On the Generalization of Representation Uncertainty in Earth Observation

Supplementary Material

A. Optimization of pretrained uncertainties

Loss prediction provides a general approach for uncertainty estimation, as any task's level of wrongness can be defined by its loss $\mathcal{L}_{task}$. In loss prediction tasks, an uncertainty module u is added after the representation layer of a standard supervised encoder, predicting the corresponding loss for each sample. Specifically, u is implemented as a small MLP head, on top of the model's representation $e(x)$ and is trained using an $\mathcal{L}_2$ loss between u and the task loss L_{task}. The main supervised task is learned together with the uncertainty module, using the objective:

$$\mathcal{L} = \mathcal{L}_{task}(y, f(x)) + (u(e(x)) - \mathcal{L}_{task}(y, f(x)))^2$$

Kirchhof *et al*. [31], adapted the loss prediction approach to develop pretrained uncertainties, introducing some modifications. Particularly, they propose two methods to perform the prediction task. First, they apply a stop-gradient mechanism before the uncertainty module to enable parallel training ensuring that its gradients do not affect the supervised classifier. Second, they pretrain the large-scale supervised classifier and extract its representations. The uncertainty module is trained on top of these frozen representations, enhancing computational efficiency. Since the task loss depends only on the representations, they can be cached once, accelerating training. In this study, we adopt the second approach.

Moreover, instead of relying on the $\mathcal{L}_2$ which is inherently tied to the scale of the supervised task loss, a ranking-based objective is introduced. This objective ensures that uncertainty values remain consistent across different tasks and loss scales. The ranking-based loss is defined as:

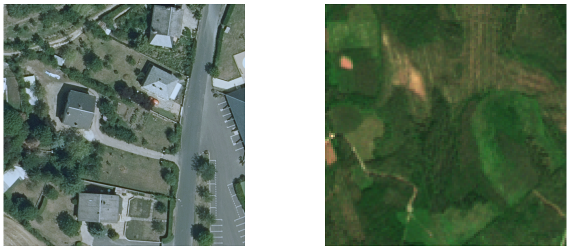
$$\mathcal{L} = \max(0, \mathbb{1}_{\mathcal{L}}(u(e(x_1)) - u(e(x_2)) + m)),$$

$$\text{s.t. } \mathbb{1}_{\mathcal{L}} = \begin{cases} +1, \text{ if } \mathcal{L}^{det}_{task}(y_1, f(x_1)) > \mathcal{L}^{det}_{task}(y_2, f(x_2)) \\ -1, else \end{cases}$$

$\mathbb{1}_{\mathcal{L}}$ is the indicator function, taking a value of $+1$ if the examined sample has higher task loss than a randomly selected sample in the batch, and -1 otherwise. During training the uncertainty values are adjusted accordingly, ensuring that the sample with higher task loss receives a higher uncertainty value, enforced by a margin of m. Following [31], we set $m = 0.1$.

B. Visual examples of Semantic Factors

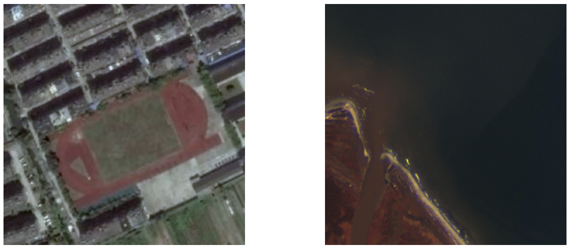
In this section, we visually inspect the variability induced by the semantic factors defined in Sec. 3.2. Figure 10 highlights the variability induced by varying the GSD (SF1).



(a) GSD: 20cm per pixel

(b) GSD: 10m per pixel

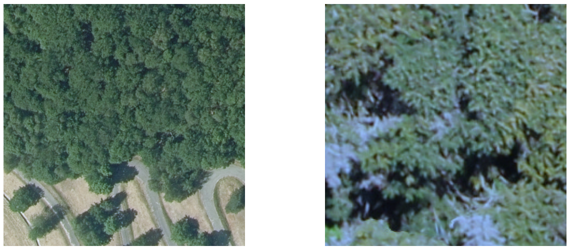
Figure 10. Example of variability induced by the GSD (SF1).



(a) Land-focused scene.

(b) Marine-focused scene.

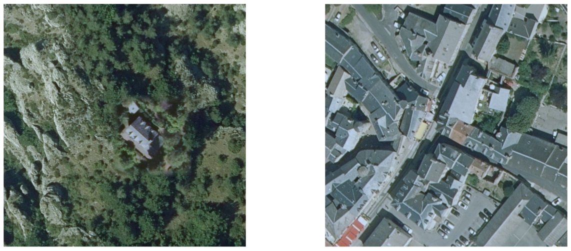
Figure 11. Example of the impact of the domain of interest (SF2).



