remote sensing

MDPI

Article

SENSERO: A Multi-Scale, Multi-Label Sentinel-2 Benchmark for Romania with Cross-Scale Land-Cover Classification Baselines

Teodor Costăchioiu [1], Vlad Vasilescu [2], Alexandru-Cosmin Grivei [1], Andreea Griparis [1] and Daniela Faur [1,*]

[1] GeoSense—Geospatial and Smart Sensors for Environmental Applications, Campus Research Institute, National University of Science and Technology POLITEHNICA Bucharest, 060042 Bucharest, Romania; teodor.costachioiu@upb.ro (T.C.); alexandru.grivei@upb.ro (A.-C.G.);andreea.griparis@upb.ro (A.G.)

[2] SIGMALab—Signal Modelling and Analysis Lab, Campus Research Institute, National University of Science and Technology POLITEHNICA Bucharest, 060042 Bucharest, Romania; vlad.vasilescu2111@upb.ro

* Correspondence: daniela.faur@upb.ro

Highlights

What are the main findings?

- We introduce SENSERO, a multi-scale, multi-label Sentinel-2 dataset for Romania: 10,000 fixed locations at eight patch sizes (64–336 pixels), sharing identical centres and splits across scales, with multi-label annotations derived from CORINE Land Cover 2018.
- ResNet-50 results show that larger patches offer little improvement at CLC Level 1, but deliver clear gains at Levels 2 and 3, with performance levelling off at around 224–256 pixels. These gains remain consistent across leave-scene-out validation, different labelling rules, and alternative backbones, and are strongest where broad CLC classes cover diverse types of terrain.

What are the implications of the main findings?

- Patch size and nomenclature should be reported as explicit design variables; results obtained under different settings are not directly comparable.
- SENSERO offers a shared, controlled basis for studying these effects and closes a coverage gap for Romania, absent from benchmarks such as BigEarthNet.

Abstract

Patch size is an important design parameter in patch-based Earth observation (EO) learning, yet it is rarely treated as an experimental variable because existing benchmarks use a single fixed scale. We introduce SENSERO, a multi-scale, multi-label Sentinel-2 benchmark for Romania consisting of 10,000 geographically fixed locations extracted at eight patch sizes, from 64 × 64 to 336 × 336 pixels, at a constant 10 m ground sampling distance, so that all scales share identical patch centres and an identical, spatially disjoint 70/15/15 partition. Each patch provides twelve-band bottom-of-atmosphere reflectance, a rasterised CORINE Land Cover (CLC) 2018 reference map from which five deterministic multi-label annotation layers are derived (three levels of the CLC hierarchy, the BigEarthNet nomenclature, and a scheme adapted to the Romanian landscape). The benchmark task is patch-level multi-label classification; the rasterised reference maps serve to derive and trace the labels, not as a segmentation target. Using ResNet-50 baselines trained from scratch, we find that larger patches bring little benefit at CLC Level 1 but clear gains at Levels 2 and 3 (macro mAP at Level 3 rising from 0.42 at 64 px to 0.61 at 336 px), with per-location gains saturating around 224–256 pixels; that transfer across scales is asymmetric, models trained on wide patches transferring to narrow ones far better than the reverse; and that the choice of nomenclature affects measured performance as much as patch size. Leave-scene-out cross-validation, an

Check for updates

Academic Editor: Xuecao Li

Received: 3 September 2026
Revised: 15 October 2026
Accepted: 22 October 2026
Published: 24 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

ablation of the labelling rule, pretrained ResNet-50 and ViT-S/16 controls and bootstrap confidence intervals show that the scale effect is not attributable to spatial leakage, to the annotation rule, or to the baseline architecture. Comparison with an independent 10 m land-cover product shows that the gain from context is concentrated where CLC labels generalise over heterogeneous ground.

Keywords: sentinel-2; CORINE land cover 2018; multi-label classification; multi-scale learning; patch size; benchmark dataset; land cover; transfer learning; earth observation; Romania

1. Introduction

Artificial intelligence for Earth observation (EO) [1,2] has enabled large-scale mapping of land cover, vegetation, and environmental dynamics. The growing availability of open EO data, particularly from the Copernicus Sentinel and Landsat missions, together with increased computational capacity, has driven substantial progress in automated environmental analysis [3,4]. Nearly all of this progress rests on a common abstraction: the satellite scene is decomposed into square patches, and the patch becomes the unit of learning, of labelling, and of evaluation.

At a fixed ground sampling distance (GSD), patch size determines the spatial extent observed by the model and influences the number of land-cover classes likely to co-occur within a sample. This choice therefore affects both the information available for classification and the complexity of the learning task, which may range from effectively single-label to genuinely multi-label. For Sentinel-2 at 10 m GSD, a 64 × 64 patch covers 40.96 ha, whereas a 336 × 336 patch covers 1128.96 ha, a factor of 27.6 in area. Against the 25 ha minimum mapping unit (MMU) of CORINE Land Cover (CLC), the former is barely larger than a single mapping unit, while the latter can accommodate more than forty. Consequently, the semantics of a patch's label depends on scale: at small extents the reference label behaves like a scene-level category, while at large extents it becomes a compositional summary of a heterogeneous landscape.

Most existing benchmarks use a predefined patch size, which limits systematic evaluation of its effect on model performance. Moreover, differences in geographic coverage, acquisition periods, nomenclatures, class distributions, and data partitioning protocols make it difficult to isolate the effect of patch size when comparing results across datasets. Table 1 compares the principal international Sentinel-2 and land-cover benchmarks. EuroSAT [5] uses 64 × 64 single-label patches; BigEarthNet [6,7] and its refined successor reBEN [8] established CLC-derived multi-label annotation on 120 × 120 patches as the standard protocol for Sentinel-2 scene understanding; SEN12MS [9] and So2Sat LCZ42 [10] added Sentinel-1 radar at 256 and 32 pixels respectively; Sentinel2GlobalLULC [11] and LandCoverNet [12] use 224 and 256 pixels for global land-cover mapping; fMoW [13] covers 62 functional categories at sub-metre resolution; and the recent pre-training corpora SSL4EO-S12 [14], SatlasPretrain [15] and MMEarth [16] each provide global coverage at a single fixed patch size. Classical scene-classification datasets such as UC Merced [17], AID [18] and NWPU-RESISC45 [19] likewise rely on a single categorical label at one scale. Comparing across these datasets confounds patch size with geography, acquisition period, nomenclature, class balance, and partitioning protocol, so the effect of scale as such cannot be isolated. What is missing is a benchmark in which the spatial extent varies while everything else is held fixed: locations, imagery, reference layer, nomenclature, and data partition.

Table 1. SENSERO in relation to international Sentinel-2 and land-cover benchmarks. Patch size refers to the 10 m bands where applicable.

Dataset	Sensor(s)	Patch (px)	Samples	Labels	Coverage
EuroSAT [5]	S-2	64	27,000	10 classes, single-label	Europe (34 countries)
BigEarthNet-S2 [6]	S-2 (S-1 in v1.1)	120	590,326	43/19 CLC classes, multi-label	10 European countries
reBEN [8]	S-1, S-2	120	549,488	19 CLC classes, multi-label; CLC map	10 European countries
SEN12MS [9]	S-1, S-2, MODIS	256	180,662	MODIS IGBP/LCCS maps	Global
So2Sat LCZ42 [10]	S-1, S-2	32	400,673	17 local climate zones	42 cities worldwide
Sentinel2GlobalLULC [11]	S-2	224	194,877	29 classes, single-label	Global
LandCoverNet [12]	S-2	256	1980 chips	7 classes, segmentation	Africa (v1), global
SSL4EO-S12 [14]	S-1, S-2	264	251,079 locations	unlabelled (pre-training)	Global
MMEarth [16]	12 modalities	128	1.24 M	pixel maps, unlabelled	Global
SENSERO (this work)	S-2	**64–336 (8 scales)**	10,000 locations × 8	5 CLC-derived multi-label layers; CLC map	Romania (national)

Counts are as reported by the respective publications and may differ between versions. SENSERO is the only entry in which the same locations are provided at several patch sizes under one partition.

The choice of patch size in this literature spans 5 × 5 to 256 × 256 pixels. At very fine scales, Fazzini et al. [20] discriminated grassland habitats from 5 × 5 Sentinel-2 patches that preserve spectral homogeneity within parcels; at intermediate scales, So2Sat's 32 × 32 patches balance local morphology against computational cost; Malczewska et al. [21] use 128 × 128. Collectively these choices indicate that no single spatial scale is universally optimal, and that in many applications the appropriate scale is not known in advance: smaller patches promote single-class consistency and spectral purity, larger patches capture the broader spatial variability required for multi-label and contextual learning. What the literature does not provide is a controlled measurement of this trade-off. Existing multi-scale work in remote sensing addresses the question from the model side, through multi-scale feature aggregation inside a network or through scale-aware pre-training objectives [22]; this is complementary to, but distinct from, varying the extent of the input sample itself. A related line of work connects EO imagery to natural language, from RSICD [23] and RSITMD [24] to RS5M [25]; captions in these datasets are short, human- or model-written and weakly aligned with authoritative taxonomies, whereas the label sets used here prioritise verifiable alignment with CLC over linguistic variety (Section 2.7). Our own companion study examined LLM-based captioning of the same patch locations across scales and CLC levels [26], and found captioning quality to improve with patch size, most at the finest CLC level and with signs of levelling off at the largest patches; whether supervised classification exhibits the same behaviour was not examined.

Existing benchmarks have another limitation: they barely cover Romania. BigEarthNet was constructed from tiles over ten countries (Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, and Switzerland); Romania is not among them, and the nearest covered tiles in neighbouring Serbia do not extend across the western border. EuroSAT samples patches across Europe, but only a handful fall within Romanian territory. Regional Romanian datasets exist, such as DACIA5 [27] for crop identification over the Braşov region (Sentinel-1/2, 2020–2024, 17 crop classes) and land-cover mapping studies of the Subcarpathians and Bucegi Mountains [28], but they are limited in extent or

not released as annotated data. Figure 1 compares the footprints of BigEarthNet, EuroSAT and DACIA5 with Romania's territory. Two consequences follow. The practical one is that no public multi-label benchmark allows model performance over this territory to be measured at national scale. The methodological one is that, because the territory is unseen by the dominant source benchmarks, it constitutes a suitable target domain for evaluating cross-dataset transfer, a direction we discuss in Section 5 and treat experimentally in a companion study.



Figure 1. Coverage of existing datasets over Romania: BigEarthNet (blue, stops at the border), EuroSAT (red, few patches), and DACIA5 (orange, Braşov region only). Basemap: Natural Earth (public domain).

We address both gaps with SENSERO, a multi-scale, multi-label Sentinel-2 benchmark for the territory of Romania. SENSERO comprises 10,000 geographically fixed patch centres, each extracted at eight spatial scales (64, 112, 120, 128, 224, 256, 280, and 336 pixels) at a constant 10 m GSD. Because the centres are shared across scales, and because a single spatially disjoint 70/15/15 partition is propagated by patch identifier to every scale, patch size is isolated as the only varying factor. Each patch is annotated with five deterministically generated multi-label sets derived from CLC 2018 [29]: three follow the CLC hierarchy (Levels 1–3), one follows the BigEarthNet nomenclature [7], and one follows a scheme regrouped for the Romanian landscape. Thirty-seven of the 44 CLC Level-3 classes occur on Romanian territory; SENSERO covers 34 of them at the canonical scale and 35 at the largest, accounting for 99.999% of the nationally mapped area (Section 5). The additional class at 336 pixels is sclerophyllous vegetation (323), which covers 129.6 ha nationally and is captured only once the footprint is large enough to contain it above the 2% threshold. The benchmark task defined on SENSERO is patch-level multi-label classification: a model receives a patch and predicts the set of land-cover classes present in it. The rasterised CLC maps distributed with every patch are the provenance of those labels, and are released so that every label can be traced to the reference raster and so that users can derive alternative targets; the baselines reported here do not perform pixel-wise segmentation.

The dataset is accompanied by a set of reproducible baseline experiments designed to address three research questions concerning the effects of spatial scale on land-cover

classification. **RQ1: How does multi-label land-cover classification performance vary with patch size, and does the answer depend on the semantic granularity of the target nomenclature?** Larger patches supply more spatial context but also a more heterogeneous and therefore harder label set. We expect the balance between these two effects to differ between coarse (CLC Level 1, 5 classes) and fine (CLC Level 3, 34 classes) targets, and we quantify it directly by training and evaluating independently at each of the eight scales for each of three nomenclature levels. **RQ2: Do representations learned at one spatial scale transfer to another?** Because all scales share identical centres, splits and reference labels, a model trained at scale i can be evaluated at scale j without any confounding change in the underlying data. The resulting 8×8 cross-scale matrix separates two effects that are otherwise entangled: sensitivity of the learned features to the extent of the field of view, and sensitivity to the change in label composition that accompanies it. **RQ3: How much of the difficulty observed at fine semantic granularity is attributable to the nomenclature rather than to the model?** CLC classes are defined on land-use and land-management grounds as much as on physical appearance, so some Level-3 distinctions are not observable in 10 m reflectance at all. Because SENSERO annotates the same imagery and the same partition under five vocabularies of differing thematic and visual coherence (Section 2.7), the ambiguity associated with the target nomenclature can be separated from classifier limitations. Because the answer to RQ1 could in principle be produced by three potential confounding factors in the experimental design: (i) proximity between training and evaluation patches that grows with the footprint, (ii) the rule by which labels are assigned, and (iii) the particular baseline architecture, the experiments also include controls for each of them.

