, global fine-tuning**. A very commonly used setup for adapting a FM to downstream applications.
- **Global pretraining, regional fine-tuning**. A less commonly used setup for regional adaptation of the FM.
- **Regional pretraining, regional fine-tuning**. The geolocation-specific pretraining puts a strong focus on regional representation learning. The model is then fine-tuned on the data from the region of interest to achieve the most region-specific setup.
- **Global pretraining, regional adaptation in the pretraining, regional fine-tuning**. This setup is inspired by the fact that global pretraining has more diversity, but region-specific representations may be useful; we introduce this setup. In this configuration, the model is first pretrained globally, and then a few epochs of additional pretraining runs are conducted with the regional data. The intention is to steer the extraction of features towards the regional representations. The model is then fine-tuned on the regional data.
- **From scratch global training**. This setup has no pretraining and serves as a global baseline to observe the impact of pretraining on each downstream task
- **From scratch regional training**. This setup also includes no pretraining and serves as a regional baseline for the regionally pretrained models.

We employ a Vision Transformer (ViT) architecture trained using the Masked Autoencoder (MAE) framework. To ensure consistency and fairness, we use $100{,}000$ training samples of size 144×144 pixels for both global and regional pretraining. Each pretraining run is conducted for $1{,}000$ epochs with a 50% masking ratio. The model processes four Sentinel-2 spectral bands – red, green, blue, and near-infrared (NIR) – with all input values normalized to the range $[0, 1]$. The ViT encoder uses the base configuration with an embedding dimension of 768, 12 transformer layers, and 12 attention heads, and a patch size of 8. This configuration results in approximately 86 million trainable parameters during pretraining. All pretrainings are conducted using AdamW optimizer, Mean Squared Error (MSELoss), and a batch size of 512. The resulting validation loss as a function of the pretraining step can be seen in Figure 3.

4.2 Fine-tuning

For the purpose of this study, we consider two tasks with differnet learning objectives and evaluation strategies:

- **Canopy Height Estimation** (pixel-wise regression).
- **Building Footprint Detection** (pixel-wise binary segmentation).

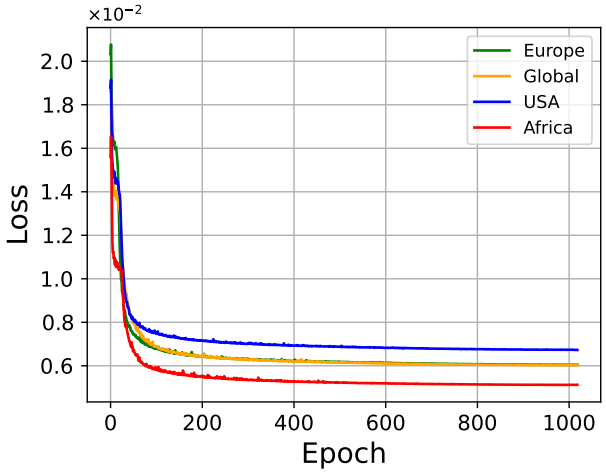

Figure 3. Reconstruction loss on the validation set for the pretraining runs.

While regression involves predicting continuous values and emphasizes numerical accuracy, segmentation requires structured spatial predictions and focuses on classification performance. This diversity allows us to evaluate the generalization ability of pretrained models across fundamentally different types of outputs and task formulations, thereby strengthening the robustness and applicability of our findings.

For the purpose of both tasks, we only use four Sentinel-2 bands to ensure high-resolution consistency (10m) and avoid potential noise from resampling coarser bands. This configuration also enables computationally efficient experiment runs while aligning with operational setups. While additional bands, such as red-edge or SWIR, may offer complementary information, our focus was to evaluate performance under minimal yet commonly available spectral inputs.