(a) Target: Forest type.

(b) Target: Tree species.

Figure 12. Example of variability in target granularity (SF3).



(a) Mountainous forest area.

(b) Urban area.

Figure 13. Example of the impact of spatial arrangement (SF4).

Figure 10a presents a sample from the FLAIR dataset at 20cm per pixel, depicting a suburban area with fine detail, while Fig. 10b shows a sample from the BigEarthNet dataset at 10m per pixel, capturing coarser objects such as a forest. Figure 11 showcases the impact of the domain of interest (SF2) in EO scenes, presenting a sample from MLRSNet focused on land cover (Fig. 11a), and a sample from the MARIDA dataset (Fig. 11b), which focuses on marine environments. Clearly, the objects existing in these two domains differ significantly. Similarly, Fig. 12 shows the impact of the target granularity, highlighting the visual difference of high-level targets like forests (Fig. 12a), which require less detail to identify, and low-level targets like tree species (Fig. 12b) that demand the highest possible level of detail.

Notably, as seen in the images, some SFs are highly interdependent; *e.g.*, GSD (SF1) influences target granularity (SF3), as spatial resolution dictates the level of available detail.

C. Uncertainty Representation Metrics

***LA*@1 for multi-label classification:** Consider a sample x and its nearest neighbor in the representation space denoted as x^*. Let their class vectors be $\mathbf{c}, \mathbf{c}^* \in \{0,1\}^K$, where K is the number of classes. Each element in the class vectors indicates the presence (1) or absence (0) of a given class. In the context of multi-label classification, we define 3 metrics:
One-LA@1 assesses whether x and x^* share at least one common class and is defined as follows:

$$\text{One-LA@}1 = \mathbb{I}\left(\mathbf{c}^\top \mathbf{c}^* > 0\right),$$

where $\mathbb{I}$ is the indicator function, that equals 1 if the condition is true, and 0 otherwise.
All-LA@1 is a stricter criterion enforcing x^* to contain all classes present in x. Formally:

$$\text{All-LA@}1 = \mathbb{I}\left(\mathbf{c}^\top \mathbf{c}^* = \|\mathbf{c}\|_1\right),$$

where $\|\mathbf{c}\|_1$ denotes the sum of the elements in the class vector $\mathbf{c}$.
%-LA@1 quantifies the proportion of classes of x^* that match the classes of x:

$$\text{\%-LA@}1 = \frac{\mathbf{c}^\top \mathbf{c}^*}{\|\mathbf{c}\|_1}$$

This metric provides a balance between One-LA@1, which may be too lenient, and All-LA@1, which can be too strict.

***LA*@1 for semantic segmentation:** In the context of semantic segmentation, labels are represented as matrices $\mathbf{C}, \mathbf{C}^* \in \mathbb{R}^{m\times n}$, where m and n denote the height and width of the image respectively, and each element in the matrix corresponds to one of the K possible classes. To capture different notions of semantic similarity in segmentation tasks, we introduce 4 metrics:
All-LA@1 is a high-level metric, assessing whether the set of classes is shared between the images. The class vectors $\mathbf{c}, \mathbf{c}^* \in \{0,1\}^K$ are calculated, where each entry indicates the presence or absence of the respective class in the image. The metric is calculated as in the multi-label scenario:

$$\text{All-LA@}1 = \mathbb{I}\left(\mathbf{c}^\top \mathbf{c}^* = \|\mathbf{c}\|_1\right)$$

Patches$_p$-LA@1 evaluates the spatial similarity between x and x^*. Each image is divided into a $p \times p$ grid of patches. For each patch, the majority class is identified, resulting in two matrices $\mathbf{M}, \mathbf{M}^* \in \mathbb{Z}^{P\times P}$. Then, the metric computes whether corresponding patches in the two images share the same majority class:

$$\text{Patches}_p\text{-LA@}1 = \mathbb{I}\left(\mathbf{M_p} = \mathbf{M_p^*}\right)$$

