PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science (2025) 93:3–18
https://doi.org/10.1007/s41064-024-00328-5

ORIGINAL ARTICLE

Geolocation-Aware Deep Coding

Infusing Geolocation Information into Deep Neural Networks for Remote Sensing

Mojgan Madadikhaljan[1] · **Michael Schmitt**[1]

Received: 9 September 2024 / Accepted: 19 November 2024 / Published online: 7 January 2025

Abstract
With more and more remote sensing data available on a global scale, the Earth observation community strives to harness the power of modern deep learning techniques by developing globally applicable models. However, remote sensing images exhibit strongly heterogeneous, geolocation-dependent characteristics, making this a challenging endeavor. In this paper, we introduce the geolocation-aware deep coding strategy to incorporate geolocation information of remote sensing data into the training of the deep learning models. The proposed method consists of defining regional subnetworks dedicated to each subset of the dataset with similar geolocational characteristics. Using two application examples, namely the mapping of building footprints from multi-spectral Sentinel-2 imagery, and the task of forest detection from single-channel thermal infrared Landsat imagery, we show that the proposed deep coding strategy stabilizes the training performance and can also improve the predictive power of deep neural networks designed for remote sensing data analysis.

Keywords Geolocation-aware Deep Coding · Global Training · Remote Sensing

1 Introduction

During the last decade, deep learning has become an established paradigm in the world of remote sensing (RS) wherein classically, the focus is on a specific scene or region of interest. Recently, the growing number of available datasets and the power of deep neural networks have initiated an interest in training more generic and global models (Schmitt et al. 2021; Hooker et al. 2018; Chen et al. 2021; Bastani et al. 2023; Vega et al. 2017; Fritz et al. 2017; Roßberg and Schmitt 2023). The global models are trained on datasets containing data from different regions all over the world and are expected to perform reasonably well in all locations. However, due to the geolocation-dependent characteristics of RS images, it is non-trivial to train a well-performing global model that consistently delivers optimal results across all regions (Beery et al. 2022).

Being a physical measurement of the Earth's surface, the appearance of an RS image is highly influenced by the geographical location of the scene. For instance, optical images of circular farmlands in Kansas, USA seem relatively different from flower farmlands in Hillegom, Netherlands in terms of shapes and colors (see Fig. 1 first row). Urban areas, and specifically the roofs of the buildings, are dissimilar in cities such as New York, USA and Munich, Germany (see Fig. 1 second row). Concurrent thermal images from a tropical forest in South India and a boreal forest in Alaska have varying gray value distributions (see Fig. 1 last row). Therefore, it is essential to take the geolocation-specific characteristics of RS data into account when training a global Earth observation (EO) model.

In addition to the appearance of the imagery, the distribution of the labels in the EO tasks can be greatly geolocation-dependent. For example, the distribution of crop types in different regions of the globe is very inconsistent as can be seen from Fig. 2. This complicates the task of classifying crop types from RS data on a global scale without consideration of regional characteristics. Except for the close-to-perfect sampling of the whole globe, adding more data to the

✉ Mojgan Madadikhaljan
mojgan.madadikhaljan@unibw.de

Michael Schmitt
michael.schmitt@unibw.de

[1] Department of Aerospace Engineering, University of the Bundeswehr Munich, Neubiberg, Germany

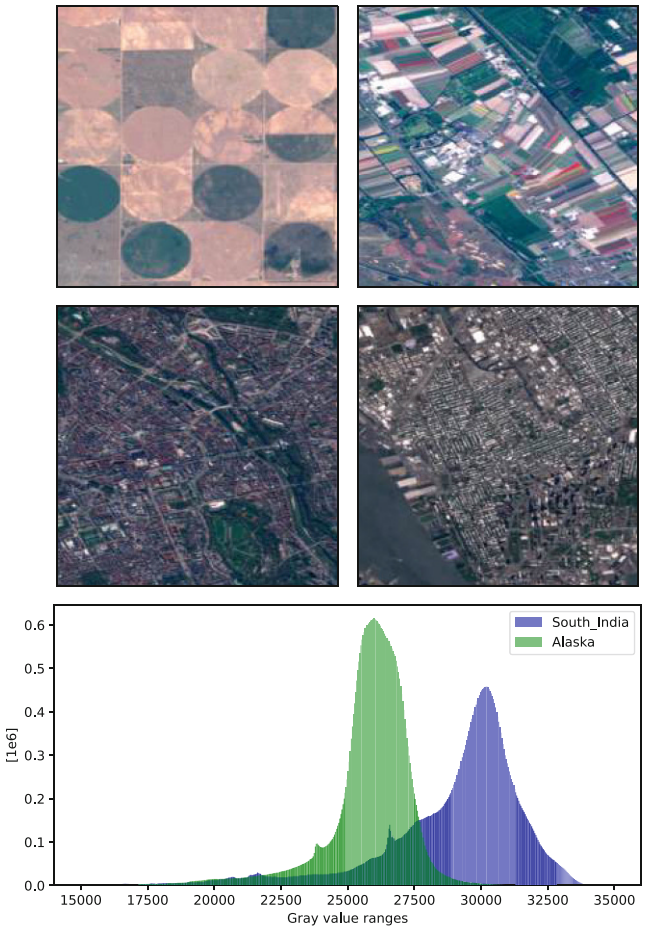

Fig. 1 Remote sensing images with geolocation-dependent appearance differences. The first row: Sentinel-2 image from farmlands in Kansas, USA (left) and in Hillegom, Netherlands (right) in the spring (2022). The second row: Sentinel-2 image from Munich City Center, Germany (left) and New York City Center, USA (right) in the spring (2022). The last row: The histograms of Landsat band 10 thermal images from forests in South India and Alaska in January (2021)

training of a classifier with location-dependent distribution will not boost the model predictions and purely data-driven approaches may result in unsatisfactory results (Von Rueden et al. 2019).

Typically, a machine learning model achieves better results when trained on a specific region—resulting in a so-called regional or region-specific model—and tested on the same region because of a certain amount of geospatial overfitting. On the other hand, for global implementation of EO tasks, it is usually not favored to train numerous region-specific models. Thus, to achieve desirable inference performance on a global scale, geolocational information should be embedded into the training of global models.

Due to the georeferenced nature of RS data, the geolocation of the image is typically known and provides valuable information about the image content e.g., the possibility of the presence of land cover classes, distribution of crop types, probability of flood or volcanic explosions, temperature range, etc. Therefore, instead of multiple regional models, it is preferred to train one geolocation-aware global model with an improved understanding of geolocation-related appearance and distribution differences.

The insertion of geolocation information can be realized in different levels of processing (Von Rueden et al. 2019; Camps-Valls et al. 2021). The infusion can be implemented at the data input level, where one or several layers containing geolocation information are concatenated to the input image channels. For instance, Mahara and Rishe (2023) convert location information, namely latitude and longitude values to geohash codes and add them as an extra raster layer concatenated to the extracted feature map. Similarly, Liu et al. (2018) add two rasters including scaled cartesian coordinates to the input image and feed all together to the rest of the network. Alternatively, Zhang et al. (2021) create coordinate vectors from location information and then convert them to coordinate features by passing them through fully connected neural nodes. The location feature vector is then stacked with the spectral feature vector and passed through the rest of the network. While the methods above are relatively simple to implement, there is a neglected problem with directly inserting coordinate values into a network. The direct coordinate values are not the optimal representors of the geolocational properties of RS data. The reason is that both cartesian and latitude/longitude coordinates are designed to enable locating in a three- or two-dimensional grided space, however, they do not necessarily reflect the geolocational similarities and differences of data with diverse microclimates caused by geographical features e.g., mountains, valleys.

The geolocation information can also be incorporated into the learning process wherein a mathematically designed module is infused into the model architecture. The mathematical formula may be derived from the formation of the natural phenomena or statistics and added to a specific depth in the model. Liu et al. (2023) design a multimodal fine-grained dual network (Dual-Net), which takes dual-date images and predicts land cover maps. They embed modal information (dates and geo-locations) into the model using a position-aware adaptive block. To dehaze optical RS images, Wen et al. (2023) define a physics-aware intralevel fusion module that compensates for insufficient elimination of haze-degraded information. The module allows the dehazing process to also consider the location-dependent parameters such as the transmission map of each pixel into account. To estimate the urban surface temperature in a high resolution, Chen et al. (2022) insert a so-called global physics feature perception branch into their network. The branch inputs atmospheric variables as well as location information—high-precision 3D point clouds representing the land surface geometry—and the authors believe that this adds a global physics guide to their model.