Each pretrained model is fine-tuned in both global and regional settings. For this, the pretrained backbones will be extended via the DPT (Ranftl et al., 2021) approach in order to predict pixel-wise quantities. For both tasks, we use $3{,}000$ samples of size 144×144 from the aforementioned hand-picked tiles without missing buildings. We fine-tune the pretrained models for canopy height estimation for 20 epochs and for building footprint segmentation for 40 epochs. In all fine-tuning, the encoder weights are frozen to avoid over-adaptation of the weights to the fine-tuning task, and have a more concrete assessment of the pretraining performance.

All models are trained using the AdamW optimizer with a batch size of 64. The Smoothed Mean Absolute Error (Smooth-L1) loss is used for canopy height estimation to combine the robustness and smoothness of the L1 and L2 losses. The Dice Score Loss for building footprint segmentation is used to directly optimize the overlap between the ground truth and prediction, counting for both false positives and false negatives, for better handling of small or sparse structures.

All training and evaluation are conducted on NVIDIA®A100 80GB PCIe GPUs.

Evaluation Metrics The evaluation metrics for canopy height estimation are chosen as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to ensure reliability and check for large errors. To evaluate the performance of the building footprint segmentations, Mean Intersection over Union (mIoU), as well as precision, is used to observe the overlap quality and the false positive rates.

5. Evaluation Results

The results of the experiments for the two downstream tasks are illustrated in Table 1 and 2. Here, the obtained metrics for the four earlier discussed pretraining approaches, as well as the two training runs without any pretraining, are presented (referred to as *from scratch*). Each table has six rows presenting the results evaluated on data from each of the three regions, with two different metrics respectively. It is to be mentioned that the higher values in the metrics used for building footprint segmentation indicate better performance, whereas in the canopy heights, lower values refer to lower errors, therefore, better results.

Several trends can be observed independently of the downstream task. For both downstream applications, the from-scratch training has the lowest performance, and pretraining improves the predictions significantly. While a globally pretrained and fine-tuned model is the most common practice among scientists, the results from the table indicate that tailoring the pretraining to the specific region of interest improves the performance of the corresponding models. For the task of building footprint segmentation, over all three regions, the best-performing models are either pretrained on the specific region or by adapting a globally pretrained checkpoint to the specific region. The same holds true for the results of Table 2, with one exception for the RMSE in Africa.

To provide the reader with a visual intuition of the performance of the models, we illustrate the best and the worst models in each region for each of the two tasks in Fig. 5. The columns from left to right show the RGB Sentinel-2 image, ground truth label for the corresponding task, the prediction of the lowest performing and the best-performing model, as well as the differences between the corresponding predictions to the ground truth. Here, it can be observed that the error rate – false negative in the case of building footprint segmentation and underestimation in the case of canopy height prediction – reduces significantly when conducting pretraining according to the results in Table 1 and 2.

5.1 Computational Cost of Regional Adapation

When studying the regional adaptation of global models, the question arises of how much adaptation is needed for a performance improvement. This, of course, represents an important question for end-users of remote sensing foundation models (RSFMs), since regional adaptation goes beyond fine-tuning additional layers on a frozen backbone, which serves as a feature extractor and hence is more computationally expensive. To address this question, we implement the adaptation with different numbers of additional pretraining epochs on a specific target region – starting from a given globally pretrained checkpoint – and observe the performance differences for the task of building footprint segmentation in Africa. Figure 4 shows the corresponding results on how the mIoU values for the fine-tuning on Africa change when additional regional adaptation (pretraining) is conducted as a function of the adaptation epochs. Here, a linear trend can be observed, which indicates an improvement of the fine-tuning results with respect to a larger number of adaptation epochs. As a baseline, the globally pretrained locally fine-tuned performance is given. It can be observed that (according

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-849-2026 | © Author(s) 2026. CC BY 4.0 License.