This metric evaluates the spatial similarity between the images, by capturing the spatial alignment of their classes. In our study, we set $p = 3$ to capture high-level context, but this can be adjusted depending on the task's requirements.
Probability Distance of LA@1 *(PD-LA@1)* measures the contextual similarity between x and x^*, by comparing the class distributions across the entire image. The distributions $\mathbf{p}, \mathbf{p}^* \in [0,1]^K$ represent the percentages of each class in samples x and x^* respectively, and the metric is calculated as:

$$\text{PD-LA@}1 = 1 - HD(\mathbf{p}, \mathbf{p}^*),$$

where HD is the Hellinger Distance $HD(x,y) = \sqrt{\frac{1}{2}\sum_{i=1}^{N}\left(\sqrt{x_i} - \sqrt{y_i}\right)^2}$, which measures the similarity between two probability distributions. This metric is used to measure the contextual similarity between the two images, regardless of the spatial position of the classes.
Patches$_p$ Probability Distance of LA@1 *(Patches$_p$-PD-LA@1)* combines the spatial focus of Patches$_p$-LA@1 with the contextual similarity of PD-LA@1. The image is divided into a $p \times p$ grid of patches and for each patch $i,j \in \{1,\ldots,p\}$, the class distributions $\mathbf{p}_{ij}$ and $\mathbf{p^*}_{ij}$ are calculated. The HD_{ij} is then calculated for each i, j, and the final metric is the average of these distances across all patches:

$$\text{Patches}_p\text{-PD-LA@}1 = 1 - \frac{1}{P^2}\sum_{i=1}^{P}\sum_{j=1}^{P} HD_{ij}.$$

This metric captures both the spatial and contextual similarities between the two images, offering a comprehensive view of how well the two images align, both in terms of pixel position and overall structure.

The LA@1 for a dataset is calculated as the mean LA@1 across its samples. Binary LA@1 metrics (One-LA@1, All-LA@1, Patches$_p$-LA@1) quantify the percentage of representations whose nearest neighbor is semantically similar. In contrast, PD-LA@1 and Patches$_p$-PD-LA@1 represent mean probability distances, while %-LA@1 reflects mean proportions. This emphasizes that the values of the metrics are not directly comparable, as they measure different elements.

D. Coefficient of Predictive Ability (CPA)

Traditional ROC analysis is designed for binary classification tasks. The Universal ROC (UROC) curves and the associated Coefficient of Predictive Ability (CPA) extend this framework to any linearly ordered outcome, including binary, ordinal, mixed discrete-continuous, and continuous variables, thereby generalizing ROC analysis [18].

Generalizing the binary setting, the problem is transformed into a sequence of binary classification tasks. Given bivariate data (x_i, y_i) for $i = 1, \dots, n$, where x_i represents a predictor and y_i a continuous outcome, m unique values of y_i $z_1 < \cdots < z_m$ with $m < n$ are defined.

To construct the UROC framework, the real-valued outcomes are converted into binary indicators $\mathbb{1}\{y_1 \geq \theta\}, \dots \mathbb{1}\{y_n \geq \theta\}$ for threshold values $\theta \in \{z_2, \dots, z_m\}$. This results in $m - 1$ derived binary classification problems of the form

$$(x_i, 1\{y_i \geq z_{c+1}\}), \quad c = 1, \dots, m - 1.$$

Each of these $m - 1$ binary classification problems admits a standard ROC curve, which can be sequentially visualized as a ”ROC movie.” To summarize these ROC curves into a single representation, the UROC curve is defined as a weighted average of the individual ROC curves, providing a unified representation of predictive performance across continuous outcomes.

The CPA, defined as the area under the UROC curve, serves as a generalization of AUROC for continuous outcomes. Particularly, CPA is a weighted average of the AUROC values for the derived binary problems in the very same way that the UROC curve is a weighted average of the classical ROC curves that constitute the ROC movie. Notably, in the case of strictly binary outcomes, CPA reduces to AUROC, preserving interpretability within the classical ROC framework.

For further details, we refer the reader to the original work by Gneiting et al. [18].

E. Discard Test

The Discard Test is a diagnostic tool used to assess the quality of a model's uncertainty estimates by iteratively removing the most uncertain predictions from a test set and measuring the resulting change in model error. The fundamental principle behind this test is that if a model's uncertainty estimates are reliable, the most uncertain predictions should correspond to higher errors, thus removing them should lead to an improvement in overall model performance. The exact steps of the test are the following:

1. Model predictions are ranked in descending order based on their associated uncertainty estimates.
2. The ranked samples are divided into equal-sized batches according to a predefined discard fraction.
3. The most uncertain batch is removed from the set.
4. The model's error is recalculated on the remaining test samples.
5. Steps 3–4 are repeated iteratively until all samples have been discarded.

This process generates a curve that visualizes how the model error changes as more uncertain predictions are excluded. An effective uncertainty estimation method should result in a monotonically decreasing error curve, indicating that the most uncertain samples also tend to have higher errors. Deviations from these trends, such as non-monotonic error curves, suggest that the uncertainty estimates are not fully reliable, as removing uncertain predictions does not consistently enhance model performance. In this study, we use 200 discard fractions, so the steps are repeated 200 times for each dataset and pre-trained model. Moreover, we use the model loss as a measure of error.

