Location Is All You Need: Continuous Spatiotemporal Neural Representations of Earth Observation Data

Supplementary Material

In this supplementary material, we provide additional technical and experimental details that complement the main paper.

- In Sec. 8, we describe the pretraining configuration of *LIANet*, including optimization settings and an analysis of hash collisions in the multiresolution hash tables.
- In Sec. 9, we outline the fine-tuning strategy across the downstream EO tasks presented in the main paper, including dataset descriptions, evaluation protocols, and complete results tables for $\mathcal{A}_+$ and $\mathcal{A}_{++}$. We further provide qualitative predictions and an ablation study on the effect of varying the fine-tuning area size.
- In Sec. 10, we report fine-tuning results on two datasets from the *PANGAEA* benchmark [35], and detail the corresponding experimental setup.
- Finally, in Sec. 11, we analyze *LIANet* from a neural compression perspective, quantifying its implicit encoding efficiency and reconstruction fidelity.

The source code and the dataset that is used for pretraining and fine-tuning of the proposed framework can be found in `https://github.com/mojganmadadi/LIANet/tree/v1.0.1`.

8. Pretraining Setup

To pretrain the proposed *LIANet* on different-sized areas, we use L_1 loss, AdamW optimizer with a base learning rate of 5×10^{-4}, and a Cosine learning rate scheduler that has 5 warm-up epochs. Every epoch is backpropagated with 1,024,000 randomly selected points, trained with a batch size of 64. The number of training epochs for $\mathcal{A}_0$ is 225, for $\mathcal{A}_+$ is 393, and for $\mathcal{A}_{++}$ is 897 where the training converges. The encoder layers of *LIANet-Base* contain *ResNet50* blocks, whereas *LIANet-Large* advantages from deeper layers of *ResNet101*.

Hash Collisions. Monitoring hash collisions is essential to ensure that the encoded embedding $\mathbf{g}_{x,y,t}$ remains discriminative for each spatial coordinate (x, y). At resolution level $\ell \in \{1, \dots, L\}$ with hash function $h_\ell(\cdot)$, a collision occurs when two distinct grid nodes map to the same table index:

$$(x_i, y_i) \neq (x_j, y_j) \quad \text{and} \quad h_\ell(x_i, y_i) = h_\ell(x_j, y_j). \quad (1)$$

Collisions become more likely at higher resolutions, where the number of grid nodes exceeds the hash table length. In principle, this may lead to conflicting gradient updates if spatially distant locations alias across all resolution levels and therefore share identical concatenated embeddings. However, the final representation is constructed by concatenating multi-resolution encodings, each generated with an independent random seed (see `LIANet/Pretraining/src/models/LIANet.py`). Consequently, simultaneous collisions across all levels are statistically unlikely. The chosen hyperparameters balance memory efficiency and representational capacity while keeping the empirical collision rate negligible.

9. Fine-Tuning on $\mathcal{A}_+$ and $\mathcal{A}_{++}$

We evaluate on five downstream tasks, including regression, binary, and multi-class segmentation, to assess the utility of the learned representations. Three visual sample patches from two seasons together with their reconstruction results of *LIANet-Base* and *LIANet-Large*, as well as their predicted labels for all tasks, are illustrated in Fig. 7.

Dynamic World Land Cover. Dynamic World provides near-real-time, $10\,\mathrm{m}$ global land-cover maps predicted from Sentinel-2 imagery [3]. Each scene is labeled into nine classes (Water, Trees, Grass, Crops, Shrub & Scrub, Flooded Vegetation, Built-up, Bare Ground, Snow & Ice). For our area of interest, we form a 6-class pixel-wise segmentation (0: Water, 1: Trees, 2: Grass, merging Shrub & Scrub, 3: Crops, 4: Built-up, 5: Bare—merging Snow & Ice). Because labels are available per timestamp, this task probes how temporal embeddings are exploited during fine-tuning.