In summary, this paper makes four contributions. First, SENSERO is, to our knowledge, the first EO benchmark in which patch size is an explicit experimental axis: eight scales are extracted around identical patch centres, under a single spatially disjoint partition, with reference labels derived from the same CLC 2018 layer at every scale. Second, it is the first national-scale multi-label land-cover benchmark for Romania. Third, each patch carries five annotation layers generated by a deterministic rule that makes every label traceable to the reference raster. Fourth, we provide a reproducible baseline suite that addresses RQ1–RQ3, together with the robustness controls described above. The dataset is released on Zenodo under CC BY 4.0, and the complete data-preparation pipeline and benchmark code are available on GitHub (see Data Availability Statement).

The remainder of the paper is organised as follows. Section 2 describes the construction and validation of SENSERO. Section 3 defines the benchmark protocol, the three main experiments, and the robustness controls. Section 4 reports the results, Section 5 discusses their implications, their transferability beyond Romania and the limitations of the dataset, and Section 6 concludes.

2. The SENSERO Dataset

2.1. Design Overview

SENSERO combines Sentinel-2 Level-2A imagery with hierarchical land-cover annotations derived from CLC 2018, providing a multi-scale representation of Romania's land-cover patterns. It contains 10,000 patch locations extracted from 38 Sentinel-2 scenes covering the national territory, each rendered at eight spatial scales. Because individual patches may contain several land-cover classes, the dataset is inherently multi-label at all scales. The unit of annotation and of evaluation is the patch: each patch carries the set of CLC classes it contains, and the rasterised CLC map is distributed alongside the imagery as the source of that set rather than as a per-pixel target. Identical geographic patch centres are used at every scale, which is the property that enables the cross-scale

comparisons of Section 3. Figure 2 shows the spatial distribution of the patch locations and their split assignment.

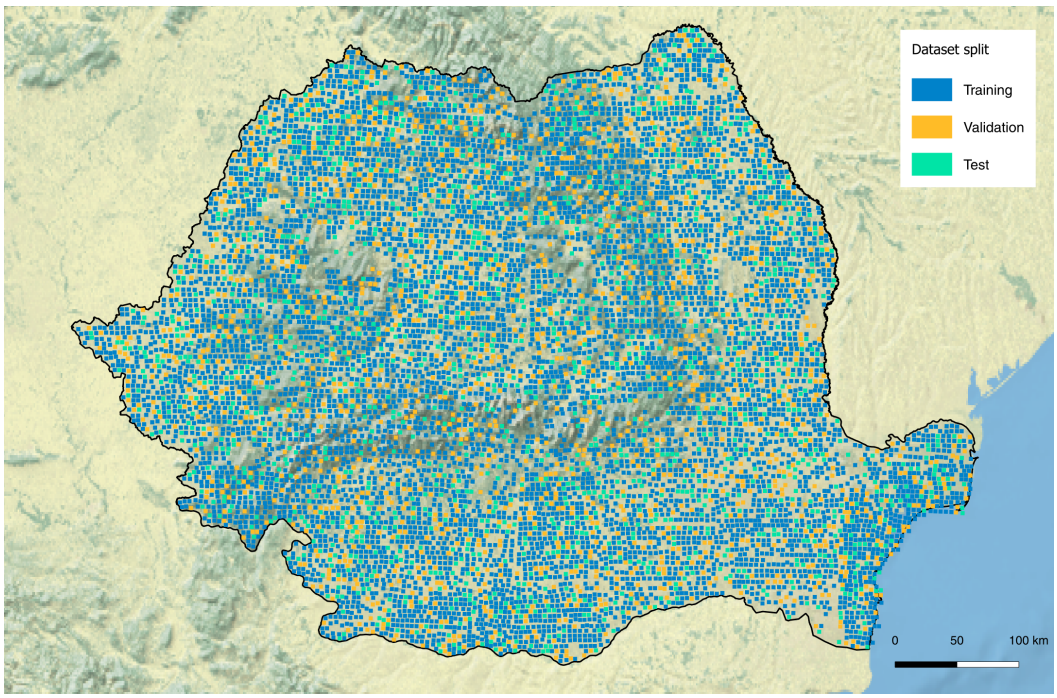


Figure 2. Spatial distribution of the 10,000 SENSERO patch locations across Romania, coloured by dataset split (training, validation, test). The locations were selected to capture the country's diverse agro-ecological and landscape conditions. Basemap: Natural Earth (public domain).

Figure 3 illustrates a single location extracted at all eight scales together with the corresponding label sets, making the effect of spatial extent on semantic composition directly visible.

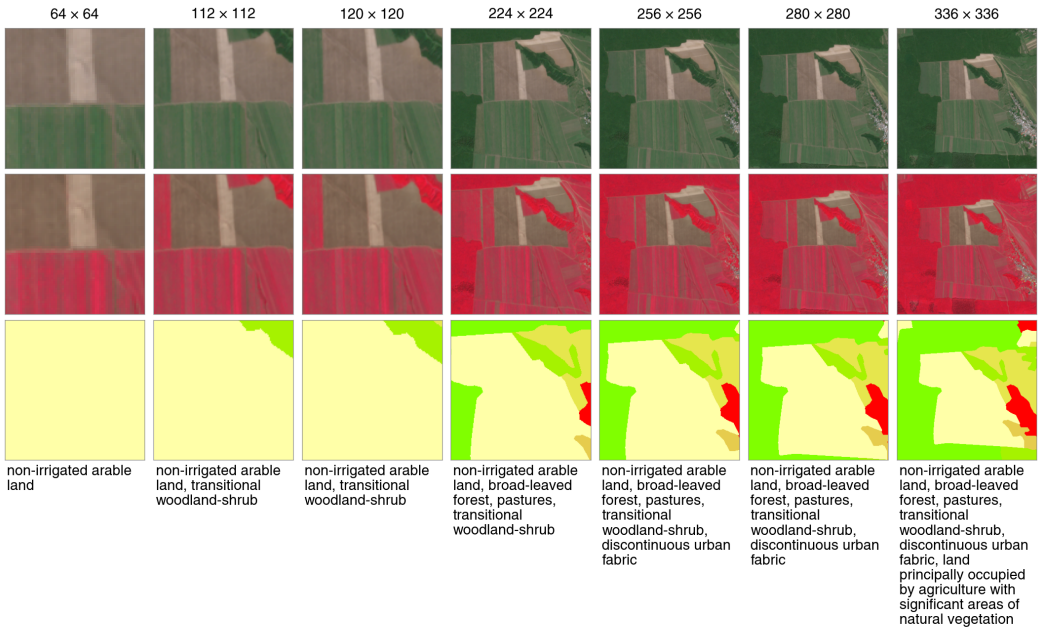


Figure 3. Multi-scale patch example from SENSERO showing a single location at all eight spatial extents (64 × 64 to 336 × 336 pixels). Top: true-colour composite (B4–B3–B2); middle: false-colour NIR composite (B8–B4–B3); bottom: CLC reference map with Level-3 class labels.

2.2. Data Sources

The dataset was derived from 38 Sentinel-2 Level-2A scenes [30] covering Romania, all acquired between 9 August and 14 October 2018 and obtained from the Copernicus Data Space Ecosystem (https://dataspace.copernicus.eu, accessed on 27 July 2026) as Collection-1 reprocessed products (processing baseline 05.00).

This acquisition window was selected on operational grounds: it corresponds to the period of typically lowest cloud and haze contamination over Romania (particularly important over mountainous terrain, where persistent cloud is common) while remaining early enough to precede the onset of frost and snow at higher elevations. The interval was kept deliberately narrow (approximately nine weeks) so that vegetation phenology remains broadly comparable across scenes. Level-2A products were used because they provide atmospherically corrected, bottom-of-atmosphere (BOA) surface reflectance, which improves radiometric comparability across scenes.

The SENSERO acquisitions correspond to the reference year of CLC 2018 [31], the most recent edition of the CORINE Land Cover inventory at the time the dataset was assembled and the standard harmonised pan-European land-cover reference. CLC is a generalised inventory with a 25 ha MMU, focused on broad, temporally stable classes, and is used throughout SENSERO as the authoritative semantic reference layer.

2.3. Sentinel-2 Preprocessing

All scenes were processed with ESA SNAP v12.0.1 following a reproducible protocol. From each Level-2A product we retained the twelve optical bands (B1–B8, B8A, B9, B11, B12). Bands originally acquired at 20 m and 60 m were resampled to 10 m by bicubic interpolation using Band 2 (Blue) as the geometric reference. Bicubic interpolation yields a smooth, continuous reflectance surface well suited to convolutional models such as the baselines evaluated here, whose spatial filters operate on local reflectance gradients. The cirrus band B10 is not part of the Level-2A product. The non-reflectance ancillary layers were excluded from the released stacks: the Scene Classification Layer (SCL), the aerosol optical thickness and water-vapour maps, the quality and cloud-probability masks, and the true-colour preview. The SCL was nevertheless used at an earlier stage of the workflow to guide patch placement.

Each scene was exported as a 12-band, 16-bit GeoTIFF (LZW compression, predictor = 2). Pixel values encode BOA reflectance with a quantification value of 10,000 and a per-band additive offset of -1000, applied to all products since Processing Baseline 04.00 to prevent negative reflectance from being clipped over dark surfaces [32]. Surface reflectance must therefore be recovered as $\rho = (\mathrm{DN} - 1000)/10{,}000$ and not as DN/10,000; omitting the offset introduces a systematic $+0.1$ reflectance bias across all bands, which would corrupt any downstream spectral index or classification model. Care is required when comparing these data against products generated under earlier baselines, in which the offset is absent. Bands B1 and B9, natively 60 m, are retained for completeness but carry coastal-aerosol and water-vapour information rather than surface reflectance; most surface-reflectance analyses will exclude them, and the baselines in Section 3 do so.

Two RGB composites were generated per scene: a true-colour composite (B4–B3–B2) and a false-colour near-infrared composite (B8–B4–B3). Each channel was independently histogram-stretched and gamma-corrected with fixed, scene-invariant parameters so that composites remain directly comparable across scenes (true colour: minimum = 0.01, maximum = 0.50, gamma = 0.45; false colour: same minimum, maximum = 0.60, gamma = 0.55). Both were exported as 8-bit GeoTIFFs with full georeferencing. Because the stretch and gamma correction are not invertible, these composites are intended for

visual inspection and lightweight workflows; quantitative work should use the 16-bit reflectance bands.

2.4. Land-Cover Reference Maps

For each Sentinel-2 scene, two land-cover reference maps were generated from CLC 2018. The CLC vector data were subset to the Romanian national boundaries and extended with a 20 km buffer to ensure complete coverage of border regions, preserving the native EPSG:3035 coordinate reference system. Rasterisation and reprojection were performed in QGIS 3.40 using a custom Python script based on GDAL 3.8: for each scene, the buffered subset was clipped to the scene extent, rasterised at 10 m, and reprojected to the coordinate reference system of the corresponding Sentinel-2 tile (EPSG:32634 or EPSG:32635), ensuring pixel-level alignment with the imagery. Two raster products were produced per scene: a 16-bit single-band map encoding CLC Level-3 codes, and an 8-bit visualisation rendered with the official Copernicus palette. Both were exported as LZW-compressed GeoTIFFs.

2.5. Patch Selection and Multi-Scale Extraction

Patch centres were defined in QGIS using a custom script that displayed the maximum patch footprint of 3360 × 3360 m (336 × 336 pixels). True-colour composites and CLC maps served as visual and semantic references, and centres were selected manually to achieve uniform national coverage and balanced representation of CLC classes while excluding clouds, smoke, aircraft trails and other artefacts.