The results are accompanied by the provision of Monotonicity Fraction (MF). MF measures how often model performance improves as more uncertain samples are discarded and is computed as:

$$MF = \frac{1}{N_f - 1} \sum_{i=1}^{N_f - 1} \mathbb{1}(\epsilon_i \geq \epsilon_{i+1}),$$

where $\mathbb{1}$ is the indicator function, and ϵ_i is the model error (here the loss) at discard fraction i. N_f denotes the total number of discard fractions considered. An MF value of 1 indicates perfect monotonicity. An ideal uncertainty estimation method would yield a high MF (indicating consistent performance improvement).

F. Overview of the experimental design

Table 2 summarizes the experimental design employed to evaluate the generalization of representation uncertainty. The table offers detailed references to the relevant investigation targets and the sections, figures, models, and dataset configurations that support each case, thereby facilitating traceability of the experiments.

G. Uncertainty Module training details

The uncertainty module was implemented as an MLP with two hidden layers, applied on top of the learned representations. Each linear layer is followed by a LeakyReLU activation, and the final layer uses a Softplus activation to ensure

Investigation Target	Section	Figure	Model	Pretraining Datasets	Inference Datasets
Impact of SF1	Sec. 4.2	Fig. 4	ViT-Large	BigEarthNet	All
Impact of SF1	Sec. 4.2	Fig. 5	ViT-Large	BigEarthNet	Flair
Impact of SF2	Sec. 4.1	Fig. 2	All	ImageNet, BigEarthNet, Flair	MARIDA
Impact of SF3	Sec. 4.2	Fig. 2	All	BigEarthNet-5	All
Impact of SF4	Sec. 4.2	Fig. 2	All	BigEarthNet-5	All
Usefulness of uncertainties	Sec. 4.3	Fig. 6	ViT-Large	ImageNet, BigEarthNet, Flair	All
Usefulness of uncertainties	Sec. 4.3	Fig. 7	ViT-Large	ImageNet, BigEarthNet, Flair	Flair
Localized Uncertainty	Sec. 4.4	Fig. 8	ViT-Large	BigEarthNet	Flair, MARIDA
Noisy Data	Sec. 4.5	Fig. 9	All	ImageNet, Flair	BigEarthNet, noisy BigEarthNet

Table 2. Overview of the experiments conducted in this study. Each investigation target is accompanied by references to the relevant sections, figures, models, and the corresponding pre-training and inference datasets used.

Figure/Table	Pretraining Dataset	Input Res.	Model	unc_width	weight decay	modality
Fig. 2	BigEarthNet	120	ViT-Tiny	512	0.1	RGB
			ViT-Small	512	0.1	RGB
			ViT-Base	512	0.1	RGB
			ViT-Large	512	0.1	RGB
Tab. 4	BigEarthNet	120	ViT-Tiny	512	0.1	SAR
			ViT-Small	512	0.1	SAR
			ViT-Base	512	0.1	SAR
			ViT-Large	512	0.1	SAR
Tab. 4	BigEarthNet	120	ViT-Tiny	512	0.1	MS
			ViT-Small	512	0.1	MS
			ViT-Base	512	0.1	MS
			ViT-Large	512	0.1	MS
Fig. 2	BigEarthNet-5	120	ViT-Tiny	256	0.01	RGB
			ViT-Small	512	0.1	RGB
			ViT-Base	512	0.01	RGB
			ViT-Large	512	0.1	RGB
Fig. 2	Flair	120	ViT-Tiny	256	0.5	RGB
			ViT-Small	256	0.5	RGB
			ViT-Base	256	0.5	RGB
			ViT-Large	256	0.5	RGB
Fig. 4	BigEarthNet	60	ViT-Large	512	0.5	RGB
Fig. 4	BigEarthNet	30	ViT-Large	512	0.01	RGB
Fig. 4	BigEarthNet	16	ViT-Large	512	0.1	RGB

Table 3. Model configurations and settings used for training the uncertainty modules. The ”Figure/Table” column indicates the corresponding figure or table in the main text or supplementary material associated with each experiment.

