50 Cities to visit before you classify - A world-wide multi-modal dataset for semantic segmentation of remote sensing imagery

Supplementary Material

Table 5. Dataset overview of 33 labeled cities.

Continent	City	Width [px]	Height [px]	Labeled [%]
Africa	Cairo	6,282	5,765	77.9
	Cape Town	2,353	2,228	79.3
	Johannesburg	2,180	2,119	73.0
	Nairobi	2,449	2,209	73.4
Asia	Istanbul	3,446	2,195	69.1
	Kyoto	4,362	4,537	96.0
	Mecca	3,955	3,766	74.3
	Mumbai	2,246	2,401	88.9
	Shanghai	3,904	4,504	49.7
	Wuhan	4,148	3,200	51.3
Europe	Barcelona	3,431	3,522	86.8
	Berlin	2,726	4,435	78.5
	Lugano	1,587	1,481	79.2
	Matterhorn	1,957	1,417	93.9
	Oslo	2,357	4,395	74.5
	Reykjavik	2,066	2,252	69.6
	Rome	3,464	4,483	62.0
	Wolfsburg	1,324	2,192	66.7
North America	Calgary	4,280	4,872	68.1
	Los Angeles	4,570	3,936	75.8
	Mexico City	3,674	3,321	75.6
	New York	2,796	3,386	70.4
	San Francisco	1,381	1,310	81.5
	Toronto	3,357	4,420	81.9
Oceania	Auckland	3,179	4,027	71.7
	Melbourne	4,482	3,789	78.3
	Perth	2,195	2,193	81.0
	Sydney	4,129	3,251	84.7
South America	Buenos Aires	2,516	2,737	88.0
	Itaituba	2,844	2,763	90.2
	Lima	4,329	3,697	71.1
	Rio de Janeiro	2,021	2,236	88.2
	Santiago	2,822	2,160	72.0
Total	**33**	**Area: 35,927 km^2**		**75.2**

Table 6. Dataset overview of 17 unlabeled cities.

Continent	City	Width [px]	Height [px]
Africa	Kigali	3,559	3,328
	Kinshasa	3,106	2,323
	Lagos	3,319	3,314
	Marakesh	2,388	1,751
	Kampala	3,101	4,101
Asia	Bangkok	6,304	4,382
	Jakarta	3,088	2,998
	Kuala Lumpur	2,232	2,761
Europe	Athens	1,705	1,827
North America	Denver	3,082	2,611
	Houston	5,549	2,882
	Kansas City	3,800	5,064
	Port-au-Prince	2,163	3,291
	San Jose	2,104	2,196
Oceania	Suva	1,574	1,783
South America	Caracas	4,402	3,291
	Montevideo	2,269	1,685
Total	**17**	**Area: 16,925 km^2**	

Table 7. Relative proportions of each class by continent.

Class Label	Relative Proportion [%]						
	Africa	Asia	Europe	North America	Oceania	South America	**Global**
LV-B	0.57	1.70	6.00	2.31	0.69	4.08	2.56
LV-NB	5.28	2.45	5.03	15.93	8.63	3.89	7.21
HV-B	0.10	0.36	1.83	1.62	0.20	–	0.78
HV-NB	1.82	24.76	25.28	0.95	12.75	15.56	13.63
ULD	17.70	13.66	12.67	13.32	16.92	5.38	13.46
UMD	17.52	9.70	6.87	30.77	22.89	26.62	18.73
UHD	0.49	8.75	4.75	4.49	0.75	4.01	4.24
I	2.61	4.57	3.84	7.94	6.50	3.48	5.00
W	4.73	14.48	17.25	17.82	30.30	27.86	18.08
SI	–	–	3.84	0.73	–	–	0.82
R	5.49	4.74	1.95	0.17	0.02	6.42	2.94
BS	2.43	4.00	10.56	3.60	0.16	1.83	4.03
S	41.24	10.83	0.14	0.17	0.19	0.87	8.48

6. Scene statistics

Labeled Cities (33 total). The labeled subset (Tab. 5) spans six continents and covers 35,927 km^2 in total. City footprints range from roughly 1.3k–6.3k px in width and 1.3k–5.8k px in height. The proportion of annotated area varies substantially across locations, with a mean labeled ratio of 75.2%. Europe (8 cities) and North America (6 cities) contribute the largest number of labeled scenes, while Africa and Oceania each provide 4, and South America provides 5. Several cities, e.g., Kyoto, Buenos Aires, and Rio de Janeiro, exceed 88% label coverage, whereas others such as Shanghai, Wuhan, Rome, and Wolfsburg exhibit notably lower coverage (≈50–67%).