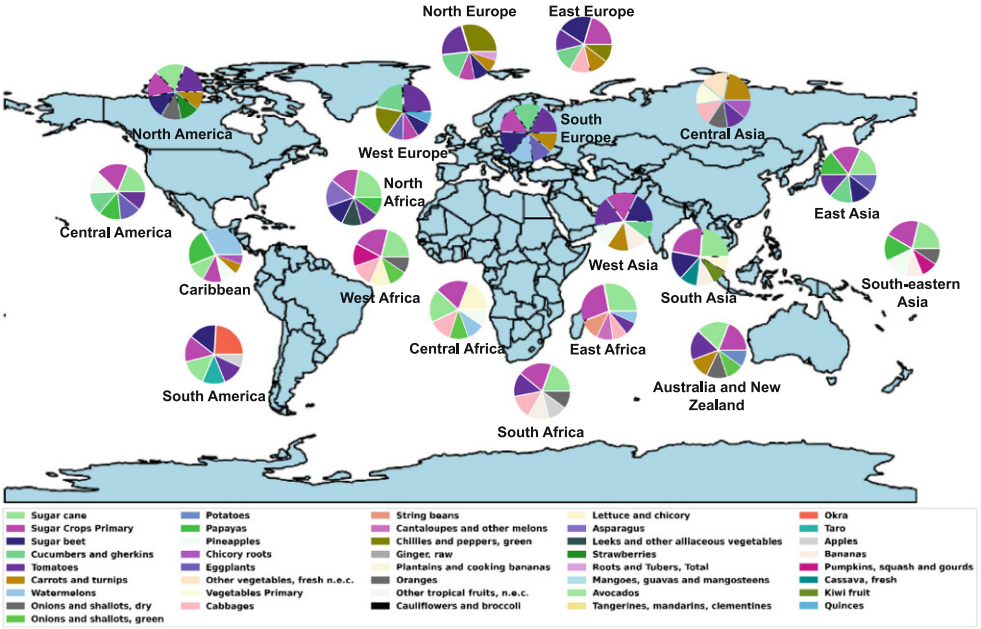

Fig. 2 Regional distribution of major crops of the world from Food and Agriculture Organization (FAO) (Food Agriculture Organization of the United Nations 2022)

Authors in Ekim and Schmitt (2024) use the cyclic coordinate encoding to infuse two additional latent feature maps derived from coordinates and extended scene content into their classification model to improve the model's predictive performance. Such models often need extra information (e.g., time, multi-temporal images, atmospheric variables, extended scene content) in addition to the geolocation properties to provide promising results. Also, the design of such modules is often very complex and sometimes not feasible. For instance, framing the EO tasks such as the classification of land covers into a mathematical formulation is not trivial. Additionally, the insertion of topographical information into training does not always improve the understanding of the model concerning the geolocational characteristics of data.

Another way of adding awareness to the model is defining a physics-related penalty to the target optimization function (Takeishi and Kalousis 2021; Diligenti et al. 2017). The effort is spent interpreting physical criteria as an optimization term to add to the optimization process. Wei et al. (2022) integrate an aerosol optical depth propagation equation to optimization function to compensate for the heterogeneous characteristics of multi-source RS data such as different data missing rates and measurement errors. Inspired by the fact that neighboring pixels tend to have similar land cover class labels, Cao et al. (2018) add a label smoothness prior to the optimization penalty. To better reconstruct the could mask, Wang et al. (2023) design a CloudMask loss consisting of two knowledge-informed loss terms to estimate cloud thickness and the number of layers. Design and optimization of a target penalty is highly task-specific and demanding in terms of fine-tuning and balancing between the labels-related penalties and the physics-related penalties.

Inspired by the research demands mentioned above, the main contributions of this work can be summarized as follows:

- We introduce a novel approach to infuse geolocation information into a deep learning architecture. The proposed approach requires the clustering of datasets into subsets with similar characteristics. Using the *deep coding* strategy, the model will have dedicated branches for each subset. The proposed model architecture in Sect. 2 enables the training of one global model that benefits from exposure to all global samples while also concentrating on geolocation-specific details.
- In Sect. 2.1, we explain our approach regarding the subdivision of a dataset and provide insights on how the division can be realized for different types of datasets. We also demonstrate two model architectures in Sect. 2.2 where *deep coding* modification is used to incorporate geolocation information.
- In Sect. 3.1, we utilize the proposed method to detect building footprints from Sentinel-2 bands for two cities. We illustrate that training a geolocation-aware model achieves more desirable qualitative and quantitative performances compared to a model with no geolocation-awareness, and a model with cyclic coordinate encoding (see Sect. 3.1.3).
- In addition, we also investigate the impact of geolocation awareness in the difficult problem of land cover classification from a single-channel image. More specifically, we intend to improve the mapping of the "tree cover" class from a single thermal infrared (TIR) satellite im-

age by adding geolocation awareness to the model (see Sect. 3.2). The proposed model is compared to globally trained models that are not geolocation-aware, and to geolocation-specific models. Inspired by Ekim and Schmitt (2024); Mai et al. (2023) and Liu et al. (2023), we also compare our results to the cyclic coordinate encoding strategy. The experiments are evaluated qualitatively and quantitatively (see Sect. 3.2.3).

2 Geolocation-Aware Deep Coding

In this section, we explain our approach to inputting the geolocation information into the model with the *deep coding* strategy. The proposed method is built upon the well-known UNet architecture (Ronneberger et al. 2015) as the backbone since it is one of the best-established deep learning models and is frequently used in RS. Nevertheless, the *deep coding* strategy is model agnostic and can be implemented on other architectures such as ResNet (He et al. 2016), AlexNet (Krizhevsky et al. 2012), and VGG (Simonyan and Zisserman 2014) as well.

Inspired by the term "hard coding" from software development practices, we introduce the *deep coding* modification to a deep convolutional neural network. The *deep coding* modification is essentially creating separate subbranches within the deep blocks of the model to demand extra focus on geolocation-specific details. The proposed methodology consists of two main steps. The first step is to conduct an initial analysis of the dataset to provide information concerning the data characteristics, biases, and distributions. Based on the outcome of the initial analysis, the dataset will be divided into subsets (i.e. regions) with similar characteristics. The term "similarity" stands for a location-dependent property of data that are not the target ground truth labels, but may be highly correlated to them. The next step is the infusion of the geolocation information e.g., regional clustering into the deep learning architecture. Firstly, in Sect. 2.1, we explain how to subdivide the dataset into regions, and next, in Sect. 2.2 we describe how the model's geolocation awareness is implemented.

2.1 Subdivision of the Data

A major step in the proposed methodology is to define the so-called "regions". The term region in the context of this paper is referred to as a spatial subdivision of a global dataset. This step requires domain knowledge about the task and a deep understanding of the geolocational characteristics of the input images and the target ground truth labels. The dataset should be studied for its geolocational similarities which are correlated to the target labels and can be used for dataset clustering and performance improvement of the model. The geolocational similarities, for instance, can be considered as the target object's shape, color, temperature, material, height, appearance in the image, etc.

The geolocation of an image can provide a wide range of information. It defines the climate properties of location and expected temperature range. It gives hints on the probability of existing land cover classes, farmed crop types, the used material in the construction, possible types of natural hazards, etc. However, where to put the boundaries and separate one region from the other is highly dependent on the tasks and the target labels. For instance, a region can be defined as a continent, country, city, climate region, geohash, etc. depending on the target labels' granularity, similarity, and distinction. For instance, if the task is to classify the crop types, climate regions will provide beneficial information on the distribution of crops and is a reasonable choice to cluster the global dataset. In the case of building footprint detection, city borders can be a wiser choice as they clarify the available construction material, culture, and policies.

A very important aspect of clustering the dataset into regions is that the characteristics of the data—from which the subsets are defined—should be known in the inference mode. The reason is the region-specific branches of the model to which the input image is forwarded depending on its georeferencing information.

When deciding on defining the region boundaries, the tradeoff between the number of regions (i.e. the granularity of clustering) as well as the density of data in each region should be well observed. The reason is that in the proposed architecture, each of the regional branches is only trained with the data from the corresponding region. In contrast to the global models where the loss from forwarding every single image is backpropagated through the whole model, in this scenario, not every single image contributes to improving the weights of the whole network. Suppose we have a dataset with m subdivisions, each of them having n_m images. In this case, every part of a global model is updated $\sum_{i=1}^{m} n_m$ times after each epoch. In contrast, the regional branches in the geolocation-aware model are updated n_m times after each epoch. Therefore, low-density regions may be combined with the most similar data to have a reasonable performance in all regions. For example, having several regions with thousands of data samples and one region with only five samples is not recommended. Drastic imbalances in the input size of each region will be further discussed in the experiment of Sect. 3.2.2.

After deciding on the number of regions, images in the dataset will be clustered to each region based on their geolocation and assigned a region index (e.g. $0, 1, 2, \ldots, m$).