		ViT - Tiny		ViT - Small		ViT - Base		ViT - Large	
Metric	Modality	LA@1	LA-CPA	LA@1	LA-CPA	LA@1	LA-CPA	LA@1	LA-CPA
One	RGB	0.995	0.429	0.996	0.437	0.997	0.427	0.998	0.544
	SAR	0.975	0.435	0.978	0.427	0.979	0.403	0.981	0.418
	MS	0.996	0.402	0.997	0.366	0.997	0.431	0.997	0.353
All	RGB	0.526	0.657	0.592	0.646	0.664	0.616	0.716	0.607
	SAR	0.397	0.665	0.42	0.659	0.463	0.641	0.463	0.641
	MS	0.558	0.650	0.63	0.632	0.609	0.613	0.723	0.615
%	RGB	0.796	0.607	0.827	0.596	0.857	0.569	0.88	0.575
	SAR	0.714	0.600	0.727	0.595	0.748	0.582	0.748	0.582
	MS	0.815	0.600	0.845	0.583	0.834	0.571	0.885	0.562

Table 4. LA@1 and LA-CPA for Multispectral (MS), Synthetic Aperture (SAR) and RGB data for BigEarthNet pretraining and inference.

the uncertainties remain positive, as described in the original paper. The uncertainty width unc_width, *i.e.* the width of the linear layers, is set to either 256 or 512, tuned individually for each model. A full overview of the hyperparameters used across all experiments can be found in Tab. 3.

Our models were trained for 1000 epochs in each config-

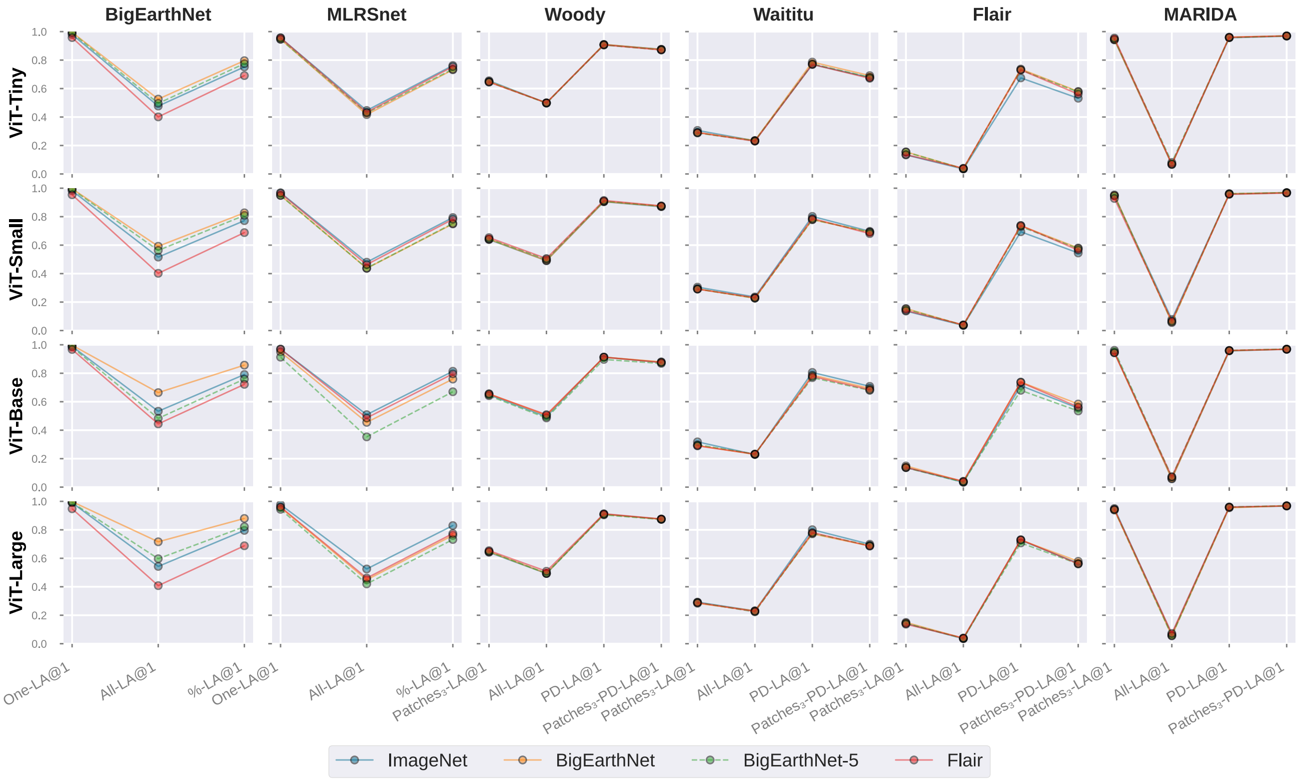


Figure 14. Performance evaluation of LA@1 for all ViT variants and inference datasets across all LA metrics, with higher values representing better performance. Colors indicate different pre-training datasets. EO-pretrained models and ImageNet pre-trained models show comparable behavior.