Unlabeled Cities (17 total). The unlabeled portion (Tab. 6) comprises 17 cities across all continents, covering 16,925 km^2. Spatial extents range from approximately 1.6k–6.3k px in width and 1.7k–5.1k px in height. Africa and North America each contribute five cities, Asia contributes three, and Europe, Oceania, and South America contribute one or two each. This set includes several large metropolitan regions such as Bangkok, Houston, and Kansas City.

7. Label statistics

The relative proportions of semantic classes vary across continents (Tab. 7). The most common classes are *Urban Medium Density* (18.7%), *Water* (18.1%), and *Sand* (8.5%), followed by substantial amounts of *High Vegetation, Non-Bare* (13.6%), *Urban Low Density* (13.5%), and *Low Vegetation, Non-Bare* (7.2%). Each continent exhibits a distinct composition: Africa contains a large share of *Sand* (41.2%) plus notable *Urban Low* and *Urban Medium Density*; Asia shows high *High Vegetation, Non-Bare* (24.8%) and *Water* (14.5%); Europe features substantial *High Vegetation, Non-Bare* (25.3%), *Water* (17.3%), and *Bare Soil* (10.6%); North America is dominated by *Urban Medium Density* (30.8%) and *Low Vegetation, Non-Bare* (15.9%); Oceania contains large shares of *Water* (30.3%) and *Urban Medium Density* (22.9%); and South America has notable *Water* (27.9%), *Urban Medium Density* (26.6%), and *High Vegetation, Non-Bare* (15.6%), with the lowest *Urban Low Density*. *Snow and Ice* appears only in Europe and North America and remains rare (¡1%). Overall, these patterns highlight the dataset's broad land-cover diversity.

8. Example Scenes

The example scenes in Figures 5–10 illustrate the geographic and visual diversity represented in the dataset. For each selected city, we show the SAR input, the corresponding optical reference, and the annotated semantic labels. The examples span all continents and highlight a wide range of urban forms, land-cover types, and environmental conditions, underscoring the dataset's global scope and heterogeneity.