Canopy Height. We use canopy height maps derived from high-resolution Maxar imagery [36]. This is a single-output regression task with one label per location across all timestamps, evaluating absolute height estimation from multispectral inputs.

Dominant Leaf Type. We use the Copernicus High Resolution Layer on tree cover/forests [13] labels that include three classes: broadleaf, coniferous, and no-forest. This multi-class segmentation leverages all 12 Sentinel-2 bands and tests the model's ability to discriminate forest functional types.

Building Coverage Percentage. We estimate the fraction of building coverage per $10\,\mathrm{m} \times 10\,\mathrm{m}$ pixel using labels derived from Microsoft building footprints (from

Table 3. The pixel-wise classification performance evaluation of two datasets from the PANGAEA benchmark, measured with Intersection-over-Union (IoU), Accuracy (Acc), and F1-score (all calculated with macro averaging), along with the tunable parameter counts. For all metrics reported, the corresponding **top three** performances are printed **bold**.

Task	Model / Setting	# Tunable Params (M)	IoU	ACC	F1
PASTIS	UNet / Micro UNet	17.3 / 0.49	0.18 / 0.13	0.26 / 0.19	0.26 / 0.19
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	**0.26** / 0.23 / **0.24**	**0.35** / 0.30 / **0.31**	**0.33** / **0.32** / **0.32**
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	0.23 / 0.21 / 0.19	0.29 / 0.28 / 0.25	**0.32** / 0.30 / 0.27
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.19 / 0.17 / 0.14	0.23 / 0.24 / 0.19	0.28 / 0.25 / 0.19
	LIANet-Modified	0.87	**0.34**	**0.46**	**0.46**
Burn Scars	UNet / Micro UNet	17.3 / 0.49	**0.45** / **0.47**	**0.65** / **0.62**	**0.62** / **0.62**
	TerraMind-base (Full / Frozen / Embedding)	92 / 5.0 / 0.39	0.39 / 0.41 / 0.35	0.58 / 0.60 / 0.52	0.55 / 0.57 / 0.51
	Prithvi v2-300 (Full / Frozen / Embedding)	313 / 7.7 / 0.46	**0.44** / 0.43 / 0.42	0.60 / **0.63** / 0.61	**0.59** / **0.59** / 0.58
	DOFA-Large (Full / Frozen / Embedding)	357 / 7.7 / 0.46	0.37 / 0.41 / 0.38	0.59 / 0.59 / 0.55	0.51 / 0.58 / 0.55
	LIANet-Modified	0.87	**0.47**	**0.63**	**0.63**

Table 4. (Top) Pixel-wise classification task performance measured with Intersection-over-Union (IoU), Accuracy (Acc), and F1-score (all calculated with macro averaging) along with the tunable parameter counts for the area of interest $\mathcal{A}_+$. (Bottom) Regression task performance (Mean Absolute Error (MAE) and Mean Squared Error (MSE)) along with the tunable parameter counts for the area of interest $\mathcal{A}_+$. The from-scratch trainings and FMs are evaluated in different configurations with a varying number of tunable parameters. For all metrics reported, the corresponding **top three** performances are printed **bold**. Due to the limited number of comparisons, we are not marking any models for the task of building footprint segmentation.