		From Scratch		Pretrained				
		-	-	Global	Global	Local	Global Adapt	← Pretrained on
Region	**Metric**	Global	Local	Global	Local	Local	Local	← Fine-Tuned on
Africa		0.53	0.59	0.66	0.67	0.65	**0.68**	
Europe	IoU	0.55	0.55	**0.63**	0.62	**0.63**	**0.63**	
USA		0.56	0.57	0.66	0.66	**0.68**	0.67	
Africa		0.65	0.70	**0.77**	0.76	**0.77**	**0.77**	
Europe	Precision	0.65	0.63	0.71	0.71	**0.72**	0.71	
USA		0.65	0.68	0.75	0.76	**0.77**	**0.77**	

Table 1. The evaluation results on the task of building footprint segmentation. Two metrics, IoU and Precision, for the evaluation of the model performance over three regions are presented for two models trained from scratch and four different pretraining and fine-tuning strategies. The higher the value, the better the model performance.

		From Scratch		Pretrained				
		-	-	Global	Global	Local	Global Adapt	← Pretrained on
Region	**Metric**	Global	Local	Global	Local	Local	Local	← Fine-Tuned on
Africa		0.165	0.136	0.102	0.102	**0.101**	**0.101**	
Europe	MAE	0.158	0.146	0.100	0.097	**0.095**	0.096	
USA		0.146	0.144	0.098	0.096	**0.093**	0.095	
Africa		0.178	0.188	**0.138**	0.143	0.142	0.144	
Europe	RMSE	0.193	0.190	0.146	0.147	0.146	**0.144**	
USA		0.197	0.198	0.143	0.140	**0.137**	0.138	

Table 2. The evaluation results on canopy height estimation. Two metrics MAE and RMSE, for the evaluation of the model performance over three regions are presented for two models trained from scratch and four different pretraining and fine-tuning strategies. As the values shown in this table indicate the error values, the lower the values, the better the perfromance of the model.

to this experimental setup) the local adaptation increases the performance when the adaptation epochs exceed around ≈ 30 epochs.

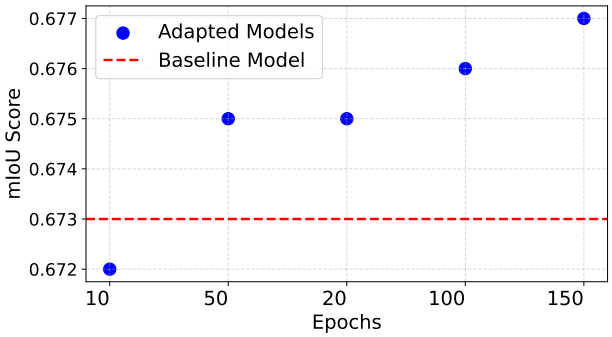

Figure 4. The evaluation results of building footprint segmentation on Africa. The baseline represents a globally pretrained and locally fine-tuned model. The Adapted models have additional pretraining on images from Africa.

5.2 Masking Ratio of Tokens in the Pretraining

In contrast to natural images with uniform content, RS images are inherently heterogeneous, including multiple land cover types such as cropland, buildings, rivers, lakes, and forests, all within the same scene. The semantic variability in RS data poses an additional challenge to inferring missing information from limited context. Hence, the effectiveness of masked autoencoder (MAE) pretraining is even more sensitive to the choice of masking ratio, where higher masking ratios may lead to greater reconstruction difficulty. Nevertheless, we also experimented the performance of the pretrained models with 75% masking ratio on the downstream applications. In Table 3, the experiment results of canopy height estimations are shown, where 75% of the tokens were masked during pretraining.

	PT:	Global	Global	Local	Global Adapt
Region	FT:	Global	Local	Local	Local
Africa		0.110	**0.106**	0.124	**0.106**
Europe		0.116	0.113	0.123	**0.112**
USA		0.114	0.113	**0.084**	0.110

Table 3. The MAE values for canopy height estimation. The pretrained model has 75% masked tokens. PT: pretrained on, FT: fine-tuned on.