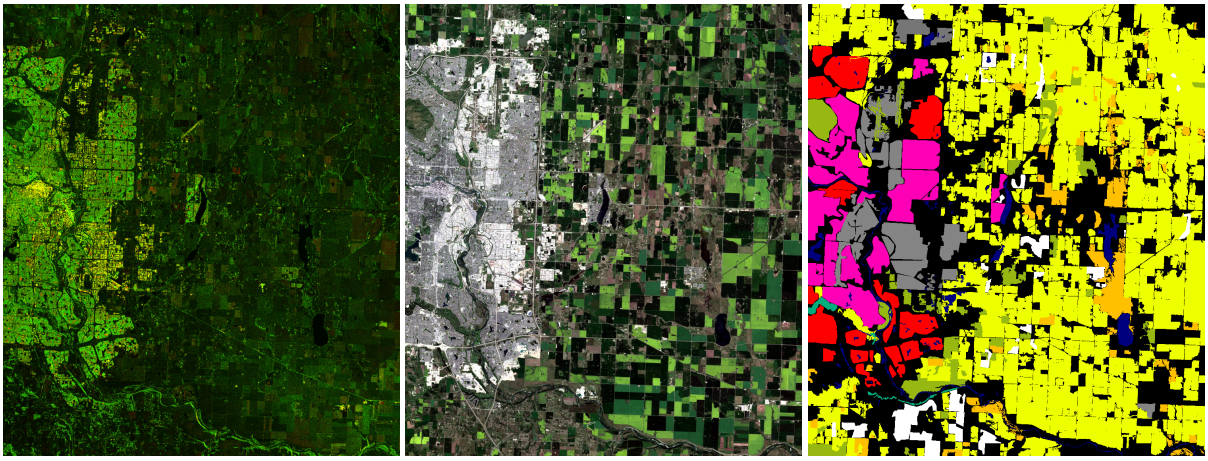


Figure 5. North America - Calgary

Figure 6. South America - Rio de Janeiro

Figure 7. Australia & Oceania - Sidney

Figure 8. Asia - Wuhan

Figure 9. Africa - Cape Town

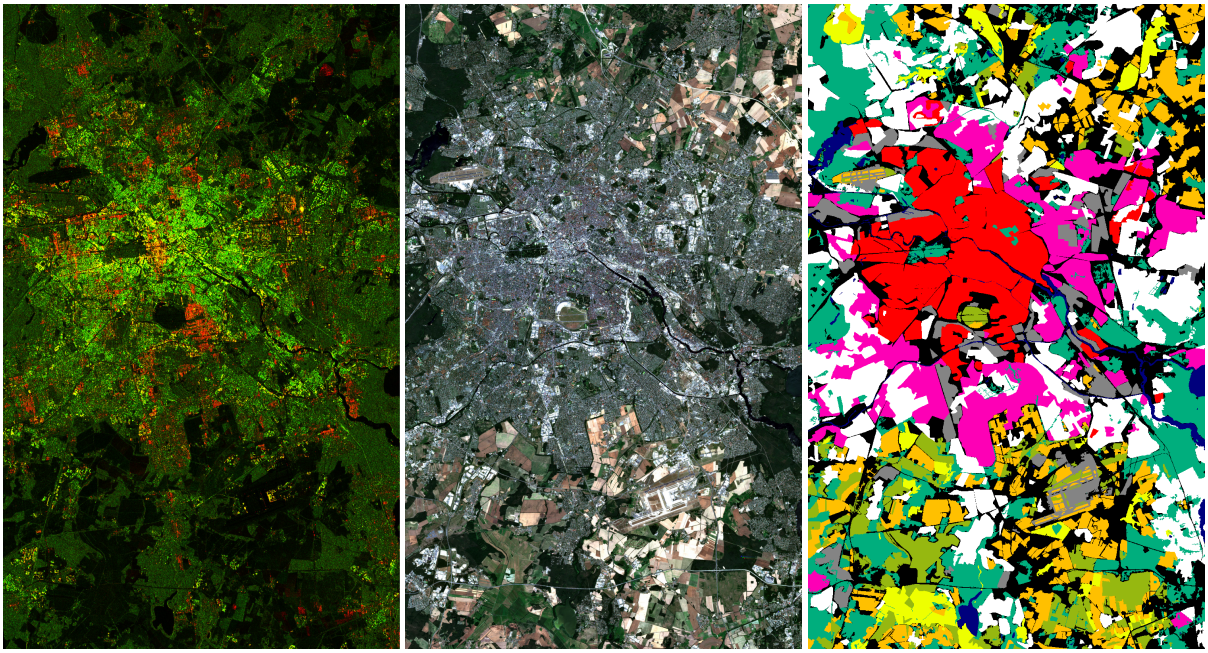
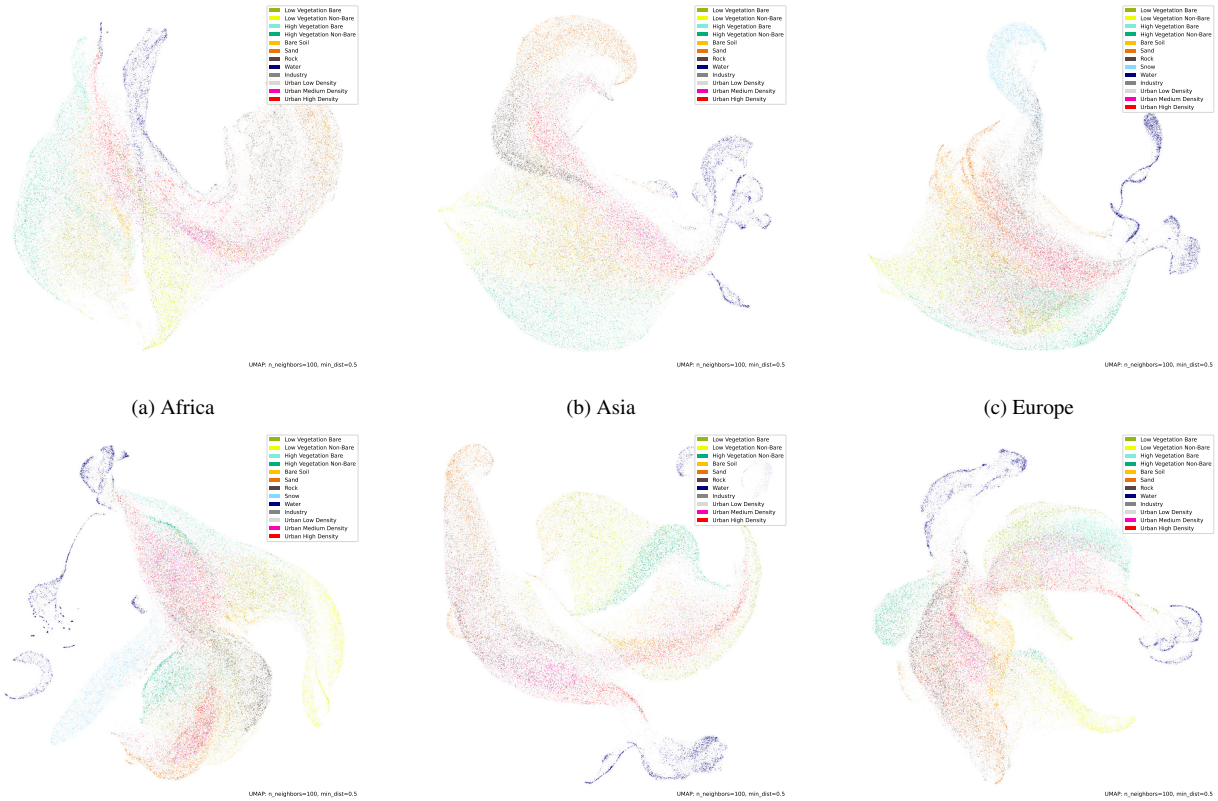