The resulting sample is an expert stratified sample optimised for land-cover class coverage, and not an areally representative sample of Romania. Rare CLC Level-3 classes such as *rice fields* (213), *salt marshes* (421) and *coastal lagoons* (521), which together occupy a negligible fraction of the national territory, were sampled deliberately so that every class occurring in the country is present in all three dataset partitions at a workable count. The objective was guaranteed class coverage, not areal over-representation: measured as a share of the label mass, these classes end up near or below their areal share (Section 5), but at counts that make training and evaluation possible. *Rice fields*, which cover 0.24% of the territory, are represented by 41 patches; *coastal lagoons* by 56 and *salt marshes* by 10, where a proportional draw would have yielded very few samples, or no samples. This necessarily reduces the relative share of the classes that dominate Romania by area, principally *non-irrigated arable land* (211) and *broad-leaved forest* (311). The trade-off was accepted because the primary design objective was to support model training and evaluation across the full CLC nomenclature, which a strictly proportional sample of 10,000 patches could not achieve: several classes would have been represented by a handful of patches, or by none at all. Users who require a distribution matching the actual national land-cover composition can reweight the samples using the per-class patch counts recorded in the metadata tables. The consequences of this design for the interpretation of the benchmark metrics are addressed in Appendix D.

A total of 10,000 centres were defined and exported as vector data, then converted to Apache Parquet tables, one per Sentinel-2 scene. Patch extraction was automated in Python with GDAL 3.8.4, transforming geographic coordinates into image-relative pixel locations and extracting square patches of 64, 112, 120, 128, 224, 256, 280 and 336 pixels. The set of scales was chosen to include the sizes used by the principal existing benchmarks—64 (EuroSAT), 120 (BigEarthNet), 224 (Sentinel2GlobalLULC and the standard input size of ImageNet-pretrained backbones), 256 (LandCoverNet)—so that results obtained on SENSERO are directly comparable to those obtained elsewhere, together with intermediate sizes that fill the range.

For each centre and scale, five raster products were generated by subsetting the scene-level products: a 16-bit multispectral cube with all 12 optical bands; a true-colour composite; a false-colour composite; a 16-bit single-band CLC Level-3 reference map; and an 8-bit CLC visualisation. All were exported as GeoTIFFs with LZW compression and complete georeferencing; the true-colour and CLC colour products are additionally provided as 8-bit PNG.

2.6. National Land-Cover Reference

To characterise the sample against the territory it was drawn from, CORINE Land Cover 2018 was clipped to the Romanian national boundary and the area of each Level-3 class integrated. The boundary was obtained as a single polygon in Stereo 70 (EPSG:3844, Pulkovo 1942(58)), the Romanian national projection, and reprojected onto the CLC coordinate system rather than the reverse: EPSG:3035 (ETRS89-LAEA Europe) is equal-area, so polygon areas can be read from it directly, whereas Stereo 70 is conformal. Along this boundary, the local areal scale factor of Stereo 70 ranges from -0.050% near the projection origin to $+0.130\%$ on the coast, a spread of 0.180 percentage points; the corresponding figures are 0.520 pp for UTM zone 35N and 1.132 pp for zone 34N, versus 0.000 pp for EPSG:3035. The national total is barely affected in Stereo 70 ($+0.004\%$), because its scale factor was chosen so that areal errors cancel across Romania, but per-class areas acquire a geographically structured bias that EPSG:3035 avoids at no cost. The change of datum from Pulkovo 1942(58) to ETRS89 was resolved by a 3 m Helmert transformation, which over a national perimeter of roughly 2500 km perturbs the area by about 7.5 km^2, or 0.003%.

The clipped reference covers 24,353,325 ha (i.e., 243,533 km^2), distributed over 37 Level-3 classes. This exceeds Romania's official land area of 238,397 km^2 because the national polygon includes territorial waters: subtracting CLC class 523 (*sea and ocean*, 5126 km^2) *leaves* 238,407 km^2 of *land* and *coastal lagoons*, within 0.004% of the official figure. Extending the boundary a further 20 km seaward, matching the buffer applied when subsetting CLC for patch extraction adds only 1055 km^2 ($+0.43\%$), almost all of it class 523, and rescales every terrestrial share by the same constant factor of 1.025; the choice of seaward extent therefore does not reorder the comparison.

Of the 37 classes occurring nationally, SENSERO covers 34. The three absent are *sclerophyllous vegetation* (323, 129.6 ha nationally), *burnt areas* (334, 48.8 ha) and *agro-forestry areas* (244, 25.1 ha), together 0.001% of the mapped territory. The dataset therefore represents essentially the entire national land-cover composition, and the classes it omits are ones for which the available mapped area was insufficient to obtain representative samples.

2.7. Multi-Label Annotation Layers

For each patch, five label sets were generated from the corresponding CLC 2018 reference patch, providing semantic information at multiple thematic levels (Table A2). Generation relied exclusively on deterministic, rule-based logic; no machine learning or language model was involved at any point.

A centre-weighted strategy was used, designed to approximate the descriptive behaviour of a human observer by prioritising salient features near the patch centre. The central region, representing approximately 20% of the patch area, was analysed to identify dominant CLC Level-3 classes, with anthropogenic classes prioritised when present. Classes were then ordered by decreasing areal proportion, first within the central region and then across the whole patch. Classes occupying less than 2% of the patch area were excluded to suppress noise from marginal features. The centre-weighted step determines only the *order* of the tokens; membership in the label set is determined by the area threshold alone. This is verified directly in Section 4.5: regenerating the Level-3 sets from the

reference rasters with the threshold and nothing else reproduces the published sets on all 10,000 patches at every scale. Since the classifiers of Section 3 are trained on binary indicator vectors, token order never reaches the training target. All label sets are lowercase, contain no internal punctuation, and present land-cover terms as comma-separated tokens, so that they serve both as human-readable descriptions and as structured targets.

The first three layers correspond to CLC Levels 1, 2 and 3, offering progressively finer semantic detail. The fourth follows the BigEarthNet nomenclature [7], which permits direct comparison with models trained on that benchmark. The fifth uses a custom scheme developed by our working group specifically for the Romanian context, introduced for two reasons. First, the BigEarthNet nomenclature was derived from a set of countries that does not include Romania: several of its classes have negligible occurrence on Romanian territory, while a number of CLC Level-3 classes that do occur in Romania fall outside it entirely and are mapped to *unlabeled* (Table A2), so relying on the fourth layer alone would discard part of the semantic content carried by the reference data. Second, the CLC Level-3 nomenclature is defined on thematic rather than visual grounds and does not always group together classes that are spatially and spectrally similar at 10 m resolution in the Romanian landscape. The fifth layer regroups Level-3 classes according to how they present themselves in Sentinel-2 imagery over Romania: the three *mine*, *dump* and *construction site* classes (131, 132, 133) are merged into a single *quarries or dump sites* category, and the *scrub and herbaceous vegetation* classes present in Romania (321, 322 and 324) into *shrubland*. The scheme is offered as a pragmatic, visually coherent target in addition to the standard CLC and BigEarthNet layers.

One property of the generation procedure should be stated because it affects how the two coarser vocabularies relate to the CLC layers. When a Level-3 class present in a patch has no counterpart in the BigEarthNet nomenclature or in the custom scheme, the corresponding slot is not left empty but filled with the next class by areal share, which may be a class below the two per cent threshold and therefore absent from the Level-3 layer. Layers four and five are consequently not reconstructible from the Level-3 labels alone; they carry a small amount of additional information. On the canonical 128-pixel scale this affects 19 of the 10,000 patches (0.19%). In the two cases where no replacement class remained, the explicit token *unknown land cover* is emitted instead.

The five annotation layers, referred to as captions in the distributed metadata for backward compatibility with the released files, consist of multi-label sets represented as text. They were designed to support vision–language models using a controlled vocabulary derived from the CLC nomenclature. Token order reflects the relative salience of land-cover classes, and each label can be traced to the corresponding reference raster. This deterministic representation ensures consistent annotations across the dataset and supports reproducible evaluation. The label sets can be converted into natural-language descriptions using templates or paraphrasing. However, recovering the original labels from generated descriptions cannot be guaranteed. Figure 4 illustrates the five annotation layers for selected patches.

2.8. Train–Validation–Test Partition

The 10,000 locations were partitioned into training, validation and test subsets (7000/1500/1500; ratio 70/15/15) through a four-stage procedure that jointly balances class representation, rare-class coverage, and spatial independence between the training and evaluation pools. All stratification was performed on the multi-label target vectors defined at CLC Level 3. Because every other label representation derives from Level 3, balancing the most granular level is the most demanding constraint and propagates balanced repre-

sentation to all derived targets. The canonical partition was constructed at the 128 × 128 pixel scale, an intermediate choice among the available sizes.

RGB	CLC Map	caption1	caption2	caption3	caption4	caption5
		agricultural areas, forest and semi-natural zones	arable land, heterogeneous agricultural areas, forests, pastures, scrub and/or herbaceous vegetation associations	non-irrigated arable land, complex cultivation patterns, broad-leaved forest, pastures, transitional woodland-shrub	arable land, complex cultivation patterns, broad-leaved forest, pastures, transitional woodland-shrub	agriculture, mixed farmland, forests, pastures, shrubland
		forest and semi-natural zones, agricultural areas	forests, arable land, scrub and/or herbaceous vegetation associations, heterogeneous agricultural areas, pastures	broad-leaved forest, non-irrigated arable land, transitional woodland-shrub, complex cultivation patterns, pastures	broad-leaved forest, arable land, transitional woodland-shrub, complex cultivation patterns, pastures	forests, agriculture, shrubland, mixed farmland, pastures
		artificial surfaces, forest and semi-natural zones, agricultural areas	mine dump and construction sites, forests, arable land, pastures, industrial commercial and transport units	dump sites, broad-leaved forest, non-irrigated arable land, pastures, industrial or commercial units	broad-leaved forest, arable land, pastures, industrial or commercial units	quarries or dump sites, forests, agriculture, pastures, urban fabric

Figure 4. Sample patches at 256 × 256 pixel scale illustrating all five annotation layers in SENSERO, progressing from CLC Level 1 through Level 3 class labels, followed by the BigEarthNet-style nomenclature and the custom SENSERO scheme.

Stage 1: Tile-level stratified sampling. Romania is covered by multiple Sentinel-2 tiles, each spanning approximately 100 × 100 km and identified by an MGRS grid code. Rather than pooling all patches into a single global partition, iterative multi-label stratified sampling [33] was applied independently within each tile, targeting the 70/15/15 ratio. This ensures that every tile contributes proportionally to all three subsets, so that the evaluation sets are geographically distributed across the country rather than concentrated in a few regions.

Stage 2: Rare-class enforcement. After concatenating the per-tile partitions, CLC classes appearing in only one to three patches received deterministic placement to guarantee full class coverage: single-patch classes were assigned to training; two-patch classes were divided between training and test; three-patch classes were distributed one per subset. Each forced placement was accompanied by a compensatory swap, preferentially exchanging patches dominated by common agricultural or forest categories (CLC codes 200–399), to preserve the target subset sizes.

Stage 3: Spatial separation. Spatial autocorrelation, the tendency of geographically proximate observations to be more similar than distant ones [34], inflates evaluation metrics when neighbouring patches appear in both the training and evaluation sets [8]. Patch centres were therefore constrained so that no two footprints overlap at any of the eight scales, which at the largest scale requires a separation of at least one patch width, 3360 m, along at least one coordinate axis. Direct geometric testing of the released footprints confirms this holds exactly: across all 10,000 patches, no pair shares any ground area at 336 pixels, the closest separation between an evaluation patch and its nearest training patch being 3376 m, or 1.005 patch widths (3392 m for test patches alone). The partition is thus non-overlapping but, at the largest scale, adjacent by construction; the consequences are quantified in Section 4.5. Six pairs are edge-adjacent, sharing a boundary to within a millimetre of numerical precision but no area.

A further semantic proximity constraint was applied in addition to geometric non-overlap. Validation or test patches with a training neighbour within 5000 m and an identical CLC Level-3 label set were resolved through label-set-matched swaps, preserving subset sizes and class distributions. In total, 214 swaps were performed. Forty-six same-label

pairs involving 40 evaluation patches remain below the threshold (minimum 3528 m), corresponding to rare label combinations for which no compliant replacement was available. The complete audit is provided in the code repository. The implications of residual spatial proximity are discussed in Section 5.3.

Stage 4: Multi-scale propagation. The canonical partition was propagated by patch identifier to all remaining scales, ensuring identical subset membership everywhere. Because all scales share the same centres, no additional spatial checks were required. This step is what makes the cross-scale experiment of Section 3.5 valid: a test patch at scale j is the same ground location as the corresponding test patch at scale i, and was never seen in training at any scale.

2.9. Data Records and Distribution

The dataset [35] is distributed as 16 compressed archives organised by data format and scale, following the naming convention `[format]_patch_[size].zip`, where `format` is `geotiff` or `png` and `size` is one of the eight patch dimensions. Each archive is self-contained, bundling the image patches with the corresponding metadata table, so that downloading a single archive is sufficient to work at a given format and scale. Within each archive, data are organised under `SENSERO/[format]/Patch_[size]`, grouped by Sentinel-2 scene identifier following the Copernicus naming convention for unambiguous traceability to the source products. The GeoTIFF archives provide full geospatial and spectral information; the PNG archives contain only composite products and are intended for lightweight visualisation and benchmarking workflows.