Task	Model / Setting	# Tunable Params (M)	IoU	ACC	F1
Dynamic World	UNet / Micro UNet	17.3 / 0.49	**0.75** / 0.66	**0.82** / 0.76	**0.84** / 0.76
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	**0.70** / **0.70** / 0.65	**0.78** / 0.77 / 0.72	**0.80** / **0.80** / 0.72
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	**0.69** / 0.65 / 0.58	0.77 / 0.73 / 0.68	**0.79** / 0.75 / 0.67
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.68 / 0.62 / 0.46	0.75 / 0.71 / 0.54	0.78 / 0.72 / 0.56
	LIANet-Base	0.5	**0.69**	**0.80**	**0.79**
	LIANet-Large	0.5	**0.70**	**0.80**	**0.80**
Dominant Leaf Type	UNet / Micro UNet	17.3 / 0.49	**0.82** / 0.78	**0.89** / 0.86	**0.90** / **0.87**
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	**0.79** / 0.78 / 0.75	**0.87** / 0.85 / 0.84	**0.88** / 0.86 / 0.85
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	**0.79** / 0.76 / 0.70	**0.87** / 0.84 / 0.79	**0.88** / 0.85 / 0.80
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.76 / 0.71 / 0.62	0.85 / 0.80 / 0.72	0.86 / 0.81 / 0.73
	LIANet-Base	0.5	**0.82**	**0.90**	**0.90**
	LIANet-Large	0.5	**0.83**	**0.90**	**0.90**
Building Footprint Segmentation	UNet / Micro UNet	17.3 / 0.49	0.76 / 0.68	0.87 / 0.79	0.84 / 0.77
	LIANet-Base	0.5	0.64	0.71	0.72
	LIANet-Large	0.5	0.67	0.87	0.76

Task	Model / Setting	# Tunable Params (M)	MAE	MSE
Canopy Height	UNet / Micro UNet	17.3 / 0.49	**0.049** / 0.065	**0.012** / 0.019
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	**0.048** / 0.050 / 0.110	**0.012** / **0.012** / 0.056
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	**0.048** / 0.053 / 0.110	**0.012** / 0.014 / 0.056
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.051 / 0.057 / 0.110	**0.013** / 0.016 / 0.056
	LIANet-Base	0.5	**0.049**	**0.011**
	LIANet-Large	0.5	**0.047**	**0.011**
Building Density	UNet / Micro UNet	17.3 / 0.49	**0.017** / 0.022	**0.006** / 0.012
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	0.022 / 0.023 / 0.023	**0.008** / **0.009** / **0.009**
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	**0.020** / 0.023 / 0.027	**0.008** / **0.008** / 0.010
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.022 / 0.024 / **0.021**	**0.008** / **0.009** / 0.013
	LIANet-Base	0.5	**0.021**	**0.009**
	LIANet-Large	0.5	**0.021**	**0.008**

Maxar/Airbus) [37]. This is a scalar regression task that requires resolving sub-pixel structure from a multispectral context.

Building Footprint Segmentation. We perform binary building segmentation at $2.5\,\mathrm{m}$ ground sampling. This task evaluates transfer to higher spatial resolution quality beyond the $10\,\mathrm{m}$ Sentinel-2 scale.

Fine-tuning Setup. We fine-tune both *LIANet-Base* and *LIANet-Large* across all downstream tasks using a batch size of 32 and input patches of 128×128 pixels. Each task is trained for 50 epochs with a Cosine learning rate sched-

uler that has 5 warm-up epochs. We employ cross-entropy loss for segmentation tasks, L1 loss for canopy height regression, and Huber loss for building coverage estimation. Learning rates are set to 1×10^{-3} for segmentation and 5×10^{-5} for regression. The training regions correspond to $500\,\mathrm{km}^2$ within each study area, resulting in 20% of $\mathcal{A}_0$, 10% of $\mathcal{A}_+$, and 4% of $\mathcal{A}_{++}$ used for training, with the remainder reserved for validation. This limited-data setup is designed to evaluate the robustness and data efficiency of the learned representations.

All benchmark models are trained under consistent configurations: batch size of 32, input patch size of 128×128 pixels, and 50 epochs. We employ cross-entropy loss for segmentation tasks, L1 loss for canopy height regression, and Huber loss for building density estimation. Optimization is performed using AdamW with a learning rate of 1×10^{-4} for segmentation and 5×10^{-5} for regression tasks. Training and validation splits follow the same setup as defined in Sec. 9.

Table 5. Comparison of the pixel-wise land cover classification performance measured with Intersection-over-Union (IoU), Accuracy (Acc), and F1-score (all calculated with macro averaging) with respect to the different training area sizes.