Dataset	Modality	ViT - Tiny		ViT - Small		ViT - Base		ViT - Large	
		micro F1 (%)	macro F1(%)	micro F1 (%)	macro F1(%)	micro F1 (%)	macro F1(%)	micro F1 (%)	macro F1(%)
BigEarthNet	RGB	73.25	59.65	73.78	61.94	73.19	60.11	73.32	60.81
	SAR	68.48	54.76	68.33	55.92	68.30	55.80	68.94	55.28
	MS	74.32	62.13	73.92	62.36	74.20	62.28	74.22	62.36
BigEarthNet-5	RGB	84.19	60.67	84.73	81.61	84.21	70.93	84.10	81.07
Flair	RGB	85.21	57.62	85.87	59.54	85.20	59.23	85.78	58.97
MLRSNet	RGB	91.50	92.37	91.91	98.45	91.45	98.35	98.46	91.94

Table 5. Performance of supervised models whose representations were used for training the uncertainty modules that were used for creating Fig. 2 of the main text and Fig. 14 of SM.

Input Res.	Modality	micro F1 (%)	macro F1(%)
120	RGB	73.78	61.94
60	RGB	70.54	57.74
30	RGB	66.00	52.06
16	RGB	60.26	44.58

Table 6. Performance of supervised models whose representations used for training the uncertainty models under varying GSD. These models refer to Fig. 4 of the main text.

Dataset	Modality	micro F1 (%)	meanIOU (%)
Flair	RGB	72.59	56.98
Marida	RGB	99.17	98.35
Waititu	RGB	84.20	72.72
Woody	RGB	93.45	87.71

Table 7. Performance of U-Net, with a ResNet-50 backbone, trained via supervised learning for semantic segmentation tasks. The alignment of zero-shot uncertainties was assessed with the losses of these models. They refer to Fig. 6 of the main text.

uration. The learning rate was warmed up with a constant value of 0.0001 for 50 epochs and then decayed to $1e-8$ for the remaining epochs. Weight decay was tuned separately for each model. AdamW was used as the optimizer with $\beta_1 = 0.8$ and $\beta_2 = 0.95$. No augmentations were applied during the training of the uncertainty module.

H. Multispectral & Synthetic Aperture Radar Data Pretraining

In the main text, we evaluated our framework's results using only the RGB channels to ensure a fair comparison with models pretrained on RGB optical data and to accommodate datasets with diverse spectral characteristics extending beyond RGB. However, recognizing the importance of applications beyond the RGB spectral bands, we also pretrain models on MS and SAR data and publicly release the pretrained weights. These models are pretrained on BigEarthNet, a dataset with both MS and SAR modalities, and their performance is evaluated on the same dataset, facilitating the comparison with the RGB training setup. In Tab. 4, we report the results of this evaluation. While LA@1 is consistently better for MS and RGB modalities compared to SAR, the LA-CPA is slightly better in SAR modalities, especially in larger models (ViT-Base, ViT-Large). This is a preliminary indication that our framework can effectively extend to other setups, yet further examination on additional datasets and setups is necessary to test the uncertainty generalization and draw a more definitive conclusion.

I. LA@1 across datasets

Figure 14 summarizes the LA@1 results across different metrics, similar to Fig. 2, which presents the LA-CPA results discussed in the main text. As highlighted in the main text, LA@1 remains consistent across models irrespective of the pre-training dataset, with ImageNet feature extractors producing robust representations that yield LA@1 values comparable to those trained on EO datasets.

J. Visualization of Samples with High/Low uncertainty

Figure 16 presents samples with high/low uncertainties across all datasets, as estimated by ViT-Large pretrained on Flair, providing a qualitative perspective on the performance of pretrained uncertainties.

K. Pretrained supervised models

In this section, we provide an overview of the performance of the supervised models used as backbone networks in our study. These models were used to extract the representations for training the uncertainty modules and to create the discard test plots. In Tab. 5, we summarize the results for the pretrained models used for creating Fig. 2 of the main text and Fig. 14 in the SM. The performance of the models used to investigate the impact of GSD on generalization, shown in Fig. 4 and 5 of the main text, is presented in Tab. 6. Finally, the results of the models used to examine the reliability of zero-shot uncertainties in downstream tasks, as shown in Fig. 6, are detailed in Tab. 7.

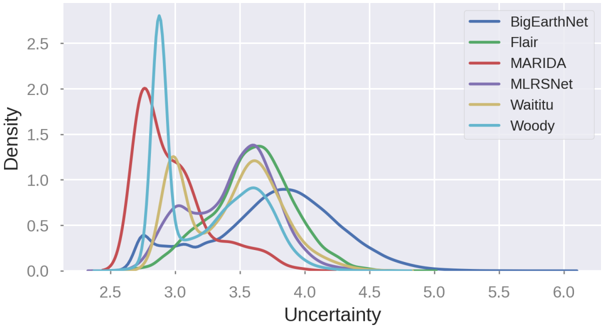


Figure 15. Densities of pretrained uncertainties, generated from a ViT-Large pretrained on BigEarthNet.