For each scale, a metadata file `metadata_[size].parquet` records, per patch, the Sentinel-2 scene identifier (`BaseFolder`), the patch base filename (`BaseFilename`), the list of CLC Level-3 codes present (`CLC_codes`), the five annotation layers (`caption1`–`caption5`), and the split assignment. The schema is identical across scales.

Two companion files are distributed alongside the archives. `sensero_patch_footprints.gpkg` is a GeoPackage vector index containing the footprints of all 10,000 patches in a single layer, in WGS-84 (EPSG:4326), with `filename` as the join key to the `BaseFilename` column of the per-scale metadata tables. GeoPackage was chosen over Shapefile because the latter's 254-character field-width limit would silently truncate the longer label sets. `sensero_technical_guide.pdf` documents the folder hierarchy, the file-naming convention, the metadata schemas, and the full CLC-to-label mapping tables.

2.10. Quality Control and Technical Validation

The dataset underwent a multi-stage quality-assurance process covering image integrity, semantic annotation accuracy, and consistency of the partition across scales.

Image inspection. Each of the 10,000 true-colour patches was visually inspected twice, in separate review cycles. In the first round, patches exhibiting clouds, haze, compression artefacts, missing data or geometric misalignment were flagged and removed; new centres were selected in QGIS, metadata were updated, and replacement patches were extracted through the same pipeline. A second inspection confirmed that all revised patches met the required radiometric and geometric standards.

Footprint independence. A geometric overlap analysis verified the spatial independence of all patch footprints. During initial extraction, a small number of samples at the largest scale (3.36 $\times$ 3.36 km) exhibited minor partial overlaps resulting from dense sampling to maximise national coverage; these were identified and corrected through the same re-extraction procedure. The published dataset contains no overlapping footprints at any of the eight scales. This was re-verified directly on the released footprint index: an exact pairwise test over all 10,000 patches, carried out in the UTM zone in which each patch was

cut, returns no pair sharing ground area at 336 pixels and therefore none at any smaller scale. Six pairs are edge-adjacent, agreeing with a shared boundary to within 6 µm, four orders of magnitude below the 10 m pixel.

Reference-map integrity. All CLC reference patches were checked programmatically for no-data values and incomplete class coverage. Missing data occurred primarily in border regions adjacent to Ukraine and the Republic of Moldova, where the spatial extent of CLC 2018 does not always fully cover the Sentinel-2 scene footprint. Affected patches were excluded, alternative centres were selected within fully covered areas, and both the imagery and the reference map were re-extracted. Geometric and semantic alignment between imagery and rasterised reference maps was further confirmed by QGIS overlay checks on a random sample of 1000 patches across scenes and scales.

Annotation verification. To assess the accuracy and robustness of the five annotation layers, 1000 randomly selected patches (10% of the dataset) were examined. For each sample, all five layers were compared against the reference Level-3 raster, verifying that the centre-weighted assignment consistently identified the dominant land-cover types, that Level-3 classes were correctly mapped to their Level-1, Level-2 and BigEarthNet counterparts, that the custom groupings were semantically coherent, that token ordering progressed logically from central to peripheral classes, and that formatting conventions were respected. Because the labels are deterministically generated from the reference maps, they introduce no model-dependent variability. For the same 1000 samples, label sets were additionally examined at every scale to confirm stability of the dominant central classes, consistency of the ordering logic, and absence of contradictory categories arising from the broader spatial context available at larger extents.

Partition validation. Class distributions were compared across all scales against the canonical 128-pixel partition. The 70/15/15 ratio was consistently maintained, with Level-3 class proportions differing by only 2–3% across scales, and no rare class was absent from any subset at any scale.

3. Benchmark Design

This section defines the evaluation protocol and the three experiments that address RQ1–RQ3. All experiments use the canonical partition of Section 2.8: training on the 7000 training patches, model selection on the 1,500 validation patches, and reporting exclusively on the 1500 test patches, which are never used for tuning.

3.1. Task Definition and Metrics

The benchmark task is a patch-level, multi-label, land-cover classification: given a patch, the goal is to predict the set of all classes present within it. The baseline models output one score vector per patch (global average pooling followed by a linear head) and are never used for creating per-pixel classification maps. The rasterised reference maps discussed in Section 2.7 enter only through the label sets derived from them. Targets are taken from the annotation layers of Section 2.7, converted to binary indicator vectors over the class vocabulary of the corresponding nomenclature. We report results for CLC Level 1 (5 classes), Level 2 (15 classes) and Level 3 (34 classes; 35 at the 336-pixel scale), which together span the range of semantic granularity required by RQ1.

Following the protocol established by BigEarthNet [6,7], models produce independent per-class sigmoid scores and are trained with binary cross-entropy. We report mean average precision (mAP), computed both micro- and macro-averaged, and F_1 score at a fixed decision threshold of 0.5, computed sample-, micro- and macro-averaged. Macro-averaged metrics are given primary weight throughout, for reasons discussed in Section 5: the SENSERO sampling design deliberately over-represents rare classes, so micro-averages are

dominated by the frequent classes and understate performance differences on the tail of the distribution.

Every configuration is trained with three random seeds (0, 1, 2), and all figures are reported as means with standard deviations over seeds. Several of the observed effects are of the order of one to two mAP points and are therefore never interpreted on the strength of a single run.

Comparisons between scales are performed using *paired* tests. Since all scales share identical patch centres and data partitions, each configuration is evaluated on the same geographic locations, allowing performance differences to be assessed for individual locations. Pairing accounts for the substantial variability in per-sample average precision across locations and improves sensitivity to consistent performance differences between adjacent scales. Treating scales as independent samples would ignore this correspondence and reduce the statistical power of the comparison. We therefore report the Wilcoxon signed-rank test [36] on per-location average precision, applying Holm correction [37] across the family of adjacent-scale comparisons. Because the expected AP of a random ranking increases with the number of positive labels, and hence with patch size (Section 4.1), per-location AP was corrected for this scale-dependent chance level before paired testing. Specifically, the expected AP of a random ranking with the same label count was subtracted from each per-location AP, and all paired tests were performed on the resulting skill-above-chance values.

3.2. Input Configuration

The benchmark uses the ten surface-reflectance bands acquired natively at 10 m and 20 m, excluding the two 60 m atmospheric bands B1 and B9 for the reasons given in Section 2. This band configuration follows BigEarthNet-S2 and subsequent Sentinel-2 benchmarks, facilitating comparison with previous work. Reflectance is recovered from the stored digital numbers as $\rho = (\mathrm{DN} - 1000)/10{,}000$.

Training augmentation is limited to horizontal and vertical flips and rotations by multiples of 90°, which preserve the spatial extent of the input patches. Random-resized cropping is excluded because it alters the effective patch scale. Per-band standardisation statistics are computed on the training partition at the canonical 128-pixel scale and reused across all scales to ensure consistent normalisation.

3.3. Architectures and Training Configuration

All results reported in this paper use a ResNet-50 [38] with the input convolution adapted to ten channels and an independent sigmoid output per class. Models are trained from random initialization. Global average pooling accommodates the varying input sizes, and inputs are not resized. Training uses AdamW with a learning rate of 10^{-3} and weight decay of 10^{-4}, cosine annealing with five warm-up epochs, at most 60 epochs, batch size 64, automatic mixed precision, and gradient-norm clipping at 1.0. The checkpoint maximising validation macro mAP is retained, with early stopping at a patience of 15 epochs; because the cosine schedule anneals the learning rate to zero within the epoch budget, runs that reach the cap are converged rather than truncated. Inference is deterministic, with cuDNN autotuning disabled during evaluation.

3.4. Experiment 1: Per-Scale Baselines

For each of the eight patch sizes and each of the three nomenclature levels, a model is trained and evaluated independently, giving 8×3 configurations per architecture. This isolates the effect of spatial extent on classification performance under otherwise identical conditions and directly addresses RQ1. Because the label sets at larger extents contain more classes on average, the comparison is not one of task difficulty held constant; the

accompanying dataset statistics of Section 4.1 quantify this shift and are required for the interpretation of the results.

3.5. Experiment 2: Cross-Scale Generalisation

Each model trained at scale i in Experiment 1 is evaluated, without any further training, on the test partition at every scale j, producing an 8×8 matrix M per nomenclature level and architecture. The diagonal reproduces Experiment 1; the off-diagonal entries measure transfer across spatial extent. The primary summary statistic is the column-fixed *transfer gap* $\Delta(i \to j) = M(j,j) - M(i,j)$: at a fixed evaluation scale j, it measures the cost of having trained at extent i rather than j against numbers computed on the same test set, so prevalence, class support and intrinsic difficulty cancel exactly. Row-based statistics—the mean degradation along a row of the matrix, or a triangle asymmetry computed from raw entries—compare one model across test sets whose chance floors differ by up to 0.069 macro mAP, more than most of the effects being measured, and are therefore not used as evidence; direction asymmetry is instead computed on the transfer gap itself, averaged separately over the cells above and below the diagonal. This experiment is only possible because scales share centres, labels and partition, and it addresses RQ2.

3.6. Experiment 3: Nomenclature Comparison

At the canonical 128-pixel scale, we compare classification performance across the five annotation vocabularies described in Section 2.7: CLC Levels 1–3, the BigEarthNet nomenclature, and the custom Romania-adapted scheme (Caption 5 of Table A2; the two are the same scheme). Results for the three CLC levels are taken directly from the corresponding 128-pixel configurations of Experiment 1 (Section 3.4), while additional models are trained for the BigEarthNet and Romania-adapted vocabularies using the same architecture, training schedule, partition, and seeds. An independent retraining of the three CLC configurations within this experiment reproduced them with 0.012 macro mAP (0.895, 0.637 and 0.472 at Levels 1–3), the run-to-run variability of a training run at fixed seed under cuDNN autotuning and mixed precision. This provides a controlled comparison of the effect of the target vocabulary on classification performance and addresses RQ3.

Two comparisons are of particular interest. First, CLC Level 3 is compared with the Romania-adapted scheme, which regroups two sets of Level-3 classes: *the mine*, *dump*, and *construction-site* classes (131–133), and *the scrub* and *herbaceous* vegetation classes present in Romania (321, 322, 324), while leaving the remaining vocabulary unchanged. This comparison examines whether reducing ambiguity among these classes improves their discriminability. Second, CLC Level 3 is compared with the BigEarthNet nomenclature, which provides a coarser representation in which several classes occurring in Romania are mapped to unlabeled. This comparison illustrates the effect of adopting the pan-European vocabulary on the semantic detail represented over Romanian territory.

Per-class F1 scores are additionally reported for the affected class groups, since aggregate metrics alone may not distinguish improved class discriminability from changes resulting from vocabulary reduction. Whether the effect of nomenclature varies with spatial extent remains an open question (Section 5.5).

3.7. Robustness Controls for the Scale Effect

The scale effect measured by Experiment 1 could, in principle, arise from three properties of the design rather than from spatial context. Each is addressed by a dedicated control at CLC Level 3, the level at which the effect is strongest. Unless stated otherwise the controls use the key scales 64, 128, 224 and 336 pixels and the training configuration of Experiment 1.

Spatial leakage. Because the partition is shared across scales, the clearance between a test patch and its nearest training patch shrinks, relative to the footprint, as the footprint grows (Section 2.8). Two analyses quantify whether this proximity contributes to the gains at large scales. The first is *distance-stratified evaluation*: every test patch is assigned the distance from its centre to the nearest training-patch centre, the test set is divided into quartiles of that distance, and the Experiment 1 models are scored inside each quartile with the paired per-location protocol. If the gain at large scales were driven by proximity, it would concentrate in the nearest quartile. The second is *leave-scene-out cross-validation*: the 38 Sentinel-2 scenes are assigned to five folds balanced by patch count (7–8 scenes per fold); for each fold, a model is trained on the scenes of three folds, selected on those of a fourth and tested on those of the fifth, so that no model ever sees the scene from which a test patch was cut. This is the strongest spatial separation the release admits, and its cost is class coverage: rare classes are geographically concentrated, so some folds lack them in their test scenes. We report the macro average over the classes present in every fold's test set, and additionally pool the five out-of-fold prediction sets, so that every one of the 10,000 patches is scored once by a model that never saw its scene, and evaluate the pooled set over the full 34-class vocabulary. This control is run at 128 and 336 pixels with one seed.