5.3 Fine-tuning with the Frozen or Non-Frozen Weights?

We conducted fine-tuning experiments both with frozen and unfrozen ViT encoder weights. When the encoder weights are not frozen, the model has more flexibility to adapt; hence, the interpretation of the global and local pretrainings is more difficult. Therefore, in our main experiments, we chose to illustrate the results using frozen encoder weights, as this setting better highlights the contribution of the pretrained representations.

	PT:	Global	Global	Local	Global Adapt
Region	FT:	Global	Local	Local	Local
Africa		**0.63**	**0.63**	0.58	**0.63**
Europe		**0.58**	**0.58**	0.57	**0.58**
USA		0.60	0.61	**0.68**	0.62

Table 4. The mIoU values for Building footprint segmentation. The pretrained model has 75% masked tokens. PT: pretrained on, FT: fine-tuned on.

6. Discussion and Conclusion

Scientific Contribution and Relevance This study investigates how different pretraining and fine-tuning approaches affect the

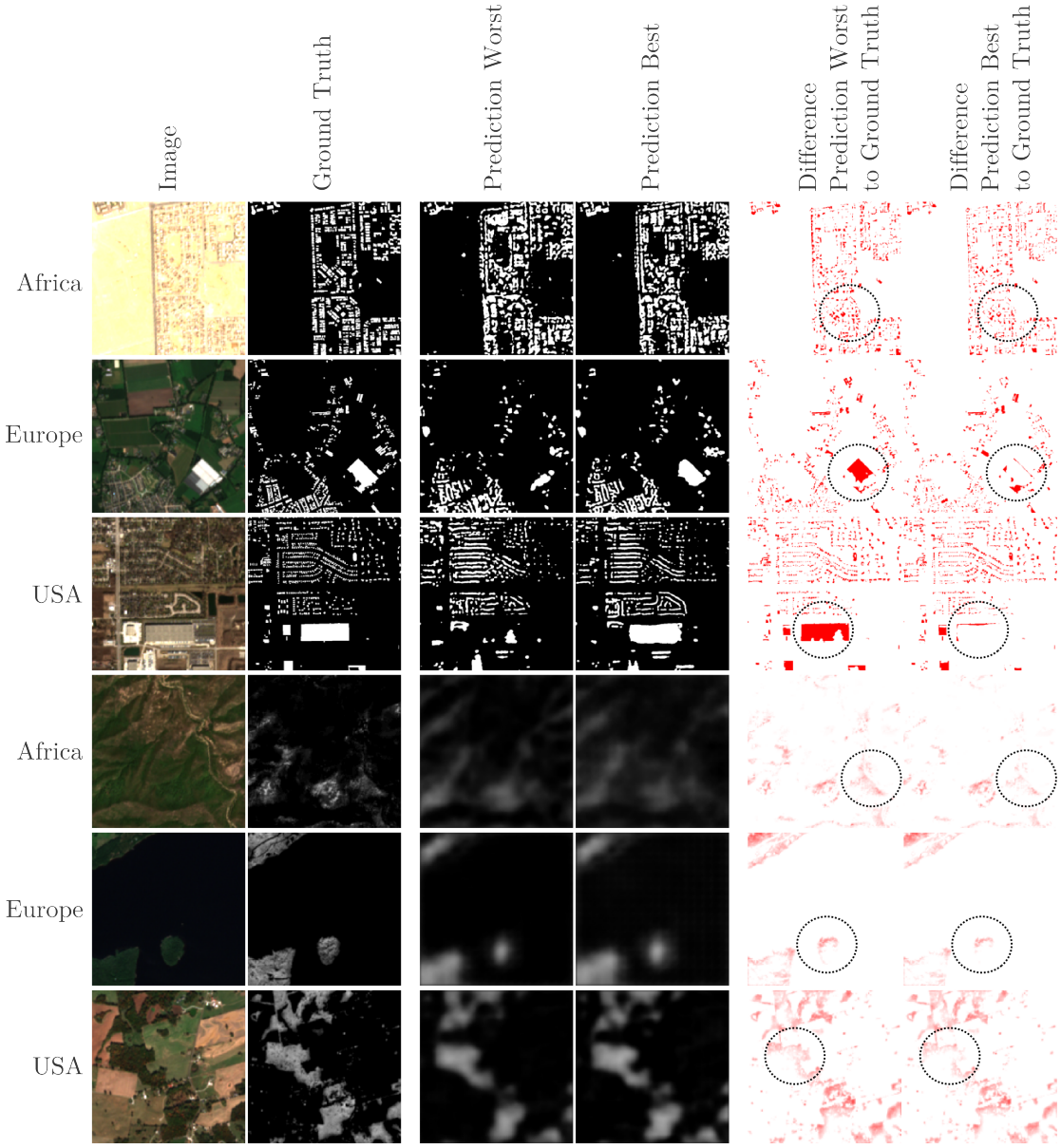

Figure 5. The sample validation images from each region for both downstream tasks. Columns from left to right are the Sentinel-2 RGB image, the ground truth mask, the lowest quality pretrained model, the best performing pretrained model predictions, the error corresponding to the predictions of the worst and the best performing models, respectively. The rows from top to bottom are selected for the building footprint detection in Africa, Europe, and the USA, and the canopy height estimation in Africa, Europe, and the USA.

model performance on the downstream applications when observed and evaluated separately across different regions. Specifically, we explore whether incorporating regional information during pretraining or tailoring fine-tuning to specific geographies leads to improved building footprint segmentation and canopy height estimation results on the regions of interest. This question is particularly relevant for applications targeting specific geographic areas, where end-users rely on and adapt pretrained models.