Model	Train Area (km^2)	IoU	ACC	F1
LIANet-Base	80	0.62	0.70	0.72
	160	0.66	0.75	0.75
	320	0.71	0.79	0.80
	500	0.72	0.82	0.81
	640	0.73	0.82	0.82
LIANet-Large	80	0.63	0.71	0.73
	160	0.65	0.74	0.75
	320	0.71	0.79	0.80
	500	0.72	0.81	0.81
	640	0.73	0.82	0.82

Ablation: Varying amount of annotated data during Fine-tuning. Table 5 presents the performance of the *LIANet-Base* and *LIANet-Large* models when fine-tuned on varying fractions of the target area for pixel-wise land cover classification. This experiment investigates how much the training area can be reduced while still maintaining reasonable accuracy. As shown in Tab. 5, although performance decreases with smaller training regions, both models still achieve satisfactory results even when fine-tuned on only 3% of the target area, $\mathcal{A}_0$.

Tab. 5 shows the performance of the *LIANet-Base* and *LIANet-Large* models in few-shot settings where the fine-tuning area varies from 25% to 3% of the target region. As observed, performance decreases as the fine-tuning area becomes smaller, yet both models maintain reasonable accuracy even with limited training data. This property is particularly valuable in scenarios with scarce labeled samples.

10. Fine-Tuning on Standard Benchmark Datasets

Selecting appropriate benchmarks for evaluating *LIANet* requires compatibility between the benchmark input modality and the pretraining modality of *LIANet*, the availability of georeferencing, and extensive spatial coverage. Since our framework models multispectral Sentinel-2 data within geographically contiguous regions, benchmarks must provide compatible imagery and sufficient spatial density. Datasets such as *Five Billion Pixels* [55], *DynamicEarthNet* [54], and *SpaceNet 7* [58] are based on very high-resolution commercial imagery (e.g., Gaofen-2, PlanetFusion, Planet), which differs substantially from the Sentinel-2 modality considered in this work. In addition, some of these datasets are not fully open-access, limiting their suitability for large-scale generative pretraining. Other benchmarks, such as *MADOS* [26], predominantly contain low-frequency spatial content (e.g., water bodies), making them less suitable for assessing the high-frequency reconstruction capabilities of *LIANet*. Furthermore, several datasets included in *GeoBench* [28] lack explicit georeferencing or require extensive preprocessing to ensure spatial alignment, which conflicts with the coordinate-based design of our approach. We therefore select two datasets from the *PANGAEA* benchmark [35], namely *PASTIS* [46] and *HLS BurnScars* [39], and adapt them to match the experimental protocol of this study.

LIANet-Modified. For benchmarking on *PASTIS* and *HLS BurnScars*, we introduce a modified configuration of *LIANet*. Since pretraining spans multiple locations and a larger number of timestamps, we increase the encoding capacity while keeping the overall parameter count comparable to *LIANet-Base*. The modified variant, referred to as *LIANet-Modified*, uses a feature dimension of $F = 128$, a hash table size of $T = 2^{19}$, and $N_{\text{grids}} = 13$ resolution levels covering a complete Sentinel-2 tile. As in the original setup, 12 spectral channels are reconstructed. To maintain a comparable total parameter count, the CNN decoder head is reduced accordingly.

***PASTIS* dataset** provides 19-class panoptic annotations of agricultural parcels together with multi-temporal Sentinel-2 image patches derived from four Sentinel-2 tiles in France. In total, the dataset contains 2,433 labeled patches. For our experiments, we select two of the four available tiles, namely `T31TFM` and `T32ULU`, and use *all* available patches and timestamps within these tiles. This corresponds to 1,279 patches, i.e., 52.6% of the full dataset. Within the selected tiles, 19 cloud-free timestamps are available for one location and 18 for the other. *LIANet-Modified* is pretrained on all timestamps of both tiles. The

Table 6. (Top) Pixel-wise classification task performance measured with Intersection-over-Union (IoU), Accuracy (Acc), and F1-score (all calculated with macro averaging) along with the tunable parameter counts for the area of interest $\mathcal{A}_{++}$. (Bottom) Regression task performance (Mean Absolute Error (MAE) and Mean Squared Error (MSE)) along with the tunable parameter counts for the area of interest $\mathcal{A}_{++}$. The from-scratch trainings and FMs are evaluated in different configurations with a varying number of tunable parameters. For all metrics reported, the corresponding **top three** performances are printed **bold**. Due to the limited number of comparisons, we are not marking any models for the task of building footprint segmentation.