L. Localized Uncertainty Samples

In Fig. 17 we provide uncertainty samples together with their localized uncertainty estimates, as extracted from a ViT-Large pretrained on BigEarthNet.

M. Uncertainty distribution between upstream and downstream tasks

In Sec. 4.5 of [31] the authors showed that predicted uncertainties capture aleatoric, and not epistemic uncertainty. The presence of aleatoric uncertainty is validated in our experiments by Fig. 9, where the noisy subset of BigEarthNet exhibits higher uncertainty than its noise-free counterpart, despite coming from the same dataset. To further solidify the lack of epistemic signals, we replicate the analysis from [31] using ViT-Large pretrained on BigEarthNet and compare its uncertainty distribution with the ones in downstream tasks (Fig. 15). Despite the distribution shift, uncertainties on BigEarthNet span a wide range and are higher than those on downstream datasets, refuting the assumption that they reflect epistemic uncertainty. Notably, MARIDA exhibits the lowest uncertainty while coming from a very distinct EO domain.

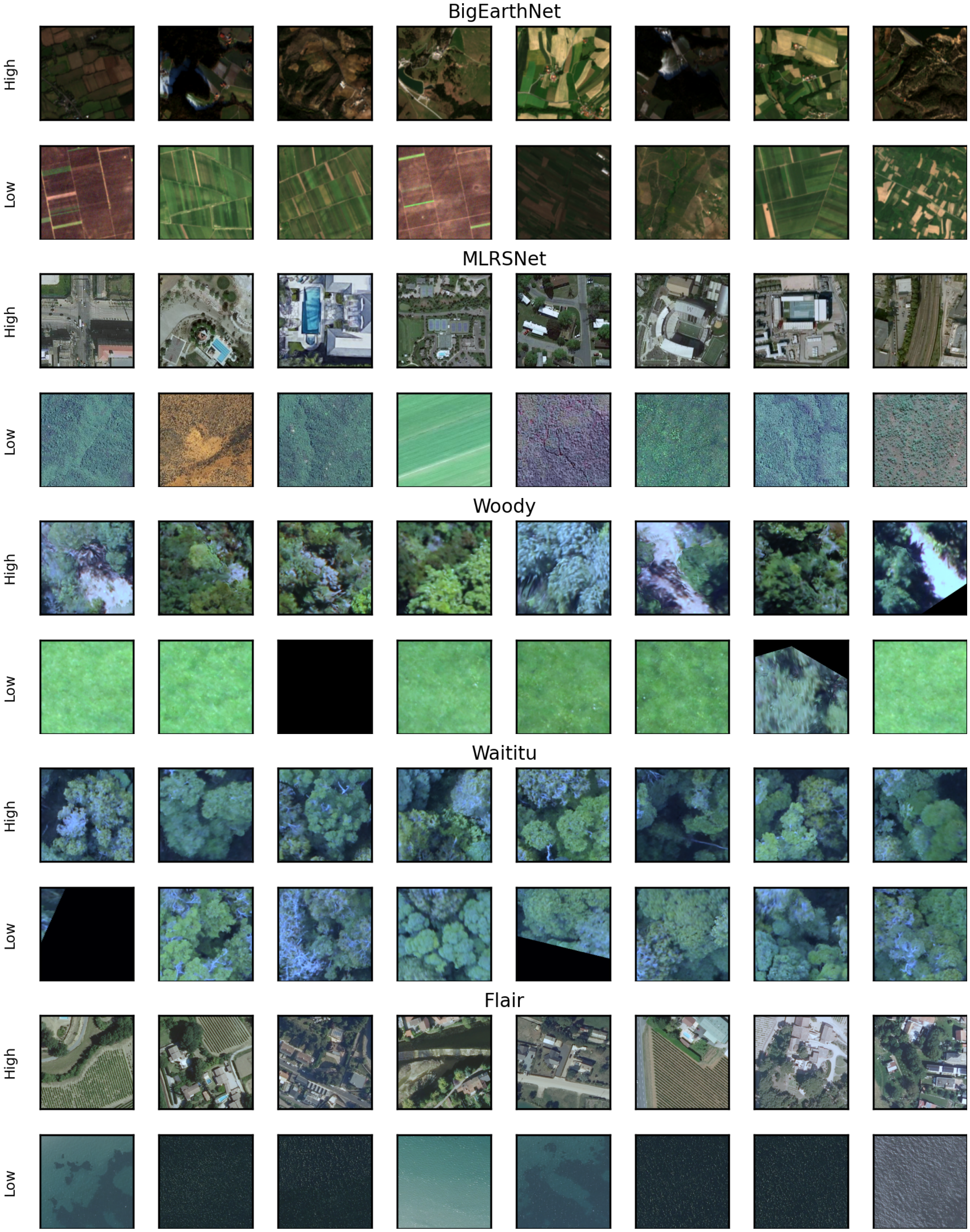


Figure 16. Samples with high/low downstream representation uncertainty across datasets used in this study. The uncertainty estimates were extracted from a ViT-Large pretrained on Flair.