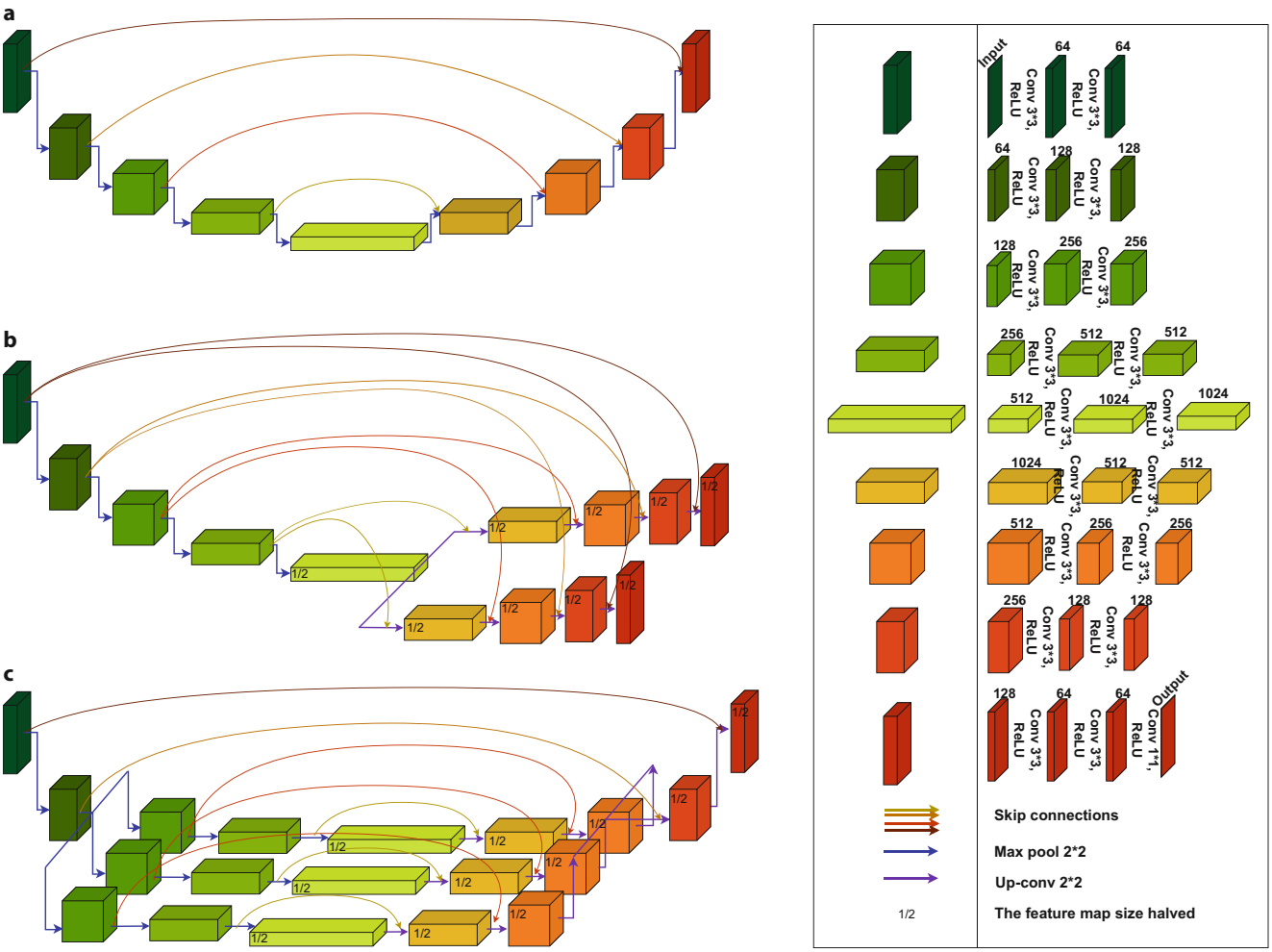

Fig. 3 Geolocation-aware deep coding modifications to the UNet backbone. **a** Original UNet architecture used as the backbone, **b** Geolocation-aware deep coding modifications to UNet with halved feature sizes in the *latter* layers of the network for $m = 2$, **c** Geolocation-aware deep coding modifications to UNet with halved feature sizes in the *intermediate* layers of the network for $m = 3$

2.2 Insertion of Geolocation Information into the Network

The *deep coding* strategy is used to have a globally trained geolocation-aware model. The *deep coding* of geolocation-dependent data clustering is realized by sub-branching deep blocks of the network. As mentioned above, we use the U-Net architecture as the backbone for our proposed *deep coding* strategy. UNet architectures use multiple convolutional neural network (CNN) blocks to extract features from input images and then apply additional CNN blocks together with skip connections to provide predictions. Fig. 3a shows the original UNet architecture. The subbranching of the model can be implemented at different levels of depth and length. It should be ensured that the branching starts where the model already has enough time to see and extract mutual features, and the branched part is also long enough to process region-specific features. The feature map size of the branches is halved to avoid the space complexity of the model.

The first implementation of *deep coding* modification is shown in Fig. 3b with $m = 2$. The architecture shares its first five CNN blocks with UNet, and it is divided into two different subnetworks (starting from the 6th CNN block) to predict the output. Each subnetwork is designated to the data from each clustered region and is only activated—backpropagated in the training phase—when images from the associated region index are fed into the model. The subnetworks therefore are exposed to less data compared to the first and last CNN blocks.

Another example of *deep coding* modification with $m = 3$ regions is demonstrated in Fig. 3c where the subdivision occurs by the end of feature extraction blocks (starting from the 3rd CNN block) and the first blocks of outcome prediction (ending in the 7th CNN block). The prior blocks of the network see images from all locations and therefore can extract the common features. In the later part of the

network, the blocks merge to output the final prediction via the two last blocks. The suggested depth of subdivision in both designs is derived empirically.

As exposure of the model to many samples robustifies the feature extraction model, we avoid placing the branching very close to the model input. The skip connections of the architecture are carried on in the *deep coding* modification to ensure the information flow through the whole network as well as each regional branch. The dimensions of CNN blocks stay unchanged as of the original UNet architecture for both feature extraction and reconstruction levels. Each regional subbranch will have different weights by the end of training.

3 Example Applications

We verify the validity of our method in two use cases. First, we incorporate geolocation awareness in detecting building footprints from Sentinel-2 images in the cities of Stuttgart, Germany, and San Francisco, USA. Next, we illustrate the effect of infusing geolocation awareness in the global classification of forests from single thermal bands of Landsat 8. For both applications,

- We compare the proposed geolocation-aware model to regional models, as well as a global model—referred to as the model which is fed with images from all available regions of interest—and another state of the geolocation-aware model containing cyclic coordinate encoding (Ekim and Schmitt 2024; Mai et al. 2023) infused to UNet. In the cyclic coordinate encoding strategy, three extra rasters containing the values of latitude, the $\sin(2\pi \times \text{longitude}/180)$, as well as the $\cos(2\pi \times \text{longitude}/180)$ are fed into an encoder block (Geo_{enc}). The extracted features from Geo_{enc} are then concatenated to the extracted image features in the pseudo bottleneck stage of UNet and forwarded to the prediction CNN blocks.
- The average Intersection over Union (IoU) is used as an established measure for the quantitative evaluation of binary segmentation tasks.
- Training is conducted on NVIDIA®A100 80GB PCIe GPU.

3.1 Building Footprint Detection Using Sentinel-2 Images

Detection and monitoring of building footprints are essential in urban planning and reconstruction, illegal building detection, three-dimensional city modeling, and disaster monitoring and response applications. While major companies such as Google and Microsoft use high-resolution RS imagery to provide building footprint data available for different locations such as Africa (Sirko et al. 2021; Microsoft 2022), many research studies extract and update building footprints using freely available Sentinel-2 data (Prexl et al. 2023; Prexl and Schmitt 2023; Corbane et al. 2021). The global free-of-charge availability, decent resolution (10 meters), and the 5-day revisit time of Sentinel-2 images make it a great source to extract building footprints and update them for general purposes. However, the worldwide differences in the appearance, the construction material, and the surroundings of buildings complicate the training of a global building footprint detection model.

In this experiment, we challenge the incorporation of geolocation of information when detecting building footprints in two cities with a strongly heterogeneous appearance of their buildings.

3.1.1 Dataset Creation and Subdivision

A dataset containing Sentinel-2 data for Stuttgart, Germany, and San Francisco, USA is downloaded together with their corresponding Microsoft building footprints (Microsoft 2022) from the year 2021. Similar to the examples shown in the second row of Fig. 1, the cities have relatively different appearances in the images in terms of roof colors, materials, and average building heights. The scenes cover mostly central urban areas and are cut into 600 patches of 256×256 pixel images for each city. Inspired by Prexl and Schmitt (2023), both the labels as well as all 13 Sentinel-2 bands are resampled to 2.5 m GSD using bi-cubic resampling. In this experiment, the subdivision of the dataset is considered by the city boundaries.

3.1.2 Experiment

To conduct the experiments, a typical UNet as shown in Fig. 3a is used as a benchmark for regional and global (here: a dataset containing both cities) training. To incorporate the geolocation information, we use the deep-coded modification of UNet from Fig. 3b with two branches—each belonging to a city—in the prediction phase. For each city, the training and validation data includes 500, and 100 of 256×256 patches from the Sentinel-2 image of the city. All trainings are run for 100 epochs with a fixed learning rate of 0.001.