Task	Model / Setting	# Tunable Params (M)	IoU	ACC	F1
Dynamic World	UNet / Micro UNet	17.3 / 0.49	**0.76** / 0.68	**0.83** / 0.76	**0.85** / 0.78
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	**0.73** / **0.72** / 0.63	**0.79** / **0.78** / 0.70	**0.82** / **0.81** / 0.71
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	0.71 / 0.68 / 0.61	**0.78** / 0.75 / 0.70	0.80 / 0.77 / 0.70
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.67 / 0.64 / 0.42	0.73 / 0.72 / 0.51	0.77 / 0.74 / 0.53
	LIANet-Base	0.5	0.70	**0.78**	0.79
	LIANet-Large	0.5	0.70	**0.78**	0.80
Dominant Leaf Type	UNet / Micro UNet	17.3 / 0.49	**0.79** / 0.76	**0.88** / 0.85	**0.88** / 0.85
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	**0.77** / 0.75 / 0.73	**0.86** / 0.83 / 0.82	**0.86** / 0.85 / 0.83
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	**0.77** / 0.73 / 0.68	**0.86** / 0.82 / 0.78	**0.86** / 0.83 / 0.79
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.74 / 0.69 / 0.59	0.83 / 0.78 / 0.70	0.84 / 0.79 / 0.71
	LIANet-Base	0.5	**0.78**	**0.87**	**0.87**
	LIANet-Large	0.5	**0.78**	**0.87**	**0.87**
Building Footprint Segmentation	UNet / Micro UNet	17.3 / 0.49	0.75 / 0.67	0.88 / 0.79	0.84 / 0.76
	LIANet-Base	0.5	0.60	0.65	0.68
	LIANet-Large	0.5	0.63	0.83	0.72

Task	Model / Setting	# Tunable Params (M)	MAE	MSE
Canopy Height	UNet / Micro UNet	17.3 / 0.49	**0.053** / 0.069	**0.013** / 0.020
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	**0.052** / 0.056 / 0.120	**0.013** / **0.014** / 0.060
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	**0.055** / 0.058 / 0.120	**0.014** / **0.015** / 0.060
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.058 / 0.062 / 0.120	**0.015** / 0.017 / 0.060
	LIANet-Base	0.5	**0.055**	**0.013**
	LIANet-Large	0.5	**0.055**	**0.013**
Building Density	UNet / Micro UNet	17.3 / 0.49	**0.021** / 0.027	**0.008** / 0.016
	TerraMind-base (Full / Frozen / Embedding)	102 / 15.5 / 0.39	0.027 / 0.028 / 0.028	**0.011** / **0.011** / **0.011**
	Prithvi v2-300 (Full / Frozen / Embedding)	324 / 20.3 / 0.46	**0.025** / 0.028 / 0.033	**0.010** / **0.011** / 0.012
	DOFA-Large (Full / Frozen / Embedding)	357 / 20.3 / 0.46	0.027 / 0.029 / **0.025**	**0.011** / 0.012 / 0.016
	LIANet-Base	0.5	0.027	0.012
	LIANet-Large	0.5	**0.026**	0.012

labels are provided at 10 m GSD with patch size 128×128 and are partitioned into five folds for cross-validation. We follow the official fold protocol, selecting each fold once for validation. The results reported in Tab. 3 correspond to the average performance across both tiles and all validation folds.