Region-Specific Representation Learning Table 1 and Table 2 provide insights into how region-specific representation learning can improve performance on tasks such as building footprint segmentation and canopy height regression. We observe that most regionally focused pretraining and fine-tuning approaches outperform globally pretrained models. This is likely because features learned during reconstruction-based pretraining (e.g., using MAEs) are more useful for downstream applications when images are self-similar. For example, features useful for reconstructing urban areas in the U.S. differ significantly from those required for urban regions in Africa due to differences in urban layout, spectral signatures, and geometric scales. Regional pretraining exposes the model to relevant data distributions, which supports better fine-tuning. Nearly all globally pretrained models see improvements when adapted to a specific region. As shown in Figure 4, these gains can be further increased by extending the number of adaptation epochs.

Limitations and Methodological Considerations Although performance differences are consistent across all three regions, the overall improvement margin remains limited. On average, improvements amount to 0.013 points in mIoU for building seg-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-849-2026 | © Author(s) 2026. CC BY 4.0 License.

mentation and 0.003 MAE for canopy height estimation. Nevertheless, the consistency of the improvements suggests they may be even more impactful for other downstream tasks or in low-label settings.

The Comparability of Regional Pretrainings Several methodological factors are crucial for fair evaluation and could further enhance the performance of regionally adapted pretraining strategies. For example, we pretrained on 30 tiles per region centered on urban areas. However, the tiles from the U.S. contained a higher diversity of urban structures, whereas many African tiles included large homogeneous (bare land) areas next to urban centers. Such uniform scenes are intuitively suboptimal for reconstruction-based MAE pretraining, as the task becomes easier and less informative. This is supported by the better reconstruction results observed in Fig. 3. To obtain more representative results, future datasets should include more geographically diverse tiles and exclude overly homogeneous areas during pretraining.