For all experiments, the loss function, optimizer, and batch size are set as Binary Cross Entropy loss, Adam optimizer, and 64 image samples.

3.1.3 Result

To evaluate the performance of the models, we conduct qualitative and quantitative evaluations of the validation

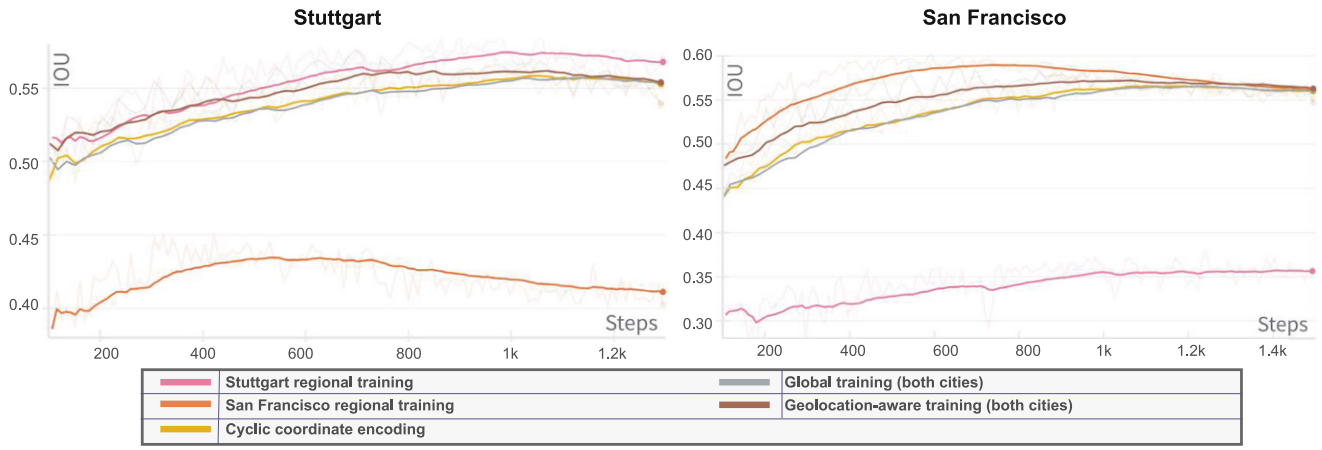

Fig. 4 The average IoU values on the validation set of each region for regional, global, and geolocation-aware training

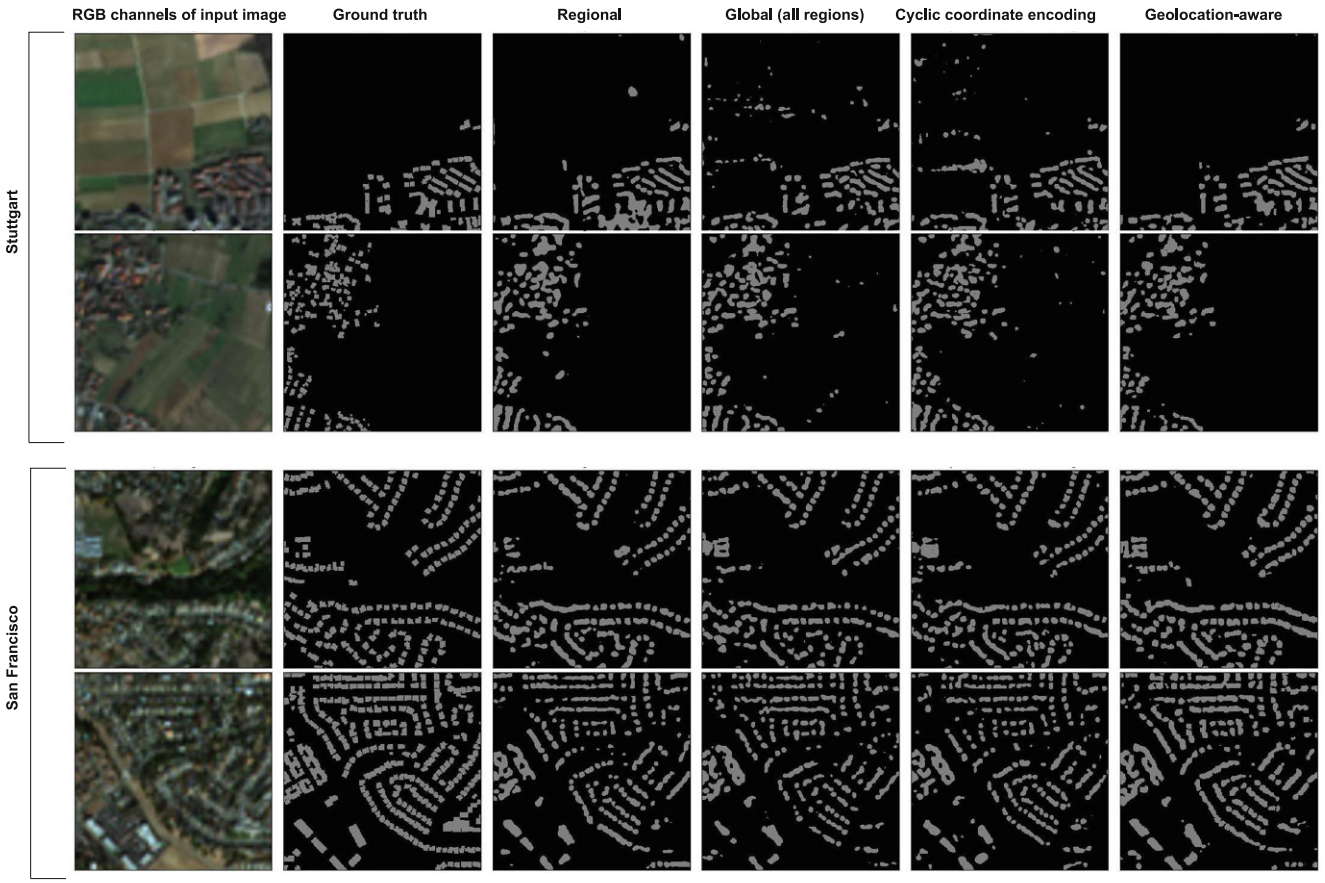

Fig. 5 Performance of the trained models on the validation data

data from both cities. The quantitative evaluation of all trained models is shown in Fig. 4 for each step of training on the evaluation set of each city. Because the number of training images is different in regional training than in other configurations, the evaluations are conducted after each step rather than epoch. As the batch size of all training is set to 64 images, a step represents the model performance after seeing 64 samples. Therefore, a stepwise comparison of the models with varying input sizes is more suitable than epochwise. Additionally, the stepwise representation allows the monitoring of the convergence speed and the stability of the trained models. The plots for each city contain a global model including both cities in the train set, a regional model that is only trained on the corresponding city, the cyclic

coordinate encoded model, and the proposed geolocation-aware model. The plots are only shown till the convergence point. The average IoU values of trained models on each city are also shown in Table 2 for final evaluation.

From Fig. 4, it can be seen that the best-performing model is the regional model which is trained on the corresponding city (see also Table 2). Achieving the highest IoU values on the validation set, the model outperforms all other configurations for both Stuttgart and San Francisco. The regional model that is trained on the nonfamiliar city, has a relatively weak performance on both regions. Converging to almost the same IoU value by the end of the training, the geolocation-aware model outperforms the global one by 0.01 in the validation set of San Francisco. While the cyclic coordinate encoding does not change the performance of the global model, the proposed geolocation-aware model follows an improved performance average during the training in both cities.

To have a visual evaluation, we tested the trained models on two sample images from each city as shown in Fig. 5. To have a relative perception of the scene, the RGB channels of input data are shown in the first column, followed by the ground truth building footprint labels in the second row. The predictions of the model which is trained on the corresponding city are shown in the 3rd column. The global model contains data from both cities, and the coordinate-encoded model and the geolocation-aware model are shown on the 4th–6th columns, respectively. Although the performance of all configurations seems to be reasonably good, there are fewer false positives and more precise predictions in the geolocation-aware predictions compared to the global and coordinate encoded models for both cities.

3.2 Forest Classification Using Single-Channel Thermal Imagery

High-resolution thermal emissions are the key to understanding and adapting to climate variability, managing water resources sustainably for agricultural production, mitigating health stress during heatwaves, predicting droughts, monitoring coastal and inland waters, and addressing natural hazards such as fires and volcanoes. Therefore, a range of high-resolution thermal missions including TRISHNA (Roujean et al. 2021), SBG-Thermal (Basilio et al. 2022), and LSTM (Koetz et al. 2018) are planned to be launched in the coming years. Furthermore, there is an increasing interest in new space initiatives preparing for smallSat missions and constellations for high-resolution thermal observations. As a result of increasing demand and interest in high-resolution thermal imagery, it becomes more crucial to have robust, reliable, and well-performing information extraction models adapted to the thermal domain.