Figure 10. Europe - Berlin

9. UMAP Visualizations

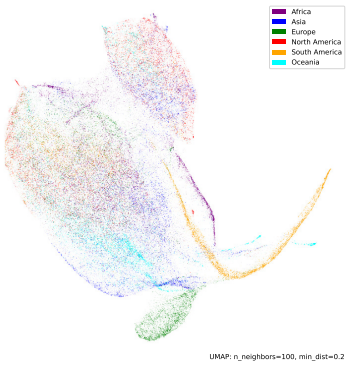
The UMAP projections in Figures 11–12 summarize the underlying structure of the dataset by embedding pixel-level feature vectors into a two-dimensional space. The continent-wise visualizations highlight how regional characteristics and land-cover distributions shape the arrangement of classes, while the class-wise plots reveal the degree of separation or overlap between surface types across different geographic contexts. Together, these embeddings provide an intuitive view of class variability, inter-class similarity, and continental heterogeneity present in the dataset.



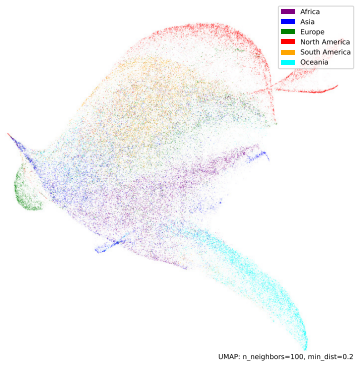
(a) Africa (b) Asia (c) Europe

(d) North America (e) South America (f) Oceania

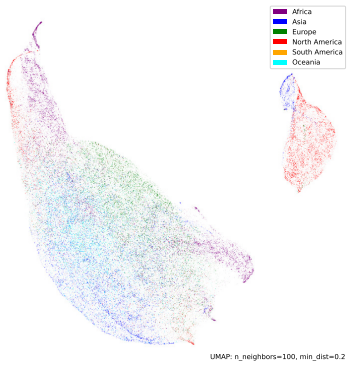
Figure 11. UMAP representations of pixel values sampled for each continent of the dataset. Each dot represents a pixel vector after dimensionality reduction of UMAP. The colors depict the class that the pixel belongs to.