Labelling rule. The area threshold of 2% (Section 2.7) admits a growing absolute area as the footprint grows: 0.8 ha at 64 pixels, 22.6 ha at 336. To separate the effect of the rule from that of scale, the Level-3 label sets were regenerated from the reference rasters at thresholds of 0%, 1%, 2% and 5% and a model was trained at each threshold and key scale (one seed). Regeneration at 2% reproduces the published sets exactly and serves as a check of faithfulness; the descriptive effect of the threshold on label density is reported separately from the trained effect on performance. The grid also contains an *iso-area* diagonal: 5% at 128 pixels, 2% at 224 and 1% at 336, all corresponding to an absolute threshold of 8–11 ha, along which the scale effect can be read at a nearly constant absolute rule.

Architecture. Experiment 1 uses one architecture trained from scratch. Two controls test whether its conclusions are specific to that choice: a ResNet-50 initialised from ImageNet weights [39], with the input convolution tiled from three to ten channels, and a ViT-S/16 [40] initialised from ImageNet weights, with positional embeddings interpolated to the input size (a 4×4 token grid at 64 pixels). Both are trained end-to-end with the Experiment 1 schedule and three seeds. The question they answer is whether the ordering of scales and the location of the plateau depend on the backbone, not whether they outperform it.

Uncertainty. Macro-averaged metrics over 34 classes are sensitive to the classes with few test positives. Alongside the seed standard deviation, which measures training variance, we report a percentile bootstrap over test patches [41] (1000 resamples; classes without positives in a resample are masked from that resample's average), which measures test-set sampling variance; the test-set support of every class at every scale; and a secondary macro average restricted to the 17 classes with at least ten test positives at every scale.

3.8. Agreement with an Independent Land-Cover Product

All labels in SENSERO derive from CLC. To characterise the benchmark against a reference that does not, the CLC map of every test patch was compared pixel by pixel with ESA WorldCover 2020 [42], a 10 m global product produced independently of CLC from Sentinel-1 and Sentinel-2 data. The two nomenclatures cannot be matched class for class: CLC is a 25 ha, partly use-based inventory (urban fabric, complex cultivation, transitional woodland) while WorldCover maps physical cover per pixel. Both were therefore reduced to a common nine-group legend (artificial, cropland, grassland, shrubland, tree, bare, snow, wetland, water; crosswalk in Table A5), WorldCover was resampled onto each patch's 10 m grid by nearest neighbour, and the fraction of jointly mapped pixels on which the two agree

was recorded per patch. Disagreement is not, in general, a CLC error; it measures how literally the CLC label describes the ground inside the patch. The Experiment 1 models were then evaluated inside four strata of this agreement, again with the paired per-location protocol, to ask whether the classifier tracks the pixel-level cover or the CLC generalisation, and how that depends on the spatial extent it observes. A patch classifier cannot be scored per pixel, so this is the form in which the question of independent validation can be posed for the present baselines.

4. Results

4.1. Dataset Statistics

Table 2 summarises how the composition of the benchmark changes with patch size. As spatial extent increases from 64 to 336 pixels, the mean number of CLC Level-3 labels per patch rises from 2.07 to 4.44, while the proportion of single-label test patches decreases from 33.5% to 7.0%. The task therefore becomes progressively more multi-label at larger spatial extents.

Mean per-class prevalence on the test partition also increases with patch size, from 0.062 at 64 pixels to 0.130 at 336 pixels. Because this quantity corresponds to the macro-mAP expected from a random scorer, it defines a scale-dependent chance floor for the comparisons reported in Sections 4.2 and 4.3. Raw between-scale differences in macro-mAP must therefore be interpreted relative to this changing baseline.

Table 2. Label composition of SENSERO as a function of patch size.

Patch Size (px)	Extent (ha)	CLC MMU	Mean Labels/Patch (L3)	Single-Label Patches (%)	L3 Classes Present	Mean Class Prevalence
64 × 64	40.96	1.6	2.07	33.5	34	0.062
112 × 112	125.44	5.0	2.64	21.1	34	0.078
120 × 120	144.00	5.8	2.73	19.9	34	0.08
128 × 128	163.84	6.6	2.81	18.7	34	0.083
224 × 224	501.76	20.1	3.66	11.9	34	0.107
256 × 256	655.36	26.2	3.90	10.1	34	0.114
280 × 280	784.00	31.4	4.07	8.9	34	0.119
336 × 336	1128.96	45.2	4.44	7.0	35	0.130

Extent is computed at 10 m GSD. The MMU column expresses the patch extent as a multiple of the 25 ha CORINE minimum mapping unit. Mean labels per patch is computed over all 10,000 patches; single-label share and mean per-class prevalence are computed on the 1500-patch test partition.

Figure 5 shows the corresponding shift in the distribution of CLC Level-3 labels per patch, confirming the transition from predominantly one- to two-label patches at the smallest extent to a clearly multi-label regime at the largest scales.

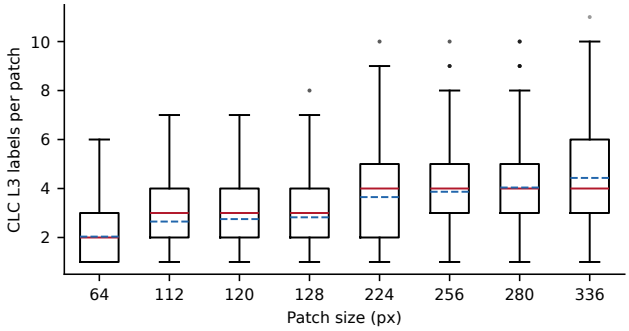


Figure 5. Distribution of the number of CLC Level-3 labels per patch on the test partition, as a function of patch size. Boxes give the interquartile range and the median (solid); the dashed line marks the mean.

4.2. Per-Scale Baselines

Table 3 and Figures 6 and 7 report the results of the first experiment, in which models were trained and evaluated independently at each of the eight patch sizes for CLC Levels 1–3. The following observations emerge from the results: (i) the benefit of increasing spatial extent depends on semantic granularity; (ii) the benefit is not monotonic and tends to saturate at intermediate-to-large patch sizes; and (iii) the effect of spatial context varies substantially across land-cover classes.

First, the benefit of increasing spatial extent depends on semantic granularity (Table 3 and Figure 6a). Macro mAP increases with patch size at all three CLC levels, but part of this increase reflects the scale-dependent chance floor identified in Section 4.1. After accounting for this effect, performance at Level 1 remains essentially stable across scales, whereas clear gains remain at Levels 2 and 3, with the strongest improvement observed at Level 3. Thus, the scale-associated performance gain is concentrated at the finer semantic levels.

Second, the benefit of increasing spatial extent is not monotonic (Figure 6b). The paired per-location analysis shows that the largest significant gains occur between 64 and 112 pixels ($p < 10^{-11}$) and between 128 and 224 pixels ($p < 10^{-12}$), after Holm correction. Other adjacent-scale changes are substantially smaller, and per-location performance reaches a plateau around 224–256 pixels. This indicates that increasing the spatial extent beyond this range provides limited additional benefit at the individual-location level.

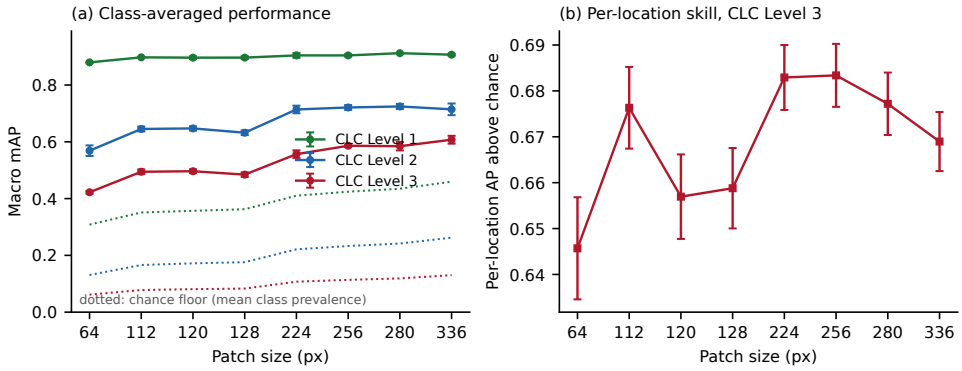


Figure 6. Experiment 1. (**a**) Macro mAP as a function of patch size for each CLC level. Solid lines with markers give the mean over three seeds, one line per level as indicated in the legend; error bars give the standard deviation over the three seeds. Dotted curves of the matching colour give the corresponding chance floor, defined as the macro mAP of a random scorer on the test set. (**b**) Per-location average precision above the empirical chance floor at CLC Level 3; markers give the mean over the 1500 test locations and error bars give the 95% confidence interval.

Third, the effect of spatial context varies across land-cover classes (Figure 7). The largest AP gains between 64 and 336 pixels are observed for land principally occupied by agriculture with significant natural vegetation, complex cultivation patterns, and transitional woodland–shrub: mosaicked classes whose CLC definition summarises a mixture of covers, so that a wider field of view shows the model the pattern that the label names rather than one of its components. Well-supported compact anthropogenic classes also gain (industrial or commercial units, AP 0.41 $\rightarrow$ 0.60; mineral extraction sites, 0.18 $\rightarrow$ 0.47), whereas water bodies is unchanged within uncertainty (a paired bootstrap of the 336- versus 128-pixel difference gives -0.03, 95% interval $[-0.14, +0.07]$). Per-class changes for the ten classes with fewer than ten test positives at every scale (Table A3) are not statistically resolvable in either direction—for instance, the 95% bootstrap interval for dump sites at 128 pixels spans $[0.00, 1.00]$—and are not interpreted here. The aggregate scale effect is therefore broad rather than uniform, concentrated in classes whose label denotes

a landscape pattern; Section 4.6 shows that it is concentrated, more generally, in patches whose CLC label generalises over heterogeneous ground.

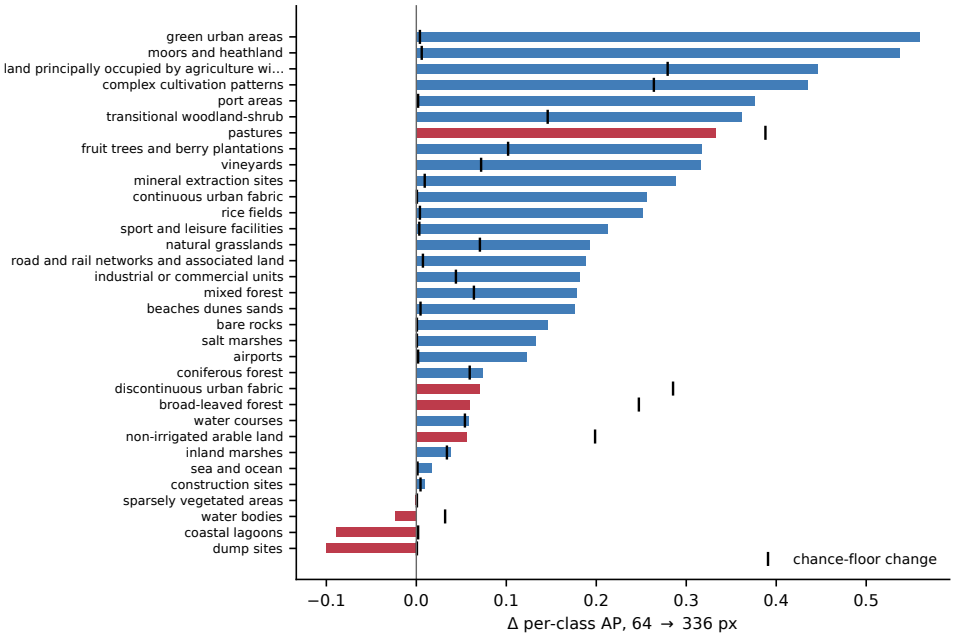


Figure 7. Change in per-class average precision at CLC Level 3 between the 64- and 336-pixel scales (bars), compared with the corresponding change in each class's chance floor, defined by its test-set prevalence (black ticks). Red bars indicate classes for which the increase in raw AP is smaller than the increase in the chance floor.

Table 3. Per-scale multi-label classification performance on the SENSERO test partition.

	CLC Level 1		CLC Level 2		CLC Level 3	
Patch Size	**mAP**	**F_1 (Macro)**	**mAP**	**F_1 (Macro)**	**mAP**	**F_1 (Macro)**
64	0.880 ± 0.002	0.773 ± 0.016	0.569 ± 0.019	0.522 ± 0.023	0.422 ± 0.005	0.385 ± 0.043
112	0.898 ± 0.002	0.775 ± 0.016	0.645 ± 0.009	0.589 ± 0.005	0.494 ± 0.008	0.428 ± 0.020
120	0.896 ± 0.005	0.801 ± 0.024	0.647 ± 0.007	0.574 ± 0.011	0.496 ± 0.007	0.412 ± 0.025
128	0.897 ± 0.005	0.782 ± 0.003	0.632 ± 0.008	0.590 ± 0.010	0.484 ± 0.008	0.406 ± 0.023
224	0.904 ± 0.009	0.826 ± 0.017	0.714 ± 0.013	0.636 ± 0.028	0.557 ± 0.013	0.497 ± 0.033
256	0.904 ± 0.004	0.813 ± 0.028	0.721 ± 0.008	0.645 ± 0.001	0.586 ± 0.004	0.522 ± 0.009
280	0.912 ± 0.001	0.833 ± 0.008	0.725 ± 0.009	0.642 ± 0.018	0.585 ± 0.015	0.503 ± 0.016
336	0.907 ± 0.004	0.820 ± 0.010	0.714 ± 0.021	0.606 ± 0.045	0.607 ± 0.014	0.513 ± 0.022

Values are reported as mean ± standard deviation over three seeds. mAP denotes macro-averaged average precision. Bootstrap confidence intervals on the Level-3 column are given in Table 4.