Being highly correlated to thermal emissions, land cover information is implicitly present in a thermal image. A land cover is distinguishable in the thermal image if, a) the difference in thermal emission of the land cover to its surroundings is higher than the radiometric resolution of the sensor, and b) the size of the land cover is greater than half of the spatial resolution of the imaging sensor. While mostly used as auxiliary information to improve the land cover predictions from multispectral imagery (Abdalkadhum et al. 2020; Sinha et al. 2015; Eisavi et al. 2015; Zhao et al. 2019; Sun and Schulz 2015), high-resolution thermal imaging missions enable the classification using single thermal imagery.

In particular, the correlation between land cover and thermal emissions plays a crucial role when it comes to wildfire detection. Many imaging methods designed to identify the heat signature of fires rely on MWIR and TIR sensors. While smoke can serve as an indicator for detection, it can also obstruct the view of the flames. TIR imaging offers a distinct advantage in this aspect as thick smoke remains transparent at these wavelengths, enabling the detection of hotspots through smoke. This capability proves valuable in monitoring active fires and locating spot fires. Additionally, TIR imaging surpasses shorter wavelength infrared imaging by providing a limited dynamic range in the presence of fire, thereby facilitating the imaging of both the fire and background without sensor saturation (Allison et al. 2016). Therefore, it can be highly beneficial to continuously monitor wildfire-prone areas with thermal sensors.

To identify the presence of a wildfire, initially, the sensor should detect wildfire-prone areas namely forests. However, single thermal channel land cover mapping is a complex task. The reason is that thermal images do not provide rich spectral and polarization-based feature space. They also lack sharp textural information and are highly dependent on the geolocation of images and therefore, the temperature emissions. On the other hand, contrary to the RGB data, thermal imaging is also useful during the nighttime. Therefore, we chose this application to showcase that it is possible to classify forests from a single thermal image and evaluate the impact of geolocational awareness of the model in providing accurate predictions.

In this experiment, we focus on classifying forests from a single thermal band. For this purpose, a globally distributed dataset including the Landsat 8 band 10 thermal images together with their corresponding ESA Worldcover data (Zanaga et al. 2022) is created in Sect. 3.2.1. The global, regional, cyclic coordinate encoded, and geolocation-aware models are trained for the configurations of different subdivisions of datasets in Sect. 3.2.2. The quantitative and quantitative results are presented in Sect. 3.2.1 and discussed in Sect. 4.

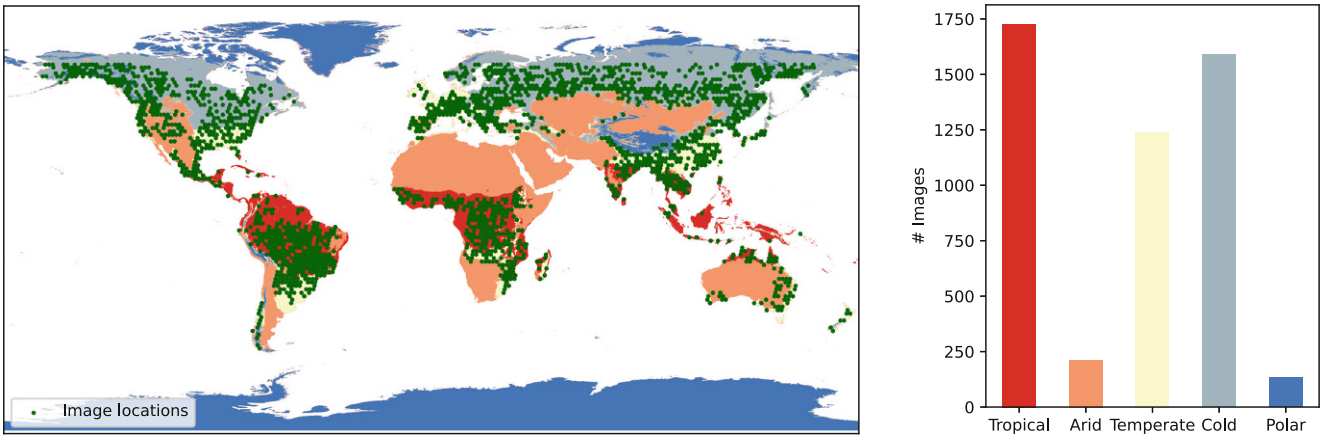

Fig. 6 The overview of the dataset. The image on the left illustrates the global spread of image locations and the image on the right shows the distribution of images over different climate regions

Table 1 The specifications of training configurations for $m = 3$ and $m = 5$ regions

Number of regions	Model	Learning rate	Schedueler	No. training images	No. validation images
$m = 5$	Tropical	0.001	None	1677	50
	Arid	0.001	Cosine annealing	84	50
	Temperate	0.001	None	1190	50
	Cold	0.001	Cosine annealing	1540	50
	Polar	0.001	Cosine annealing	162	50
	Global	0.001	None	4653	250
	Geolocatoin-aware	0.0001	None	4653	250
$m = 3$	Tropical	0.001	Cosine annealing	1427	300
	Temperate	0.001	Cosine annealing	940	300
	Cold	0.001	Cosine annealing	1290	300
	Global	0.001	Cosine annealing	3657	900
	Geolocation-aware	0.0001	None	3657	900

3.2.1 Dataset Creation and Subdivision

The dataset is created using globally distributed sparse data points for which Landsat thermal band 10 emissions together with corresponding ESA world data from 2021 are downloaded. Images have the size of 512×512 pixels and ESA World Cover data are downsampled to the pixel spacing of 100 meters. The images are selected in a way that they at least include 5% of tree cover data in their corresponding land cover labels. The tree cover class of ESA land cover contains any geographic area dominated by trees including forests. Thus, the tree cover class will also be referred to as the forest class in the context of this paper.

Considering Sect. 2.1 and the significant impact of the climate on the type, temperature, and appearance of forests, the well-known Köppen-Geiger climate classification scheme (Cui et al. 2021) is used to divide the dataset into subsets with similar characteristics. The Köppen-Geiger climate classification consists of 5 major climate regions namely, tropical, arid, temperate, cold, and polar, each covering several sub-climate regions. The major climate zones are an indicator of major forest types (i.e. tropical, temperate, and boreal) (Martone et al. 2018), and temperature expectations. Therefore, Köppen-Geiger's major climate zones are used for the division of the dataset.

The global distribution of the train and validation images and the distribution of data within 5 major climate regions are shown in Fig. 6 from left to right respectively. It can be seen that the majority of forest data are in the topical, temperate, and cold areas causing high imbalances in the different subsets.

3.2.2 Experiment

To validate the impact of region awareness through the proposed methodology in Sect. 2, we consider two main training configurations namely $m = 5$ and $m = 3$. The latter configuration includes data from tropical, temperate, and cold regions with the highest amount of forest data and provides a balanced training configuration. In both config-