***HLS Burn Scars* dataset** contains 804 labeled wildfire burn-scar patches derived from Harmonized Landsat and Sentinel-2 (HLS) imagery (2018–2021), sparsely distributed across the United States. Since *LIANet* models a geographically contiguous Sentinel-2 tile, we select the tile containing the largest number of labeled samples. Within this tile, seven training patches and two validation patches are available, resulting in a total of 9 labeled patches used for fine-tuning. This corresponds to approximately 1.1% of the full labeled dataset. For each labeled patch, a cloud-free post-event Sentinel-2 acquisition is used. Temporally close labels may share the same post-event image. In total, seven timestamps are used for generative pretraining over the selected tile. The original train/validation split is preserved. To ensure spatial consistency, label masks are aligned to the generated imagery by matching resolution, coordinate reference system, and patch dimensions. The resulting performance is reported in Tab. 3.

Pretraining and Fine-Tuning Setup. For the pretraining on both tasks, the loss, optimizer, scheduler, and batch size are as reported in Sec. 8. We implement the fine-tuning on 50 epochs with cross-entropy loss, AdamW optimizer, batch size of 32, and a fixed learning rate of 1×10^{-4}. Baselines are fine-tuned using the same setup as *LIANet*, similar to Sec. 9. To improve learning stability for the end-to-end GFM methods on the relatively small BurnScars dataset, we reduce the number of channels in the U-Net decoder. This decreases the number of trainable parameters compared to the other benchmark tasks.

11. Neural Compression Analysis

Implicit neural representations are related to the field of neural compression [11, 18]. Although *LIANet* is not meant to serve as an EO image compressor, we study compression performance in terms of reconstruction error as complementary information. Given coordinates (x, y, t), *LIANet* is able to regenerate images $\mathbf{I}$ over an area $\mathcal{A}$. Consequently, the image data is encoded by $D \circ E$ in a number of neural network parameters consuming a given amount of B bytes of disk space. Provided the number of image channels C(=12 for Sentinel-2 L2A), and assuming a total number of P_c spatial pixels in channel $c = 1 \dots C$ over area $\mathcal{A}$ for which Q number of timestamps have been encoded, we derive the bits per pixel (bpp) according to:

$$1/\text{bpp} = \frac{Q}{8B} \sum_{c=1}^{C} P_c \quad . \tag{2}$$

It is worth noting that bpp $\propto 1/Q$. Compression methods that efficiently handle the temporal dimension t in (x, y, t) have the potential to significantly lower the bpp. *LIANet*'s number of parameters to encode time, $Q \cdot N_{\text{grid}} \cdot F$, is substantially smaller compared to the remaining number of parameters of the network, in particular wrt. the decoder D. The *peak-signal-to-noise-ratio* with the definition of spatio-spectral averaging, $\langle\cdot\rangle$, reads

$$\text{PSNR}(\mathbf{I}|\mathbf{S}) = 10 \cdot \log_{10}\left(\frac{\max \mathbf{I}^2}{\langle(\mathbf{I}-\mathbf{S})^2\rangle}\right) \quad . \tag{3}$$

It quantifies the deviation of the *LIANet*-generated image $\mathbf{I}$ against the corresponding Sentinel-2 L2A data $\mathbf{S}$ by computing the logarithm of the ratio: maximal image amplitude over the mean-squared error of reconstruction. In the limit where $\mathbf{I} \to \mathbf{S}$, PSNR$\to \infty$. If the *LIANet* image strongly deviates from the Sentinel-2 reference, PSNR$\to -\infty$. When the root mean square error (RMSE) $\sqrt{\langle(\mathbf{I}-\mathbf{S})^2\rangle} \gtrsim \max \mathbf{I}$ dominates over the maximum image signal strength $|\mathbf{I}|$, PSNR is negative. At PSNR=0, the RMSE is equal to the maximal amplitude of pixel values in the reconstructed image. For a maximum signal ten times stronger than the error, PSNR$\approx$20. For excellent reconstruction where PSNR$\gtrsim$40, the error is at least 100 times weaker than the maximum signal.