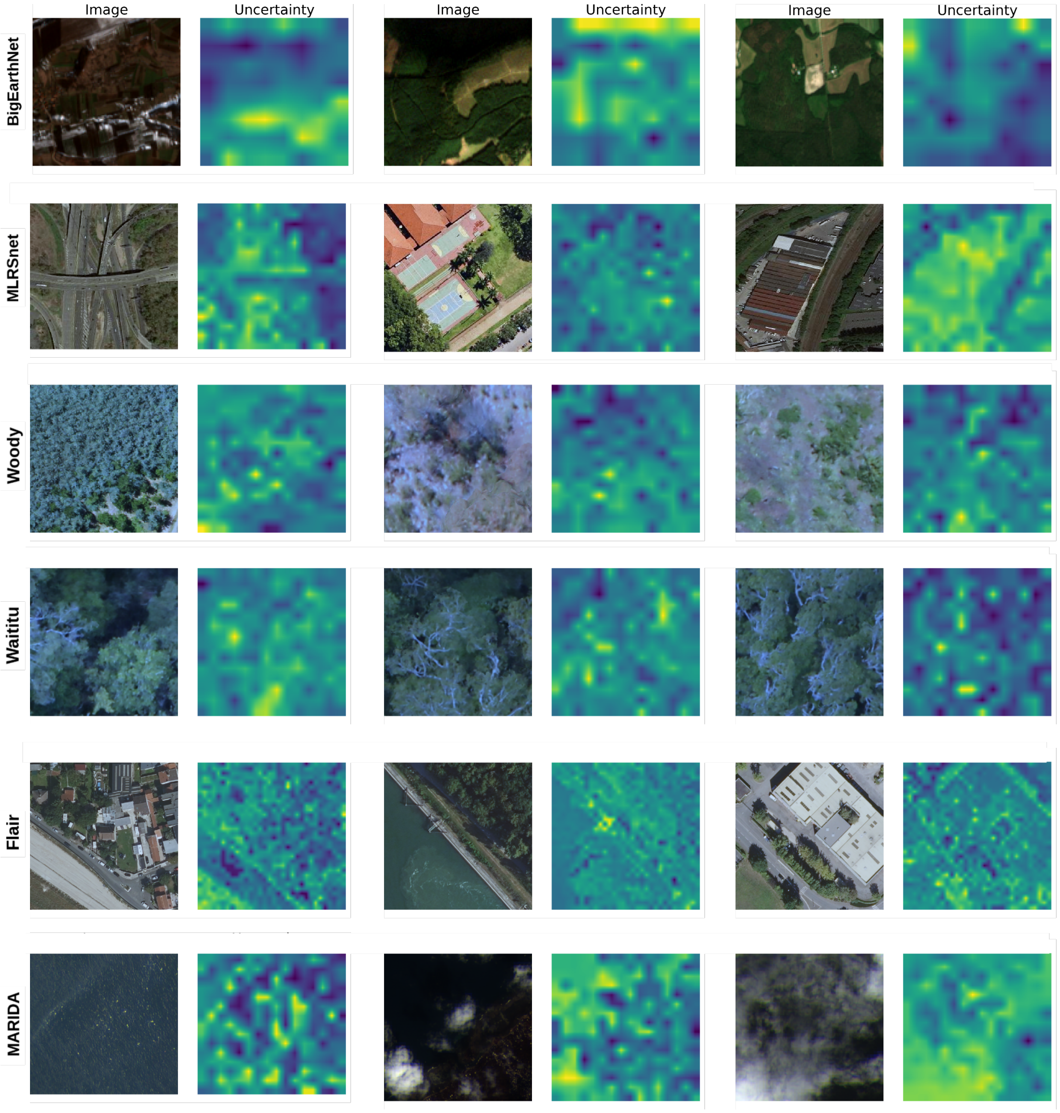


Figure 17. Example samples along with their zero-shot localized uncertainty estimates, as extracted by a ViT-Large pretrained on BigEarthNet.