Reconsidering the Definition of *Region* In this study, regions were defined by continent. However, alternative region definitions may be more appropriate depending on the application. For instance, like other works (Madadikhaljan and Schmitt, 2025), in tasks where climate is a critical factor, regions could be defined using Köppen-Geiger climate zones (Cui et al., 2021). Granularity should also be task-dependent: defining regions by shared task-relevant geolocation characteristics (e.g., similar land use or environmental conditions) is a reasonable assumption. Additionally, similarity-based clustering could be performed using land-cover classes (e.g., urban, forest) or image similarity metrics to define fine-tuned models.

Pretraining Setup In addition to the experiments presented in Section 5, we also investigate alternative pretraining configurations, including a higher masking ratio of 75% and a smaller number of pretraining samples, namely, $10{,}000$. Due to space constraints, we only report a representative subset of these additional results. As illustrated in Table 4, the canopy height regression task further confirms that incorporating regional representations during pretraining yields improved performance. Models pretrained with a 50% masking ratio and $100{,}000$ samples show improved overall performance on the downstream applications, and therefore, are selected as the main results of this paper.

Fine-tuning Setup We conducted experiments using both frozen and unfrozen encoder weights during fine-tuning. Table 4 presents the exemplary building footprint segmentation results for the configuration in which the encoder weights were allowed to update. In most cases, models with unfrozen encoders converged to similar performance levels, likely due to their increased capacity to adapt when supervised labels are provided. To better isolate the contribution of the pretrained representations and to reflect more practical deployment scenarios—where large-scale fine-tuning may not always be feasible—we report our main results using frozen encoder weights.

Why Existing GFMs Are Not Used in This Study? While recent geospatial foundation models have demonstrated strong performance on various downstream Earth observation tasks, they are predominantly trained on large-scale, globally distributed datasets. To date, no publicly available region-specific foundation model exists that would be suitable for the type of regional analysis presented in this work. Consequently, we opted to pretrain our own model in order to retain full control over the data used for both pretraining and fine-tuning. This approach allowed us to ensure fairness in terms of data quality and quantity across regions and to conduct a controlled and interpretable comparison of region-specific representations. Importantly, our architecture and training configurations were designed to remain closely aligned with those used in state-of-the-art foundation models.

Why Location Encoders Are Not Used in This Study? Recent work has also explored incorporating explicit geolocation information through learned embeddings, such as SatCLIP and related approaches. These methods typically leverage large-scale, globally distributed datasets and introduce auxiliary location-based features to guide representation learning. In contrast, our study is designed as a controlled comparison in which both global and regional models are trained on the same underlying data pool, differing only in their geographic composition. This allows us to isolate the effect of geolocation-aware data selection without introducing additional sources of information or variability.

Implications for RSFMs The results of this study suggest that focusing on specific regions – through either pretraining or fine-tuning – can enhance downstream task performance. This partially challenges the premise of RSFMs, which posits that large-scale global pretraining will universally improve performance. Further research should investigate how regionally adapted models compare to globally trained ones in terms of performance and generalization. While regional models introduce bias and limit general applicability, they offer performance gains that may justify their use in localized applications. Managing multiple regionally adapted models poses a tradeoff: increased complexity for improved local accuracy. The regional adaptation strategy proposed in this manuscript aims to balance both global diversity and improved regional performance.

Future work should further strengthen and generalize the findings presented in this study. First, expanding the experimental setup to include locally validated benchmark datasets, such as *PASTIS* (Sainte Fare Garnot and Landrieu, 2021) or *AI4SmallFarms* (Persello et al., 2023), would provide a more direct evaluation of region-specific performance and better align with real-world deployment scenarios. Second, incorporating statistical analysis such as multiple training runs with different random seeds and significance testing would enable a more rigorous assessment of the observed performance differences, particularly given their relatively small margins. Finally, extending the experiments to additional downstream tasks, including multi-class classification and more complex segmentation settings, would help assess the generality of the proposed strategies across diverse application domains.

References

Bountos, N. I., Ouaknine, A., Rolnick, D., 2023. FoMo-Bench: a multi-modal, multi-scale and multi-task Forest Monitoring Benchmark for remote sensing foundation models. *arXiv preprint arXiv:2312.10114*.