Figure 5 provides PSNR as a function of bpp for various flavors of *LIANet*. Our results demonstrate that the reconstruction quality reaches acceptable (PSNR>30, signal approx. 30 times stronger than mean error) to excellent (PSNR>40) levels with minor inaccuracies, with the latter often imperceptible to the human eye.

The advancement in neural compression will guide future design updates of *LIANet* to improve compression while maintaining downstream task utility. Based on the compression–quality trade-off summarized in Fig. 5, the best model to pick is currently *LIANet-Large*$_{\mathcal{A}_+}$: it provides enough parameter capacity to properly model a mid-size area of $|\mathcal{A}_+| = 5000km^2$ for (close to) excellent reconstruction across the entire Sentinel-2 spectral bands while keeping variability low[2], *i.e.*, the model deals well with diverse remotely sensed scenes. To further assess the quality of reconstructed images by two *LIANet* variants on $\mathcal{A}_0$, we illustrate two more examples in Fig. 6 with high-frequency details as well as clouds present in the scene. Each of the reconstructed images in Fig. 6 is generated with four forward passes and then stitching the patches horizontally.

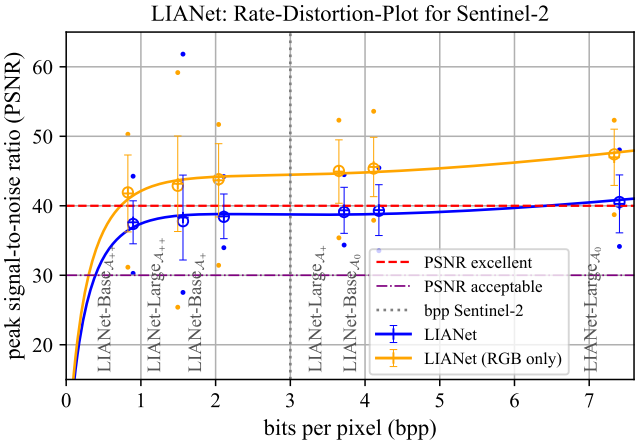

Figure 5. *Rate-Distortion analysis for LIANet's Sentinel-2 image generation.* In summary, our models perform decent (purple dash-dotted line) to excellent (red dashed line) with a distribution of PSNR-values derived from 100 randomly sampled spatial patches of size 128x128 pixels over the areas $\mathcal{A}_0$, $\mathcal{A}_+$, and $\mathcal{A}_{++}$. The label *LIANet-Large*$_{\mathcal{A}_{++}}$ for instance, refers to the *LIANet-Large* model that is pretrained on area $\mathcal{A}_{++}$. Blue and orange labels correspond to PSNR statistics for all Sentinel-2 bands and RGB channels, respectively. Dots indicate minimum and maximum values, bars cover the 1.5σ interval, the cross marks the mean, and the open circle indicates the median. The bold solid blue and orange lines provide a logarithmic interpolation PSNR(bpp) $\sim P_3\,[\log(1 + \text{bpp})]$ of the median PSNR (RGB) values with $Q_{\textit{LIANet}} = 4$ where we fixed PSNR(bpp=0)=0. $P_3[z]$ denotes a polynomial of third degree in z. A vertical dashed line (gray) indicates the bpp of the raw Sentinel-2 data in JPEG2000 format as a reference.

[2]The standard deviation of all models is about $\Delta\text{PSNR} \gtrsim 2$ around PSNR=40 which corresponds to signal-to-noise fluctuations of about two orders of magnitude.

Figure 6. Two examples of reconstruction quality of *LIANet-Base* and *LIANet-Large* on $\mathcal{A}_0$ area. It can be seen that *LIANet-Large* has an improved performance in reconstructing the high-frequency details, including buildings within the city. It can also be seen that cloudy areas are reconstructed as they are present in the Sentinel-2 image.