Table 4 places the two sources of uncertainty in the Level-3 macro mAP side by side. The bootstrap interval over test patches has a half-width of about 0.05 at every scale, four to ten times the seed standard deviation, and is skewed upwards: resamples that happen to omit one of the classes with one to three test positives lose a low AP term from the average. Restricting the macro average to the 17 classes with at least ten positives at every scale halves the interval, makes it symmetric, and leaves the ordering of scales unchanged, with the steps 128 $\rightarrow$ 224 and 224 $\rightarrow$ 336 pixels separated by non-overlapping intervals even without pairing. Absolute levels should therefore be read with the ± 0.05 of the full vocabulary in mind, while differences between scales rest on the paired per-location tests, which are far more powerful than a comparison of marginal intervals.

Table 4. Uncertainty of the CLC Level-3 macro mAP of Experiment 1: seed standard deviation and 95% bootstrap interval over test patches, over the full 34-class vocabulary and over the 17 classes with at least ten test positives at every scale.

	34 Classes			17 Well-Supported Classes		
Patch Size	**mAP**	**Seed SD**	**Bootstrap 95% CI**	**mAP**	**Seed SD**	**Bootstrap 95% CI**
64	0.422	0.005	[0.391, 0.509]	0.530	0.004	[0.515, 0.557]
112	0.494	0.008	[0.465, 0.570]	0.613	0.008	[0.597, 0.639]
120	0.496	0.007	[0.468, 0.575]	0.614	0.012	[0.599, 0.638]
128	0.484	0.008	[0.454, 0.562]	0.614	0.007	[0.597, 0.637]
224	0.557	0.013	[0.531, 0.624]	0.686	0.004	[0.670, 0.705]
256	0.586	0.004	[0.552, 0.653]	0.707	0.003	[0.691, 0.726]
280	0.585	0.015	[0.552, 0.650]	0.714	0.002	[0.698, 0.733]
336	0.607	0.014	[0.574, 0.665]	0.730	0.004	[0.714, 0.747]

Bootstrap: 1000 resamples of the 1500 test patches, interval limits averaged over the three seeds. At 64 pixels one class has no test positive and the full-vocabulary average runs over 33 classes. Test support per class and scale is given in Table A3.

4.3. Cross-Scale Generalisation

Figure 8 reports the cross-scale generalisation results at CLC Level 3, while the corresponding numerical values are provided in Table A1. The experiments yield three main findings: (i) cross-scale transfer generally incurs a performance cost; (ii) this cost is asymmetric with respect to scale; and (iii) transfer is substantially better between nearby scales than across the small-to-large scale range.

First, cross-scale transfer generally incurs a performance cost. Across the 56 off-diagonal train-test combinations, the mean transfer gap is 0.035 macro mAP, indicating that models evaluated at a different spatial extent generally perform below the corresponding model trained directly at that evaluation scale. The diagonal therefore remains the appropriate reference for assessing cross-scale transfer.

Second, the transfer cost is asymmetric with respect to scale, Figure 8b. Training at a smaller extent and evaluating at a larger one results in a mean transfer gap of 0.059 macro mAP, compared with 0.021 in the opposite direction. Models trained at larger spatial extents therefore transfer more effectively to smaller inputs than models trained at smaller extents transfer to larger ones. The asymmetry has a simple reading. A model trained on wide patches has learnt to recognise several classes that co-occur in one scene and to score them jointly; a zoomed-in crop presents it with a familiar problem that has fewer answers. A model trained on narrow patches has learnt to recognise one or two dominant classes and has never had to represent a dense co-occurrence; a wide landscape asks it a question it has not seen, and its scores for the additional classes are poorly calibrated.

Third, transfer is substantially better between nearby scales than across the small-to-large scale range. Models trained within the 64–128-pixel range transfer relatively well across these scales, and the same pattern is observed within the 224–336-pixel range, whereas transfer between the two ranges incurs larger gaps. Among the evaluated training scales, 256 pixels shows the smallest mean outgoing transfer gap. A few off-diagonal configurations marginally exceed their corresponding diagonal values, but these differences remain within the observed seed-level variability and do not provide evidence that cross-scale training outperforms matched-scale training.

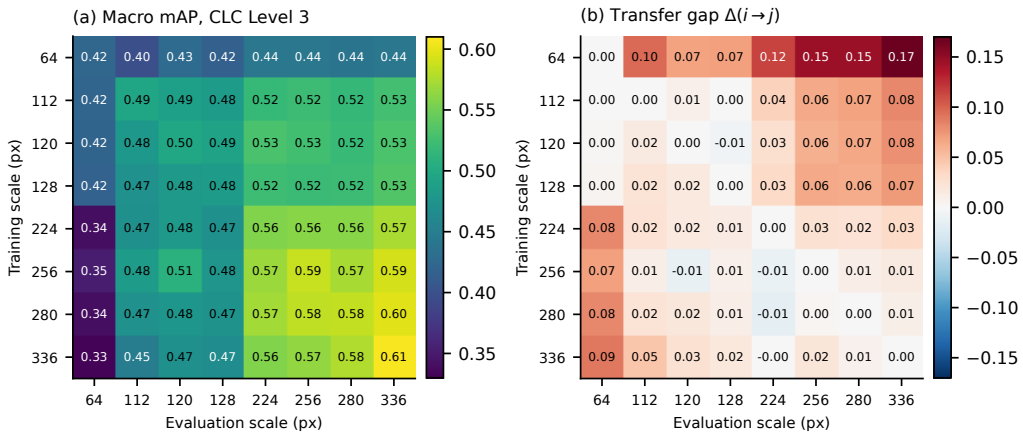


Figure 8. Experiment 2 at CLC Level 3. (**a**) Cross-scale macro mAP, with training scale by row and evaluation scale by column; values are means over three seeds. (**b**) Column-fixed transfer gap $\Delta(i \to j) = M(j, j) - M(i, j)$; positive values indicate the performance cost of training at scale i rather than at the match evaluation scale j.

4.4. Nomenclature Comparison

Table 5 compares classification performance across the five annotation vocabularies at the fixed 128-pixel scale. Aggregate performance varies substantially across vocabularies, with macro mAP ranging from 0.484 for CLC Level 3 to 0.672 for the custom SENSERO scheme. These values are not directly comparable in absolute terms, however, because the vocabularies differ in both class cardinality and class prevalence. Their corresponding chance floors range from 0.083 for CLC Level 3 to 0.363 for CLC Level 1. The results therefore show that the choice of target vocabulary has a substantial effect on the measured classification performance, but aggregate scores alone do not indicate whether this effect results from improved class discriminability or from differences in vocabulary composition.

Table 5. Classification performance across the five annotation vocabularies at the fixed 128-pixel scale. Values are means ± standard deviation over three seeds.

Annotation Layer	Classes	mAP	F_1 (Micro)	F_1 (Macro)
CLC Level 1	5	0.897 ± 0.005	0.893 ± 0.001	0.782 ± 0.003
CLC Level 2	15	0.632 ± 0.008	0.755 ± 0.001	0.590 ± 0.010
CLC Level 3	34	0.484 ± 0.008	0.641 ± 0.024	0.406 ± 0.023
BigEarthNet	19	0.659 ± 0.007	0.673 ± 0.005	0.573 ± 0.020
Custom (SENSERO)	11	0.672 ± 0.020	0.744 ± 0.016	0.618 ± 0.025

Absolute mAP values are not directly comparable across vocabularies because class cardinality and prevalence differ. The corresponding macro-mAP chance floor ranges from 0.083 for CLC Level 3 to 0.363 for CLC Level 1. The CLC rows are the 128-pixel configurations of Experiment 1 and repeat the corresponding row of Table 3.

The comparison between CLC Level 3 and the custom SENSERO scheme provides a more targeted assessment of the regrouping introduced in Section 2.7 (Caption 5 of Table A2). At 128 pixels, macro mAP increases from 0.484 for CLC Level 3 to 0.672 for the custom scheme, a difference of 0.188. However, excluding the six Level-3 classes involved in the two regrouped families raises the Level-3 macro mAP only to 0.511. The affected classes therefore account for approximately 0.027, or 14%, of the aggregate difference between the two vocabularies, indicating that the overall performance gap cannot be attributed primarily to these regroupings.

At the level of the affected class families, the regrouping produces a clear local improvement in one family and an unresolvable one in the other (Table 6). For the shrubland family, where two of the three members are well supported (115 and 133 test positives;

236 for the merged class), the merged shrubland category reaches an F1 score of 0.475, above all three CLC Level-3 members (0.392, 0.448 and 0.317). For the artificial-land family, the merged quarries or dump sites category reaches 0.320, above dump sites (0.222) and construction sites (0.104) but below mineral extraction sites (0.354); with 4, 7 and 23 test positives respectively, these differences are within training variability and do not support a conclusion in either direction. The custom regrouping therefore improves classification performance where the merged family is adequately represented in the test set, while accounting for only a limited fraction of the aggregate difference between the custom and CLC Level-3 vocabularies.

Table 6. Per-class F1 scores for the two CLC Level-3 class families regrouped by the custom SENSERO scheme at the 128-pixel scale. The Level-3 column is computed from the CLC Level-3 models of Experiment 1 (Table 3), the merged column from the custom-scheme models of Experiment 3; values are means over three seeds using a decision threshold of 0.5.

CLC Code	CLC Level-3 Class	F_1 (Level 3)	F_1 (Custom, Merged)
131	mineral extraction sites	0.354	
132	dump sites	0.222	
133	construction sites	0.104	
quarries or dump sites (merged)		—	0.320
321	natural grasslands	0.392	
322	moors and heathland	0.448	
323	sclerophyllous vegetation	—	
324	transitional woodland-shrub	0.317	
shrubland (merged)		—	0.475

The custom scheme merges classes 131–133 into *quarries* or *dump sites* and classes 321, 322, and 324 into *shrubland*. Class 323 (*sclerophyllous vegetation*) is not represented at the 128-pixel scale. Test-set support for classes 131, 132, 133, 321, 322, and 324 is 23, 4, 7, 115, 8, and 133 samples, respectively; the merged classes have 34 and 236.

4.5. Robustness of the Scale Effect

This section reports the controls of Section 3.7. All of them concern CLC Level 3.

4.5.1. Spatial Leakage

Distance-stratified evaluation. The distance from a test patch to its nearest training patch ranges from 3392 m to 10,927 m, with quartile boundaries at 3753, 4008 and 4361 m: at 336 pixels, three quarters of the test patches lie within 1.3 patch widths of a training patch. Table A7 (Appendix B) gives the mean per-location AP of the Experiment 1 models inside each quartile. Two features stand out. Performance is flat across distance at every scale, the largest within-scale range being 0.03 at 64 pixels and 0.008 at 336; and the ordering of scales is identical in all four quartiles, with the plateau at 224–336 pixels visible in each column. The paired difference between 336 and 128 pixels is positive in every quartile and every seed (12/12; 11 of the 12 Wilcoxon tests significant after Holm correction, the exception being the nearest quartile of one seed, $p = 0.049$), averaging $+0.041$, $+0.050$, $+0.065$ and $+0.058$ from the nearest to the farthest quartile. The gain is thus smallest, not largest, where training patches are closest. Proximity does not account for the scale effect within the canonical partition.

Leave-scene-out cross-validation. Because the canonical partition places training and test patches in the same scenes, distance stratification can only show that the gain is stable *within* the neighbourhood of the training set; it cannot manufacture a far stratum that the design does not contain. Table A8 (Appendix B) therefore reports the leave-scene-out control. Under full scene separation, 336 pixels exceeds 128 pixels in every fold ($+0.071$, $+0.116$, $+0.137$, $+0.054$ and $+0.122$ macro mAP over the 24 classes present in every fold's test set), and on the pooled out-of-fold predictions over the full vocabulary

the two bootstrap intervals do not overlap (0.316 [0.309, 0.336] versus 0.410 [0.401, 0.427]; +0.094, against +0.123 on the canonical partition). The absolute level is 0.17–0.20 lower than on the canonical partition. Part of that drop is mechanical (one seed, 60% instead of 70% of the patches for training, and every rare class now evaluated only in scenes the model has never seen), and part is genuine within-scene similarity from which the canonical partition benefits. The canonical numbers should therefore be read as optimistic estimates of scene-level generalisation; the scale effect is not affected.