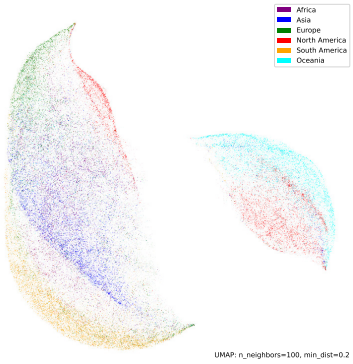
(a) Low Vegetation Bare



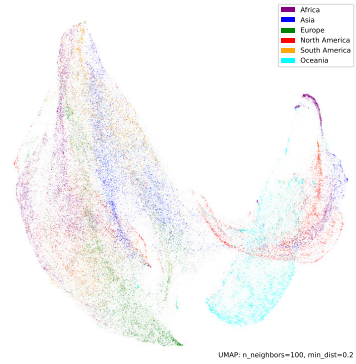
(b) Low Vegetation Non-Bare



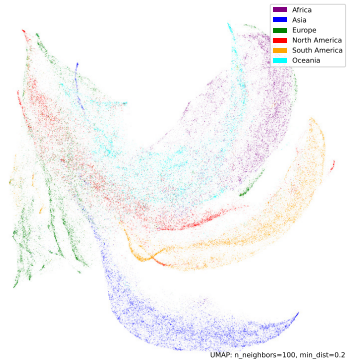
(c) High Vegetation Bare



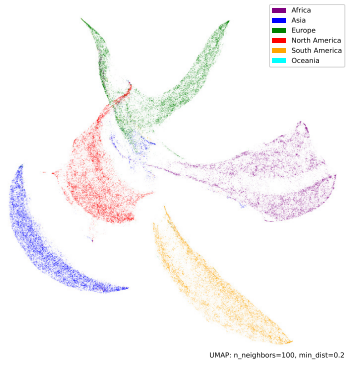
(d) High Vegetation Non-Bare



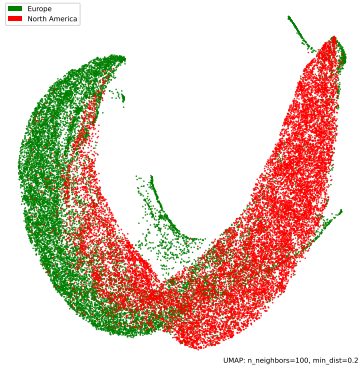
(e) Bare Soil



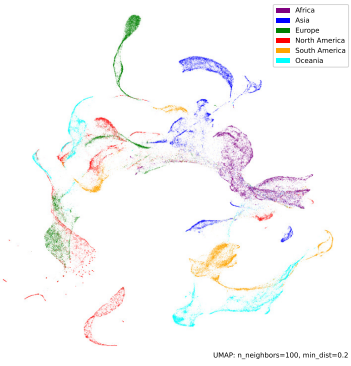
(f) Sand



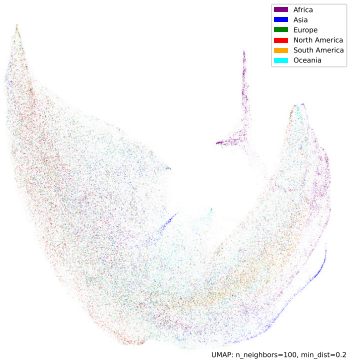
(g) Rock



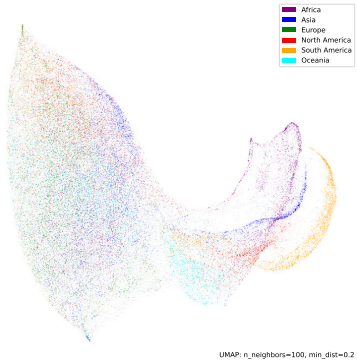
(h) Snow



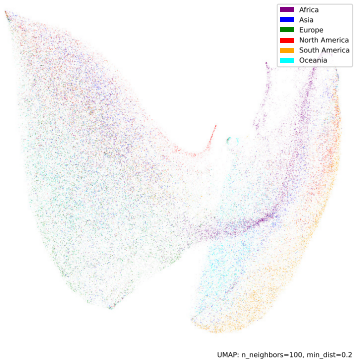
(i) Water



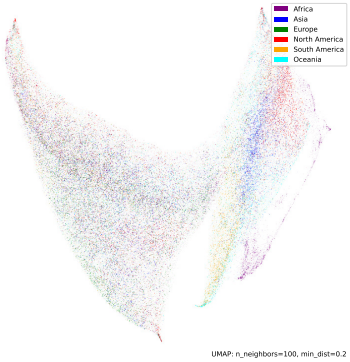
(j) Industry



(k) Urban Low Density



(l) Urban Medium Density



(m) Urban High Density

Figure 12. UMAP representations of pixel values of selected classes of the dataset. Each dot represents a pixel vector after dimensionality reduction of UMAP. The colors depict the continent that the pixel originated from.

10. Details on baseline models and training

Tables 8 and 9 summarize the model configurations used in our experiments.

Table 8 reports the number of trainable parameters for each architecture under different sensor inputs. As expected, fusing S1 and S2 increases model capacity only slightly. The SSL model introduces substantially more parameters due to its larger backbone.

Table 9 outlines the training settings for each architecture, including learning rates, batch sizes, patch sizes, and maximum epochs. U-Net is trained for more epochs with early stopping, while SegFormer and the SSL model rely on shorter, fixed-length training schedules with smaller batch sizes due to their higher computational demands.

Table 8. Trainable parameters by model and input modality.

	# Trainable Parameters		
Sensors	U-Net	SegFormer	SSL
S1	31,043,725	84,601,882	NA
S2	31,048,333	84,633,242	NA
S1+S2	31,049,485	84,639,514	138,708,182

Table 9. Training hyperparameters for each model architecture.

Hyperparameter	U-Net	SegFormer	SSL
Max. Epochs	100	30	30
Learning Rate	1×10^{-4}	6×10^{-5}	3×10^{-6}
Batch Size	16	8	4
Patch Size	256×256	224×224	224×224
Early Stopping	10 Epochs on val. loss	None	None