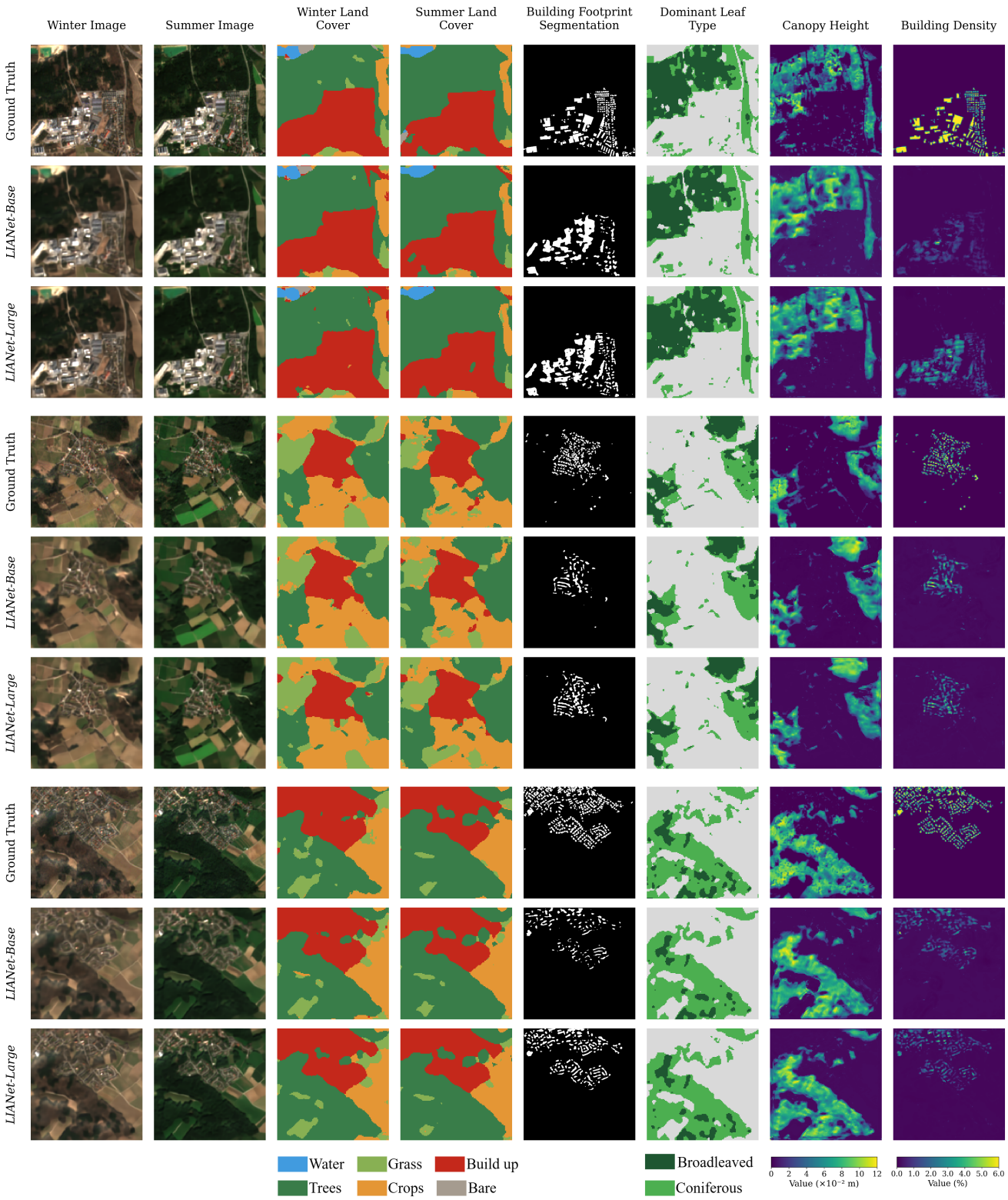

Figure 7. Three example visualization of *LIANet-Base* and *LIANet-Large* reconstruction and prediction performances on downstream applications. The landcover labels have seasonal predictions, whereas the other tasks have a single label for all timestamps. The improvement in the prediction of *LIANet-Large* can clearly be seen in the tasks concerning the segmentation and density estimation of building footprints.