Christie, G., Fendley, N., Wilson, J., Mukherjee, R., 2018. Functional map of the world. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6172–6180.

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C., Skakun, S. V., Justice, C., 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote sensing of environment*, 219, 145–161.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S., 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35, 197–211.

Cui, D., Liang, S., Wang, D., 2021. Observed and projected changes in global climate zones based on Köppen climate classification. *Wiley Interdisciplinary Reviews: Climate Change*, 12(3), e701.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 172–181.

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M. Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M. et al., 2021. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11), 7778–7796.

Ekim, B., Tadesse, G. A., Robinson, C., Hacheme, G., Schmitt, M., Dodhia, R., Ferres, J. M. L., 2025. Distribution shifts at scale: Out-of-distribution detection in earth observation. *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2265–2274.

Fibaek, C., Camilleri, L., Luyts, A., Dionelis, N., Le Saux, B., 2024. Phileo bench: Evaluating geo-spatial foundation models. *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2739–2744.

Fuller, A., Millard, K., Green, J., 2023. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 5506–5538.

Helber, P., Bischke, B., Dengel, A., Borth, D., 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2217–2226.

Jakubik, J., Roy, S., Phillips, C., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B. et al., 2023. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*.

Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N. et al., 2025. Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*.

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E., Kerner, H., Lütjens, B., Irvin, J., Dao, D., Alemohammad, H., Drouin, A. et al., 2023. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36, 51080–51093.

Madadikhaljan, M., Schmitt, M., 2025. Geolocation-Aware Deep Coding. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 1–16.

Marsocci, V., Jia, Y., Bellier, G. L., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A. et al., 2024. PANGAEA: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*.

Meta, 2024. High resolution canopy height maps. Accessed: 2025-06-13.

Microsoft, 2022. Microsoft building fooptprints. Accessed: 2022-11-01.

Nedungadi, V., Kariryaa, A., Oehmcke, S., Belongie, S., Igel, C., Lang, N., 2024. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. *European Conference on Computer Vision*, Springer, 164–182.

Noman, M., Naseer, M., Cholakkal, H., Anwer, R. M., Khan, S., Khan, F. S., 2024. Rethinking transformers pre-training for multi-spectral satellite imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27811–27819.

Persello, C., Grift, J., Fan, X., Paris, C., 2023. AI4SmallFarms: A Data Set for Crop Field Delineation in Southeast Asian Smallholder Farms.

Prexl, J., Baumann, A., Schmitt, M., 2024. A Comparison of Uncertainty Estimation Methods for Building Footprint Change Detection from Sentinel-2 Imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 339–346.

Prexl, J., Schmitt, M., 2023. The potential of sentinel-2 data for global building footprint mapping with high temporal resolution. *2023 Joint Urban Remote Sensing Event (JURSE)*, IEEE, 1–4.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. *Proceedings of CVPR*, 12179–12188.

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T., 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4088–4099.

Roscher, R., Russwurm, M., Gevaert, C., Kampffmeyer, M., Dos Santos, J. A., Vakalopoulou, M., Hänsch, R., Hansen, S., Nogueira, K., Prexl, J. et al., 2024. Better, not just more: Data-centric machine learning for Earth observation. *IEEE Geoscience and Remote Sensing Magazine*.

Sainte Fare Garnot, V., Landrieu, L., 2021. Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. *ICCV*.

Tong, X.-Y., Xia, G.-S., Zhu, X. X., 2023. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196, 178–196.

Xiong, Z., Wang, Y., Zhang, F., Stewart, A. J., Hanna, J., Borth, D., Papoutsis, I., Le Saux, B., Camps-Valls, G., Zhu, X. X., 2024. Neural plasticity-inspired foundation model for observing the Earth crossing modalities. *arXiv e-prints*, arXiv–2403.