4.5.2. Labelling Rule

Regenerating the Level-3 label sets from the reference rasters with the area threshold alone reproduces the published sets on 10,000/10,000 patches at 64, 128, 224 and 336 pixels, which confirms that the centre-weighted step of the generator affects only token order (Section 2.7). Table A4 (Appendix A) describes how the sets change with the threshold: within a scale, moving from 0% to 5% reduces the mean number of labels by 15% at 64 pixels and 36% at 336, acting almost entirely on fragmented and linear classes (transitional woodland–shrub, roads, vineyards, complex cultivation) and leaving compact classes untouched; across scales, the mean number of labels grows by a factor of 2.2–2.4 at every threshold, and the most permissive rule at 64 pixels still yields fewer labels (2.24) than the most restrictive rule at 336 (3.39). No choice of threshold makes the label spaces comparable across scales, because a larger window contains more classes.

Figure 9 gives the trained half of the ablation. Across the whole 0–5% range, macro mAP moves by at most 0.020–0.042 within a scale, with no consistent direction (1% is best at 128 and 224 pixels, 2% at 336), whereas at any fixed threshold the effect of scale is 0.155–0.208 and the ordering 64 < 128 < 224 < 336 holds throughout. Along the iso-area diagonal, where the absolute threshold is held at 8–11 ha, the ordering is preserved (0.470 < 0.561 < 0.617). The runs regenerated at 2% agree with the published-label runs of Experiment 1 within seed variability (hollow circles). The labelling rule cannot, therefore, account for the scale effect.

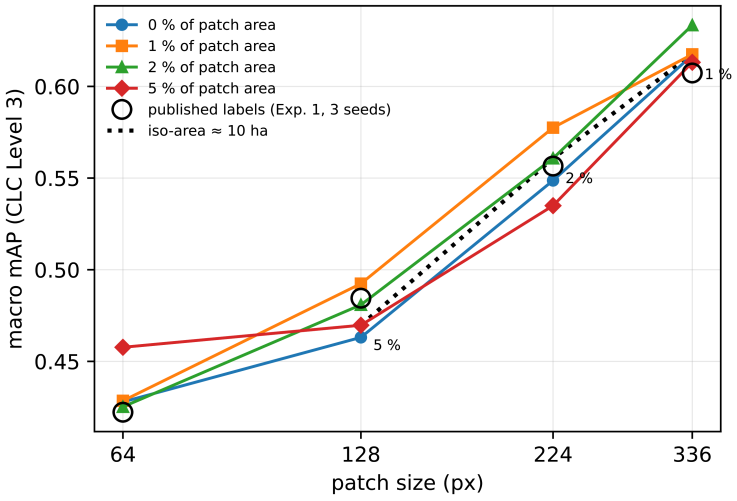


Figure 9. Labelling-rule ablation at CLC Level 3. Macro mAP against patch size for label sets generated at four area thresholds (one seed each); hollow circles are the published-label runs of Experiment 1 (three seeds); the dotted line joins the iso-area diagonal (≈10 ha absolute threshold).

4.5.3. Architecture

Table 7 compares the ResNet-50 trained from scratch for Experiment 1 with the two pretrained controls at the key scales. All three backbones rank the scales identically (64 < 128 < 224 < 336), with a 64 → 336 gain of +0.185, +0.282 and +0.179 macro mAP

respectively. Initialisation changes the level and the shape of the curve: the ImageNet-initialised ResNet-50 is below the from-scratch model at 64 pixels (0.384 ± 0.020 against 0.422 ± 0.005), where the $32\times$ downsampling of the stem leaves a 2×2 feature map and there is little to transfer, and above it from 128 pixels on ($+0.031$, $+0.031$ and $+0.059$, each at least two seed standard deviations); its $224 \rightarrow 336$ step ($+0.078$) is larger than the from-scratch model's ($+0.050$), so with pretraining the curve has not levelled off at the largest scale available. ViT-S/16 follows the same ordering at a lower level; it was trained with the schedule tuned for the convolutional baseline, stopped early on an unstable validation curve at 128 and 336 pixels (best epochs 15–17 against 33–53 for the ResNets) and observes only a 4×4 token grid at 64 pixels, so its absolute values are a lower bound rather than a comparison of architectures. The scale effect is therefore not an artefact of one backbone trained from scratch, while the location of the plateau depends on initialisation.

Table 7. Architecture control at CLC Level 3: macro mAP (mean $\pm$ SD over three seeds) at the key scales.

Backbone (Initialisation)	64	128	224	336
ResNet-50 (random; Experiment 1)	0.422 ± 0.005	0.484 ± 0.008	0.557 ± 0.013	0.607 ± 0.014
ResNet-50 (ImageNet)	0.384 ± 0.020	0.515 ± 0.016	0.588 ± 0.005	0.666 ± 0.014
ViT-S/16 (ImageNet)	0.374 ± 0.005	0.448 ± 0.027	0.533 ± 0.013	0.553 ± 0.028

4.6. Agreement with an Independent Land-Cover Product

On the test patches at 128 pixels, the median pixel agreement between CLC and WorldCover 2020 under the nine-group legend is 0.76 (mean 0.71; 0.74 at 336 pixels). The group confusion (Table A6, Appendix A) shows where the two products agree and where they differ: tree cover (96%), wetland (84%), water (75%), cropland (71%) and grassland (67%) agree as expected for two independent products at different minimum mapping units, while the artificial group agrees on only 27% of its pixels (the remainder being grassland, tree and cropland pixels inside CLC's use-based urban polygons) and CLC's shrubland group falls almost entirely on WorldCover tree (69%) and grassland (28%), with WorldCover 2020 mapping virtually no shrubland over Romania. Disagreement is thus concentrated in exactly the classes whose CLC definition summarises a mosaic.

Table A9 (Appendix C) evaluates the Experiment 1 models inside four strata of agreement. At 64 pixels performance tracks agreement almost linearly, from 0.70 where the CLC label is least literal to 0.91 where it is nearly pixel-exact. At 336 pixels the row is flat (0.89, 0.87, 0.86, 0.89). The scale gain is therefore concentrated where CLC generalises: the paired 336–64 pixel difference is $+0.18$ to $+0.21$ in the lowest stratum (three seeds, $p < 10^{-40}$), $+0.12$ to $+0.13$ in the second, $+0.03$ in the third, and -0.02 to -0.03 in the highest, significant in all three seeds. Two readings are consistent with this pattern and both are informative for practice. A model with a narrow field of view sees the pixel cover, and fails where the label departs from it; a model with a wide field of view sees the pattern that the CLC class names, and reproduces the label regardless of how literal it is. Where the patch is homogeneous and the label literal, the narrow window already scores 0.91, and the wide one gives up a little to the other classes it now has to carry. The classifier trained at large extent has thus learnt CLC's semantic generalisations along with its content, which is what a CLC-supervised classifier should do; the independent product makes that visible. Note that the lowest stratum is dominated by urban and transitional patches, which are also the classes that gain most from scale, so the stratum effect and the class effect are not fully separable; what is clean is the trend within each stratum.

5. Discussion

In the following, we discuss our findings and results with respect to several topics of interest, along with existing limitations. Additional discussions regarding the implications of the sampling design for metric interpretation, and the transferability of our methodology beyond Romania region can be found in Appendix D and Appendix E, respectively.

5.1. Interpreting the Effect of Spatial Scale

Varying patch size at a fixed ground sampling distance changes both the amount of spatial context available to the model and the composition of the target itself. As the footprint expands, additional land-cover classes enter the patch and the label vector becomes progressively denser. The scale effect observed in Section 4 should therefore be interpreted as the joint effect of increased spatial context and changing label composition, rather than as a measure of context alone. This distinction is important because models evaluated at different patch sizes operate on the same geographic locations and nomenclature, but not on targets of identical complexity.

The results further show that this scale effect depends on semantic granularity. After accounting for the scale-dependent chance floor, performance at CLC Level 1 remains essentially stable across the evaluated extents, whereas clear gains remain at Levels 2 and 3, with the strongest effect at Level 3. Additional spatial extent therefore provides little measurable benefit for the broadest land-cover categories, while becoming increasingly useful when the target vocabulary requires finer distinctions. The class-level analysis supports this interpretation: the largest gains occur for heterogeneous or spatially composite classes such as complex cultivation patterns, land principally occupied by agriculture with significant natural vegetation, and transitional woodland–shrub, for which a broader neighbourhood provides information about landscape composition rather than only local spectral appearance. The results presented in Sections 4.2 and 4.6 provide complementary perspectives on the same underlying mechanism. Many CLC Level-3 classes name a landscape pattern rather than a surface: *complex cultivation patterns* is a mosaic of small parcels, *discontinuous urban fabric* is houses interleaved with gardens and streets, *transitional woodland–shrub* is regrowth of mixed height. A 64-pixel window (640 m across) often falls on one component of such a pattern: a field, a roof, a stand of young trees, and a model that only sees the component cannot know which pattern it belongs to; this is why performance at small extents is lowest exactly where the CLC label agrees least with pixel-level cover (Table A9). A 336-pixel window (3.4 km across) shows the pattern itself, and the model reproduces the CLC label regardless of how literally it describes each pixel. Where the patch is a single homogeneous surface: a forest block, an arable plain, the small window already suffices and the large one gains nothing; its slight loss there is the cost of carrying several more classes in the same score vector. For practical Earth-observation work this means that the patch size a CLC-supervised classifier needs is set by the spatial grain of the nomenclature, not by the sensor: classes defined as mosaics require a field of view that contains the mosaic.

The benefit of increasing spatial extent is nevertheless limited. The paired per-location analysis shows its largest gains at the transitions from 64 to 112 pixels and from 128 to 224 pixels, followed by a plateau around 224–256 pixels. Beyond this range, increasing the footprint adds little additional per-location benefit, even though class-averaged mAP may continue to increase. This difference between the two views of performance suggests that larger patches do not improve all locations or classes uniformly: once sufficient spatial context is available, further increases in extent primarily affect particular classes and the composition of the multi-label target rather than producing a general improvement across

the test set. The plateau should therefore be interpreted as a property of this benchmark and model configuration, rather than as a universal optimal patch size.

5.2. Relation to Generative Captioning on the Same Patches

The scale dependence observed in the classification benchmark is consistent with our companion study on generative captioning using the same SENSERO patches, spatial scales, and CLC annotation levels [26]. In that study, captioning quality generally improved with increasing patch size, with the strongest scale dependence at CLC Level 3 and saturation at the larger extents for two of the three evaluated models. The present discriminative benchmark shows a similar overall pattern: the effect of scale is strongest at the finest semantic level, while the per-location improvement reaches a plateau at approximately 224–256 pixels.

The agreement between discriminative classification and generative captioning suggests that sensitivity to spatial extent is not specific to a particular prediction paradigm. Both tasks benefit from larger spatial support when fine-grained semantic distinctions are required, but the gains diminish once sufficient context is available. The correspondence is not exact: at CLC Level 3, class-averaged mAP continues to increase at the largest scales, whereas the paired per-location analysis plateaus earlier. This difference is consistent with the class-dependent effects observed in Figure 7 and reinforces the need to distinguish aggregate improvements from gains that occur consistently across individual locations.

5.3. Spatial Leakage, the Labelling Rule and the Baseline Architecture

The scale effect reported in Section 4.2 could have three mundane explanations, and the controls of Section 4.5 were designed to test each. The first is spatial leakage. The spatial partition prevents overlap between training and evaluation footprints at all scales, but because the same centres are used everywhere, the clearance between a test patch and its nearest training patch shrinks relative to the footprint as the footprint grows: at 336 pixels the closest test patch lies 1.01 patch widths from a training patch and three quarters lie within 1.3. A wide field of view might then sample a landscape that training patches also sample, and spatial autocorrelation [8,34] would inflate the numbers at large scales. The distance-stratified analysis shows no such signature: per-location AP does not vary with distance to the training set at any scale, and the 336-versus-128 gain is smallest in the nearest quartile. Leave-scene-out cross-validation, the strongest separation the release admits, confirms the gain ($+0.094$ macro mAP, non-overlapping bootstrap intervals) at an absolute level 0.17–0.20 below the canonical partition. The honest summary is therefore two-fold: within-scene similarity does raise the absolute numbers reported on the canonical partition, which should be read as optimistic estimates of scene-level generalisation, and it does not produce the difference between scales.