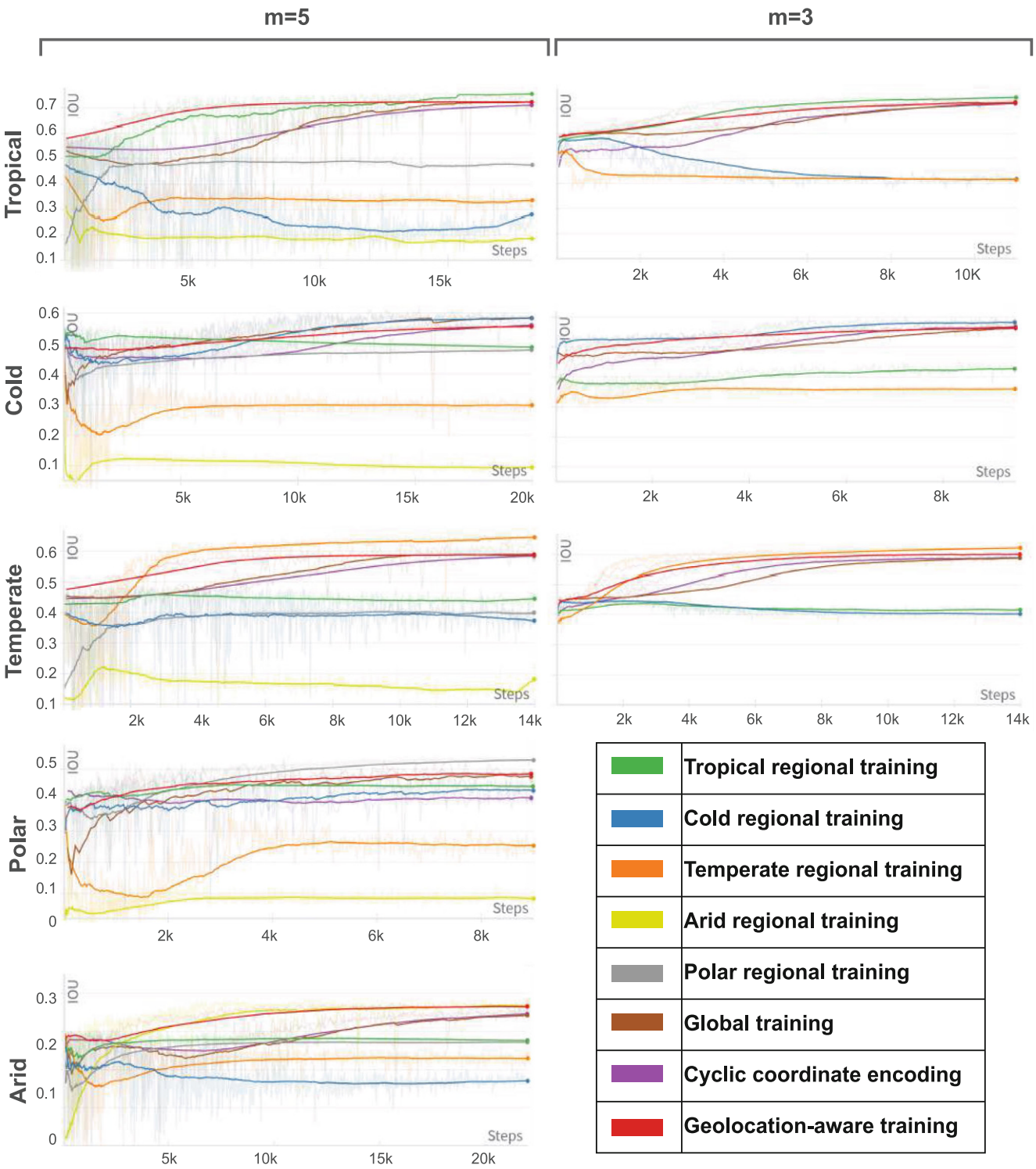

Fig. 7 The average IoU values on the validation set of each region for regional, global, and geolocation-aware training

urations, a global model—including data from all m regions, a cyclic coordinate encoded model, a geolocation-aware model, and m regional models—i.e., UNets trained on the corresponding region—are trained.

All experiments' loss function, optimizer, and scheduler (if used) are set as Binary Cross Entropy loss, Adam optimizers, and Cosine annealing respectively. For all the experiments, the batch size is fixed to 64 image samples and several learning rates with and without a scheduler are used

Table 2 The IoU evaluation of training on the validation set

Applications	Regions	Global	Cyclic coordinate	Geolocation-aware	Regional
Building footprint detection	Stuttgart	**0.56**	**0.56**	**0.56**	0.57
	San Francisco	0.56	0.56	**0.57**	0.59
Forest classification	Tropical	**0.72**	**0.72**	**0.72**	0.74
	Cold	0.55	**0.56**	**0.56**	0.58
	Temperate	0.59	0.59	**0.60**	0.62
	Tropical	**0.72**	0.71	**0.72**	0.75
	Cold	**0.57**	0.55	0.55	0.57
	Temperate	**0.59**	**0.59**	**0.59**	0.64
	Polar	0.47	0.41	**0.48**	0.53
	Arid	0.34	0.35	**0.37**	0.37

and the best-performing ones are selected as final training. The training parameters for each configuration are illustrated in Table 1.

3.2.3 Result

All trained models are evaluated on the validation set of each region. The quantitative results of the trained models are illustrated in Fig. 7 where the average IoU of each model is computed on the validation set after each step. Due to the varying number of training images in each configuration, the training steps are used to illustrate and compare the performance of models. This helps to have a comparison point where all models have been fed with a similar number of images. The plots for each region include the global model which is trained on all regions, the regional model which is only trained on the corresponding region, the cyclic coordinate encoded model, and the geolocation-aware model. The plots are only shown till the convergence point. The average IoU values of trained models on each region are also shown in Table 2 for final evaluation.

Fig. 7 shows that almost for all regions, the regional training is not only the quickest to converge but also achieves higher IoU values after convergence (see also Table 2). In polar, temperate, and tropical regions, the best-performing model remains the regional ones. In the tropical and temperate regions of $m = 5$ configuration, the global and geolocation-aware models converge to the same IoU values. In contrast, in the Polar and Arid regions, the geolocation-aware model outperforms the global and cyclic coordinate-encoded models. In the cold climate zone, however, the global model has a better performance (0.02) than the geolocation-aware one. In the $m = 3$ configuration, the geolocation-aware as well as the cyclic coordinate encoded models outperform the global model in the Cold region while converging to the same IoU value in the Tropical region. The proposed geolocation-aware model outperforms the other two configurations in the Temperate region. For each region, the regional training in other regions has the weakest performance.

To qualitatively evaluate the model, one sample validation image per region is predicted for the forest class with global, regional, cyclic coordinate encoded, and geolocation-aware models as shown in Fig. 9. From the predictions illustrated in this figure, it can also be seen that the regional predictions are relatively detailed and close to ground truth labels. The global model as well as the coordinate encoded models eliminate details and come up with general predictions. The geolocation-aware outputs, however, are rather detailed and close to the regional and therefore, the ground truth labels.

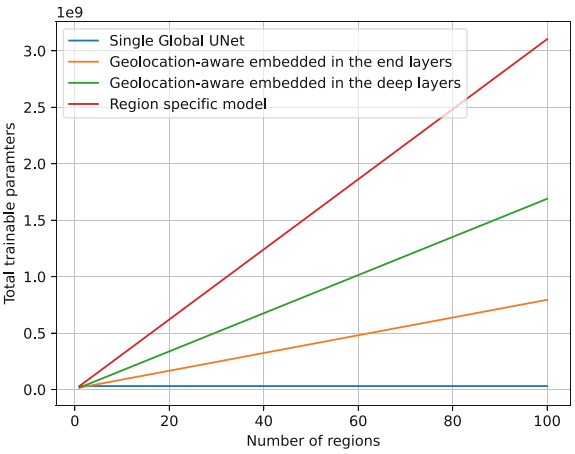

Fig. 8 The comparison of total parameters needed for a single globally trained UNet, the proposed architectures, and the regional models trained on each of the divided subsets

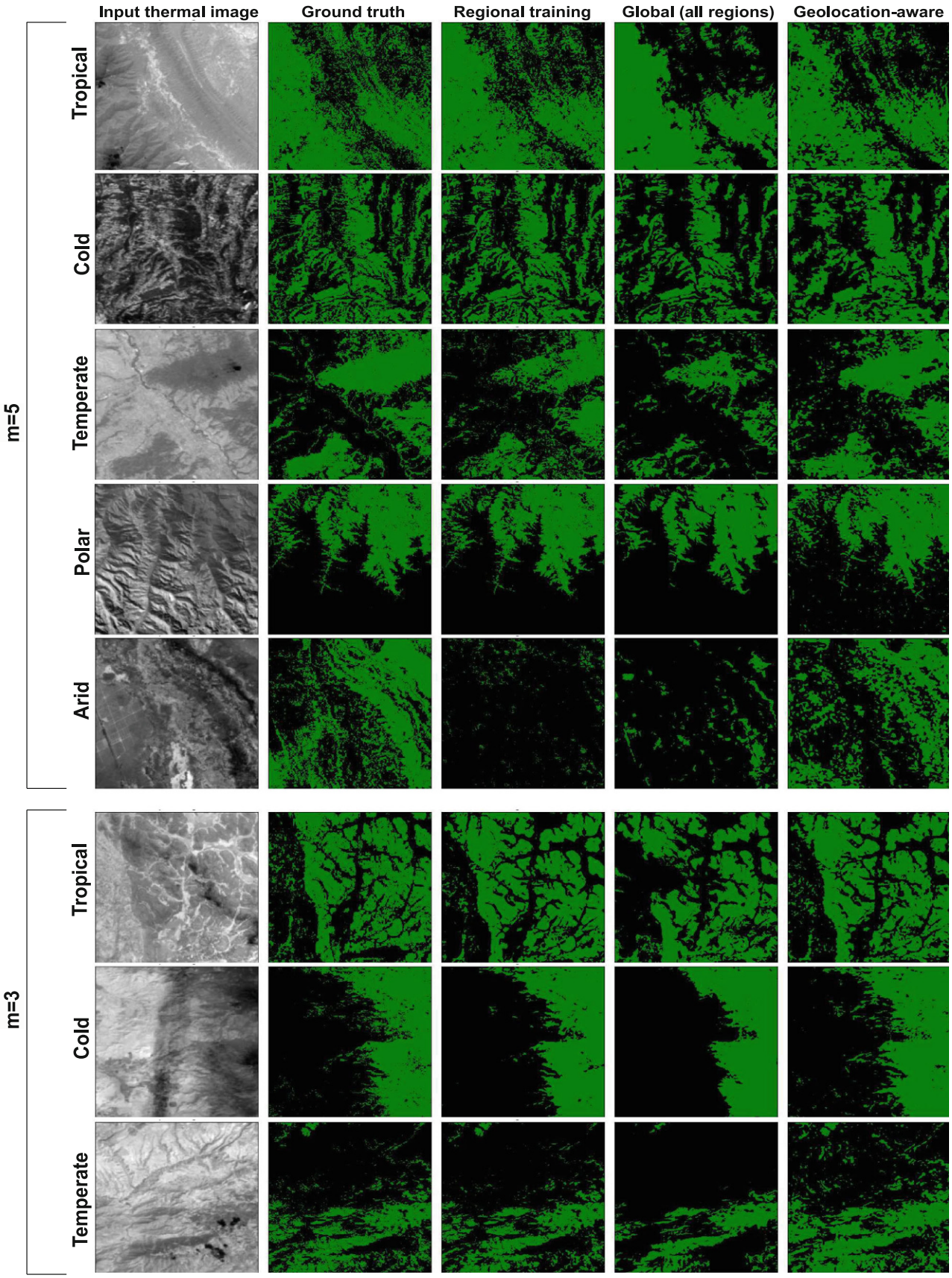

Fig. 9 Performance of the trained models on the validation data

4 Discussion

4.1 Computational Complexity

The computational complexity of proposed *deep coding* architectures is similar to that of standard UNet in the forward pass of batch processing in both training and inference modes. As the branches with no corresponding input within a batch are inactive, depending on the batch size bs, there will be bs times forward passes in both global and geolocation-aware models. The backward pass however is repeated k times (k being the number of regions from which the images are present in the batch). To optimize the backpropagation of the network in case large number of regions, we suggest selecting each batch from one region so that the the backpropagation will have an equal burden to one global UNet. To avoid the bias of the feature extractor module towards the latest batch, the batches in the last epochs should have a balanced distribution over regions.

To have an overview of the space complexity of the proposed method, Fig. 8 shows the trainable parameters needed for a maximum of 100 regions in a single-global UNet, proposed deep-coded architectures, and regional models where one UNet per region is trained. The *deep coding* modification in the deep layers of the network leads to a higher number of trainable parameters, whereas the architecture with the branches in the latter blocks of the network has a much smaller number of trainable parameters in comparison to the training of multiple regional networks. As the space complexity of the deep-coded architecture increases for a large number of regions, the balance between the number of subdivisions, the sufficient amount of data within each region, and the space complexity of the model should be observed when applying the methodology. To avoid an overcomplex model, we suggest having the *deep coding* modification in the latter parts of the network or further decreasing the feature map size of the branches.

4.2 Experiment Outcomes

Intuitively, very accurate predictions are expected from a model that is trained on one region and tested on the same region due to some extent of geolocational overfitting. Therefore, it is anticipated that the regional models outperform all other models. On the other hand, if a model is trained on one region and tested on an unalike region, we will expect a weak performance since the model is not exposed to any samples from the test region. Moreover, a global model that is exposed to samples from all regions is expected to be more generalizable and robust against overfitting. However, when training a global model, to achieve reasonable results on all regions, the model is forced to treat the region-specific characteristics as some sort of augmentation, and therefore overlook the region-specific features. This unwanted augmentation reduces the model's focus on region-specific details and therefore, decreases the model's performance to the regional models. We contemplate that the geolocation-aware model benefits from

- being exposed to samples from all regions and therefore, is generalizable and robust against overfitting
- meanwhile, considering region-specific features and using them to achieve higher performance in each region.

The plots in Figs. 4 and 7 confirm those assumptions. As expected, the regional models achieve the highest performance in both applications, followed by geolocation-aware and then either of the global or coordinate encoded models. The weakest performances belong to the regional models trained in unmatched regions.

In addition to the quantitative results, the visual results from Figs. 5 and 9 show that the regional models are capable of predicting building footprints and forest class with the highest level of detail, whereas the global and coordinate encoded models tend to generalize the predictions and eliminate details. The geolocation-aware model is relatively close to the regional model in catching the details.

Performing nearly as well as the regional ones, geolocation aware models are less prone to overfitting as they are fed with a higher number of training samples with diverse information. This can be observed in Fig. 4, where the regional training of San Francisco tends to overfit towards the end of the training.

The clustering strategy plays a crucial role in the effectiveness of geolocation-aware model training. In the building footprint use case, the dataset division is rather trivial whereas in the second use case, further analyses are needed for an optimal subdivision. The correlation of the climate zones to the forest formation and the impact of adding the climate regions to the training can be sensed from the result in Figs. 7 and 9. Both visual and graphical results confirm that when a model is trained on one climate region, it is better capable of detecting the forest from their thermal emissions and when the model has not seen data from that region, it fails to have precise extraction of forest data from thermal images.

The dataset of building footprint detection and the $m = 3$ region configurations of forest detection have a relatively balanced distribution of data over each region, whereas, in the $m = 5$ scenario, the polar and arid regions have drastically fewer images. The comparison of both scenarios shows that the geolocation-aware model can bear some extent of imbalances in the clustering of the dataset and perform desirably well. However, we do not recommend having drastic imbalances between regional subdivisions. The reason is that the low number of training data for a re-

gional branch can cause a low amount of backpropagations in the corresponding branch and can negatively impact the regional performance of the model. While data imbalance was not an issue within the scope of experiments in this paper, we suggest merging very small subsets with well-represented ones or applying best practices such as data augmentation, and weighted batch sampling to compensate for the lack of data in one region.

A critical conclusion that can be derived from comparing both experiments is that if the measurement data itself does not contain any clue regarding the image geolocation, the geolocation-aware model can significantly increase the understanding and therefore the performance of the model over all regions. For instance, in the forest detection use case, where only single band thermal emission measurements are provided to the model, realizing the geolocation of the image is not trivial and geolocation awareness can significantly increase the prediction performance. However, in case geolocational clues exist in the original measurements, e.g. by geolocation-specific differences in the visual appearance of building roofs, the global model can understand and incorporate the image location and improve its performance after a while. In this case, the geolocation-aware model reaches higher validation performances in earlier steps and leads to improved overall results. For example, in the building footprint application where 13 channels are fed into the model, the difference in roof colors and materials will be clear to the model after several epochs and it decodes the hidden geolocation information. However, the geolocation-aware model grasps the information in earlier steps and therefore has a higher IoU average over the validation set.

The method divides the dataset into hard clusters. However, geographical features change gradually rather than abruptly. The common CNN blocks as well as the skip connections help consider this graduality of transitional areas in the training and inference phase of the model. In the case of the neighboring regions with possible transitional areas as in the experiment of Sect. 3.2, we recommend using the architecture from Fig. 3c with common CNN blocks in feature extraction and prediction stages to avoid ignoring subtleties of transitional areas. In case of distinct separation of regions in the dataset with no neighboring regions similar to Sect. 3.1 where the cities are further apart, the suggested architecture in Fig. 3b can also be considered.

Fig. 7 highlights the averaged IoU values on the validation set. However, the fluctuations can still be seen in the stepwise predictions. Gradually decreasing over training, drastic performance fluctuations can be seen for almost all training configurations. The reason is the random batch sampling and batch-wise optimization of training after each step. The geolocation-aware models, however, are relatively stable with their predictions even in the first epochs of training. The reason is that the geolocation-aware model benefits from seeing more data in the feature extraction module and vigilant predictions based on the region index of images.

To summarize:

- **Advantages**. While trained and tested globally, the geolocation-aware model can catch the regional characteristics and details. It has a higher prediction power compared to the model with cyclic coordinate encoding. It has a relatively early convergence and enables single global model training instead of multiple regional models.
- **Disadvantages**. Additional effort should be made to thoroughly analyze the dataset and determine the proper clustering. The tradeoff between the amount of data in each region and the number of branches should be observed to avoid an ill-posed complex model.

The proposed methodology specifically focuses on spatial information due to its importance in EO data and tasks, nevertheless, the same strategy can be employed to incorporate temporal, spatial, and camera-related hyperparameters such as look angle. To avoid any regional bias in the mutual blocks of the network and further improve the model performance, a new batch sampling with an equal number of regional data may be employed.

5 Conclusion

In this paper, we propose a methodology that enables incorporating geolocational information into training a global model for EO tasks. To do so, we first subdivide datasets into regions with similar characteristics and then, introduce the *deep coding* strategy that involves assigning lighter deep branches to each geolocational region. The deep branches are only fed with their associated regional data in both forward and backward passes, and the geolocational characteristics of the corresponding region are implicitly learned. We examined our methodology in the applications of building footprint detection from Sentinel-2 bands and forest classification from a single thermal image. The city boundaries and climate regions are used for the division of the dataset, respectively. For both applications, we train and validate regional, global, and cyclic coordinate encoded, and the proposed geolocation-aware models, and compare the performances of each region. While regional models always achieve the best performances, the proposed geolocation-aware models have a relatively improved performance compared to the global and coordinate-encoded models. The results show that the methodology can be used to incorporate the geolocational characteristics of data in the training while keeping the advantage of model exposure to all samples. We also illustrated that employing the geolocation-aware deep coder improves the prediction performance

of global models by incorporating geolocational information.

Acknowledgements Not applicable.

Funding This document is the result of the research project funded by the Bavarian State Ministry for Economic Affairs, Regional Development, and Energy in the frame of the SERAFIM project (RaFo21-016-C).

Author Contribution Not applicable

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not applicable.

Code Availability The code can be provided upon request.

Conflict of interest Not applicable.

References

Abdalkadhum AJ, Salih MM, Jasim OZ (2020) Combination of visible and thermal remotely sensed data for enhancement of land cover classification by using satellite imagery. In: IOP Conference Series: Materials Science and Engineering, vol 737. IOP Publishing, p 12226

Allison RS, Johnston JM, Craig G, Jennings S (2016) Airborne optical and thermal remote sensing for wildfire detection and monitoring. Sensors 16(8):1310

Basilio RR, Hook SJ, Zoffoli S, Buongiorno MF (2022) Surface biology and geology (SBG) thermal infrared (TIR) free -flyer concept. In: IEEE Aerospace Conference (AERO), p 9843292

Bastani F, Wolters P, Gupta R, Ferdinando J, Kembhavi A (2023) SatlasPretrain: a large-scale dataset for remote sensing image understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 16772–16782

Beery S, Wu G, Edwards T, Pavetic F, Majewski B, Mukherjee S, Chan S, Morgan J, Rathod V, Huang J (2022) The auto arborist dataset: a large-scale benchmark for multiview urban forest monitoring under domain shift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 21294–21307

Camps-Valls G, Svendsen DH, Cortés-Andrés J, Mareno-Martínez Á, Pérez-Suay A, Adsuara J, Martín I, Piles M, Muñoz-Marí J, Martino L (2021) Physics-aware machine learning for geosciences and remote sensing. In: International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp 2086–2089

Cao X, Zhou F, Xu L, Meng D, Xu Z, Paisley J (2018) Hyperspectral image classification with Markov random fields and a convolutional neural network. IEEE Trans Image Process 27(5):2354–2367

Chen L, Fang B, Zhao L, Zang Y, Liu W, Chen Y, Wang C, Li J (2022) DeepUrbanDownscale: A physics informed deep learning framework for high-resolution urban surface temperature estimation via 3D point clouds. Int J Appl Earth Obs Geoinform 106:102650

Chen Y, Feng X, Fu B (2021) An improved global remote-sensing-based surface soil moisture (RSSSM) dataset covering 2003–2018. Earth Syst Sci Data 13(1):1–31

Corbane C, Syrris V, Sabo F, Politis P, Melchiorri M, Pesaresi M, Soille P, Kemper T (2021) Convolutional neural networks for global human settlements mapping from Sentinel-2 satellite imagery. Neural Comput Appl 33:6697–6720

Cui D, Liang S, Wang D (2021) Observed and projected changes in global climate zones based on Köppen climate classification. Wiley Interdiscip Rev Clim Chang 12(3):e701

Diligenti M, Roychowdhury S, Gori M (2017) Integrating prior knowledge into deep learning. In: IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp 920–923

Eisavi V, Homayouni S, Yazdi AM, Alimohammadi A (2015) Land cover mapping based on random forest classification of multitemporal spectral and thermal images. Environ Monit Assess 187:291

Ekim B, Schmitt M (2024) Mapping land naturalness from sentinel-2 using deep contextual and geographical priors. arXiv preprint arXiv:240619302

Food Agriculture Organization of the United Nations (2022) FAOSTAT statistical database. https://www.fao.org/faostat/en/data/QCL

Fritz S, See L, Perger C, McCallum I, Schill C, Schepaschenko D, Duerauer M, Karner M, Dresel C, Laso-Bayas JC et al (2017) A global dataset of crowdsourced land cover and land use reference data. Sci Data 4:170075

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778

Hooker J, Duveiller G, Cescatti A (2018) A global dataset of air temperature derived from satellite remote sensing and weather stations. Sci Data 5:180246

Koetz B, Bastiaanssen W, Berger M, Defourney P, Del Bello U, Drusch M, Drinkwater M, Duca R, Fernandez V, Ghent D, Guzinski R, Hoogeveen J, Hook S, Lagouarde JP, Lemoine G, Manolis I, Martimort P, Masek J, Massart M, Notarnicola C, Sobrino J, Udelhoven T (2018) High spatio- temporal resolution land surface temperature mission – a Copernicus candidate mission in support of agricultural monitoring. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp 8160–8162

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, vol 25. Curran Associates

Liu R, Lehman J, Molino P, Petroski Such F, Frank E, Sergeev A, Yosinski J (2018) An intriguing failing of convolutional neural networks and the coordconv solution. Adv Neural Inf Process Syst 31

Liu S, Wang H, Hu Y, Zhang M, Zhu Y, Wang Z, Li D, Yang M, Wang F (2023) Land use and land cover mapping in china using multimodal fine-grained dual network. IEEE Trans Geosci Remote Sens 61:4405219

Mahara A, Rishe N (2023) Integrating location information as geohash codes in convolutional neural network-based satellite image classification. IPSI Trans Internet Res 19:24–30

Mai G, Xuan Y, Zuo W, He Y, Song J, Ermon S, Janowicz K, Lao N (2023) Sphere2vec: a general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. ISPRS J Photogramm Remote Sens 202:439–462

Martone M, Rizzoli P, Wecklich C, González C, Bueso-Bello JL, Valdo P, Schulze D, Zink M, Krieger G, Moreira A (2018) The

global forest/non-forest map from TanDEM-X interferometric SAR data. Remote Sens Environ 205:352–373

Microsoft (2022) Microsoft building fooptprints. https://github.com/microsoft/USBuildingFootprints. Accessed 2022-11-01

Prexl J, Schmitt M (2023) The potential of Sentinel-2 data for global building footprint mapping with high temporal resolution. In: Joint Urban Remote Sensing Event (JURSE). IEEE, p 10144166

Prexl J, Saha S, Schmitt M (2023) High precision mapping of building changes using Sentinel-2. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp 6744–6747

Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention (MICCAI). Springer, pp 234–241

Roßberg T, Schmitt M (2023) A globally applicable method for NDVI estimation from Sentinel-1 SAR backscatter using a deep neural network and the SEN12TP dataset. J Photogramm Remote Sens Geoinformation Sci: 171–188

Roujean JL, Bhattacharya B, Gamet P, Pandya M, Boulet G, Olioso A, Singh S, Shukla M, Mishra M, Babu S et al (2021) TRISHNA: an Indo-French space mission to study the thermography of the earth at fine spatio-temporal resolution. In: IEEE International India Geoscience and Remote Sensing Symposium (InGARSS). IEEE, pp 49–52

Schmitt M, Ahmadi SA, Hänsch R (2021) There is no data like more data-current status of machine learning datasets in remote sensing. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp 1206–1209

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556

Sinha S, Sharma LK, Nathawat MS (2015) Improved land-use/land-cover classification of semi-arid deciduous forest landscape using thermal remote sensing. Egypt J Remote Sens Space Sci 18(2):217–233

Sirko W, Kashubin S, Ritter M, Annkah A, Bouchareb YSE, Dauphin Y, Keysers D, Neumann M, Cisse M, Quinn J (2021) Continental-scale building detection from high resolution satellite imagery. arXiv preprint arXiv:210712283

Sun L, Schulz K (2015) The improvement of land cover classification by thermal remote sensing. Remote Sens 7(7):8368–8390

Takeishi N, Kalousis A (2021) Physics-integrated variational autoencoders for robust and interpretable generative modeling. Adv Neural Inf Process Syst 34:14809–14821

Vega G, Pertierra LR, Olalla-Tárraga MÁ (2017) MERRAclim, a high-resolution global dataset of remotely sensed bioclimatic variables for ecological modeling. Sci Data 4:170078

Von Rueden L, Mayer S, Garcke J, Bauckhage C, Schuecker J (2019) Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning. Learning 18:19–20

Wang Y, Gong J, Wu DL, Ding L (2023) Toward physics-informed neural networks for 3D multi-layer cloud mask reconstruction. IEEE Trans Geosci Remote Sens: 4107414

Wei G, Krishnan V, Xie Y, Sengupta M, Zhang Y, Liao H, Liu X (2022) Physics-informed statistical modeling for wildfire aerosols process using multi-source geostationary satellite remote-sensing data streams. arXiv preprint arXiv:220611766

Wen Y, Gao T, Zhang J, Li Z, Chen T (2023) Encoder-free multi-axis physics-aware fusion network for remote sensing image dehazing. IEEE Trans Geosci Remote Sens: 4705915

Zanaga D, Van De Kerchove R, Daems D, De Keersmaecker W, Brockmann C, Kirches G, Wevers J, Cartus O, Santoro M, Fritz S, Lesiv M, Herold M, Tsendbazar NE, Xu P, Ramoino F, Arino O (2022) ESA Worldcover 10 m 2021 v200. Zenodo

Zhang F, Yan M, Hu C, Ni J, Zhou Y (2021) Integrating coordinate features in CNN-based remote sensing imagery classification. IEEE Geosci Remote Sens Lett 19:5502505

Zhao J, Yu L, Xu Y, Ren H, Huang X, Gong P (2019) Exploring the addition of Landsat 8 thermal band in land-cover mapping. Int J Remote Sens 40(12):4544–4559