The second explanation is the labelling rule. A relative threshold admits a growing absolute area as the footprint grows, and one could suspect that the label sets at large scales are simply easier. The ablation shows that the threshold acts on fragmented classes and hardly at all on compact ones, that no threshold makes the label spaces of different scales comparable, and that the scale ordering is unchanged under every rule, including along an iso-area diagonal. The centre-weighted ordering step was shown to have no effect on set membership at all. The third explanation is the baseline architecture. The two pretrained controls rank the scales identically; what they change is the level and the location of the plateau, which under ImageNet initialisation has not been reached by 336 pixels. The plateau reported for the from-scratch ResNet-50 should therefore be read as a property of that configuration, whereas the monotone ordering of scales at Level 3 is not.

5.4. Nomenclature and Classification Difficulty

CLC classes are defined not only by physical land-cover characteristics but also by land-use and land-management criteria, and some fine-grained distinctions may therefore be difficult to infer from Sentinel-2 reflectance alone. This is particularly relevant at Level 3, where visually similar categories can correspond to distinct thematic classes. The lower performance observed for the finer CLC vocabulary should consequently be interpreted as reflecting both model limitations and the semantic difficulty of the target nomenclature, rather than classifier performance alone.

Experiment 3 provides a more specific view of this effect. Regrouping the two targeted Level-3 class families yields a higher F1 score for the merged shrubland category, and a difference within training variability for the merged quarries or dump sites category, whose members have too few test positives (Table 6); where it can be measured, the distinction contributes to classification difficulty. Their contribution to the aggregate difference is nevertheless limited: excluding the six affected Level-3 classes increases macro mAP by 0.027, corresponding to approximately 14% of the 0.188 gap between CLC Level 3 and the custom scheme. The remaining difference cannot therefore be attributed primarily to these two class families and is consistent with broader differences in vocabulary cardinality, class prevalence, and semantic composition. For this reason, the higher aggregate performance of the custom scheme should not be interpreted as evidence of a generally superior nomenclature, but as evidence that the choice and structure of the target vocabulary substantially influence measured benchmark difficulty.

5.5. Limitations

SENSERO inherits the radiometric and geometric characteristics of Sentinel-2 Level-2A imagery at 10 m. Although all scenes were processed using the same workflow and acquired within a nine-week period, residual variability due to atmospheric conditions, illumination geometry, and acquisition date remains. The reference annotations are derived from CLC 2018, whose reference year corresponds to that of the Sentinel-2 acquisitions, but CLC is a generalised land-cover inventory with a 25 ha minimum mapping unit. Fine-scale heterogeneity and features not represented in CLC are therefore absent from the target labels even when they are visible in the imagery. The benchmark should consequently be interpreted with respect to the CLC reference representation rather than as an exhaustive description of all observable land-cover detail.

SENSERO is currently a single-epoch, optical-only dataset and therefore does not support the evaluation of temporal robustness, change detection, or optical–SAR fusion. Its fixed geographic centres nevertheless provide a direct basis for future multi-temporal or multi-sensor extensions. The benchmark experiments are designed to isolate the effects of spatial extent and target vocabulary under a common training and evaluation protocol. The reported ResNet-50 results should therefore be interpreted as reproducible reference baselines for these controlled comparisons rather than as estimates of the best attainable performance on SENSERO.

In addition, per-location predictions were retained only for the matched-scale configurations of Experiment 2, so off-diagonal cross-scale comparisons are evaluated across seeds rather than through paired per-location tests. The robustness controls of Section 4.5 were run at CLC Level 3 and, with the exception of the architecture control, with one seed; the leave-scene-out and labelling-rule controls at the remaining scales and levels, and a ViT schedule tuned for that backbone, are left to future work. Finally, the comparison with WorldCover characterises where CLC labels generalise over the ground; it is not a per-pixel validation of the classifiers which, as patch-level models, cannot be scored per pixel.

6. Conclusions

We introduced SENSERO, a multi-scale, multi-label Sentinel-2 benchmark designed to study spatial extent as an explicit experimental variable while preserving the same geographic locations, reference source, and data partition across scales. The dataset provides eight spatial extents and five annotation vocabularies derived from CLC 2018, together with a reproducible benchmark protocol for patch-level multi-label classification. It also provides national-scale coverage of Romania, a territory largely absent from existing European and global EO benchmarks. Its contributions are as follows: the first EO benchmark in which patch size is an experimental axis rather than a fixed design constant; the first national-scale multi-label land-cover benchmark for Romania; deterministic, traceable annotation at three CLC levels, in the BigEarthNet nomenclature and in a Romania-adapted scheme; and a baseline suite that answers three questions no single-scale benchmark can pose, with controls for the three artefacts that could have produced the same answers.

The experiments show that the effect of spatial extent depends strongly on semantic granularity. Increasing patch size provides little measurable benefit for coarse CLC Level-1 classification after accounting for the changing prevalence baseline, whereas finer Level-2 and Level-3 targets benefit clearly from additional extent. At Level 3, the largest per-location gains occur at the transition from smaller to intermediate spatial extents and diminish around 224–256 pixels for the from-scratch baseline. Cross-scale evaluation shows that models are sensitive to mismatches between training and evaluation extent, with transfer from smaller to larger patches incurring a greater cost than the reverse. The nomenclature comparison shows that regrouping selected visually ambiguous Level-3 classes improves their local discriminability but explains only a limited fraction of the aggregate difference between the Level-3 and custom vocabularies. The finding that spatial context helps in proportion to how much the label generalises over heterogeneous ground is a statement about nomenclature and landscape grain, not about Romania. It should transfer to regions with a comparable mix of small-parcel agriculture, fragmented forest and dispersed settlement; much of Central and Eastern Europe, the Mediterranean and the Balkans; and to any nomenclature with mosaic classes.

The scale effect remains consistent across the three robustness evaluations. It does not depend on the distance between test and training patches, and it survives leave-scene-out cross-validation at a lower absolute level; it is unchanged under every area threshold of the labelling rule, including at constant absolute area; and it has the same ordering under an ImageNet-initialised ResNet-50 and a ViT-S/16, although the location of the plateau depends on initialisation. Comparison with an independent 10 m product locates the effect: spatial context helps in proportion to how much the CLC label generalises over heterogeneous ground, and slightly reduces performance where CLC labels closely correspond to observed surface characteristics.

Taken together, these results show that patch size and target nomenclature are coupled components of benchmark design rather than independent implementation choices: the field of view a classifier needs is set by the spatial grain of the classes it is asked to recognise. SENSERO provides a controlled setting in which these effects can be studied while keeping the underlying geographic sampling fixed, and its construction is a reproducible, low-cost recipe for benchmarks in other regions. Its common patch geometry supports extensions towards multi-temporal imagery, additional sensing modalities, and cross-dataset transfer. The dataset, processing pipeline, and benchmark protocol are released openly to support reproducible research on spatial scale and semantic representation in Earth observation.

Author Contributions: Conceptualization, T.C. and D.F.; methodology, software, data curation and visualization, T.C.; validation and investigation, T.C., V.V., A.-C.G. and A.G.; formal analysis, T.C.

and A.G.; writing—original draft preparation, T.C.; writing—review and editing, V.V., A.-C.G., A.G. and D.F.; supervision, project administration and funding acquisition, D.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the project "Platform for Fusion and Management of Multi-Source Data Collections Exploitable by Artificial Intelligence Models for Risk Estimation and Predictive Analysis—AI4RISK", Project Code: SOL3/2024—PN-IV-P6-6.3-SOL-2024-2-0251, funded by UEFISCDI (Executive Agency for Higher Education, Research, Development and Innovation Funding), Romania.

Data Availability Statement: The SENSERO dataset and the full processing pipeline are released openly under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The dataset is publicly available on Zenodo at https://doi.org/10.5281/zenodo.18847389 [35]. This is the concept DOI of the record, which always resolves to the most recent version; version 1, described and used in this paper, is archived under https://doi.org/10.5281/zenodo.18847390. The underlying Sentinel-2 and CORINE Land Cover 2018 data are provided under the Copernicus open data policy [43]. The dataset is distributed as 16 compressed archives (`.zip`), organised by data format (GeoTIFF or PNG) and patch size (64–336 pixels). Each archive contains 10,000 patches with their corresponding reference maps and a per-scale metadata file in Apache Parquet format (`metadata_<size>.parquet`) recording the CLC Level-3 codes, the five annotation layers, and the train/validation/test assignment. Two companion files are distributed alongside the archives: `sensero_patch_footprints.gpkg`, a GeoPackage vector index containing the spatial footprints and attributes of all 10,000 patches in a single layer (`patches`), projected in WGS-84 (EPSG:4326); and `sensero_technical_guide.pdf`, a technical guide documenting the folder hierarchy, file-naming conventions, metadata schemas, and the five labelling-scheme mapping tables. All custom Python scripts supporting the data-preparation workflow—Sentinel-2 patch extraction, CLC reference-map rasterisation, multi-label generation, and train/validation/test split construction—together with the code required to reproduce every experiment reported in Section 4, including training configurations, random seeds, and evaluation scripts, are openly available on GitHub at https://github.com/geosense-lab/SENSERO (accessed on 27 July 2026). Code examples for exploring and visualising the dataset are provided for both local execution and Google Colab. The pipeline was developed and tested using QGIS 3.40 "Bratislava", ESA SNAP v12.0.1, GDAL 3.8, and Python 3 on Ubuntu 22.04 LTS; only open-source software was used throughout.

Acknowledgments: This work contains modified Copernicus Sentinel data (2018) and CORINE Land Cover 2018 data produced with funding by the European Union.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BOA	Bottom of Atmosphere
CDSE	Copernicus Data Space Ecosystem
CLC	CORINE Land Cover
EO	Earth Observation
GSD	Ground Sampling Distance
mAP	mean Average Precision
MGRS	Military Grid Reference System
MMU	Minimum Mapping Unit
SCL	Scene Classification Layer
VLM	Vision–Language Model

Appendix A

Table A1. Cross-scale generalisation: mAP at CLC Level 3 for models trained at the row scale and evaluated at the column scale. Values are means over three seeds; bold entries on the diagonal correspond to matched training and evaluation scales and reproduce the CLC Level-3 results of Table 3).

Train/Test	64	112	120	128	224	256	280	336
64	**0.422**	0.399	0.428	0.415	0.441	0.440	0.438	0.441
112	0.422	**0.494**	0.488	0.484	0.518	0.522	0.516	0.526
120	0.421	0.479	**0.496**	0.491	0.529	0.526	0.519	0.530
128	0.418	0.471	0.480	**0.484**	0.524	0.521	0.522	0.534
224	0.341	0.470	0.480	0.471	**0.557**	0.560	0.560	0.574
256	0.352	0.483	0.509	0.476	0.568	**0.586**	0.574	0.592
280	0.341	0.472	0.477	0.474	0.571	0.583	**0.585**	0.597
336	0.333	0.448	0.471	0.467	0.557	0.565	0.577	**0.607**

Table A2. Combined correspondence between CLC 2018 classes and Captions 1–5.

CLC	Caption 1 (CLC L1)	Caption 2 (CLC L2)	Caption 3 (CLC L3)	Caption 4 (BigEarthNet)	Caption 5 (Custom)
111	artificial surfaces	urban fabric	continuous urban fabric	urban fabric	urban fabric
112	artificial surfaces	urban fabric	discontinuous urban fabric	urban fabric	urban fabric
121	artificial surfaces	industrial, commercial and transport units	industrial or commercial units	industrial or commercial units	urban fabric
122	artificial surfaces	industrial, commercial and transport units	road and rail networks and associated land	unlabeled	unlabeled
123	artificial surfaces	industrial, commercial and transport units	port areas	unlabeled	unlabeled
124	artificial surfaces	industrial, commercial and transport units	airports	unlabeled	unlabeled
131	artificial surfaces	mine, dump and construction sites	mineral extraction sites	unlabeled	quarries or dump sites
132	artificial surfaces	mine, dump and construction sites	dump sites	unlabeled	quarries or dump sites
133	artificial surfaces	mine, dump and construction sites	construction sites	unlabeled	quarries or dump sites
141	artificial surfaces	artificial non-agricultural vegetated areas	green urban areas	unlabeled	unlabeled
142	artificial surfaces	artificial non-agricultural vegetated areas	sport and leisure facilities	unlabeled	unlabeled
211	agricultural areas	arable land	non-irrigated arable land	arable land	agriculture
212	agricultural areas	arable land	permanently irrigated land	arable land	agriculture
213	agricultural areas	arable land	rice fields	arable land	agriculture
221	agricultural areas	permanent crops	vineyards	permanent crops	permanent crops
222	agricultural areas	permanent crops	fruit trees and berry plantations	permanent crops	permanent crops
231	agricultural areas	pastures	pastures	pastures	pastures
242	agricultural areas	heterogeneous agricultural areas	complex cultivation patterns	complex cultivation patterns	mixed farmland
243	agricultural areas	heterogeneous agricultural areas	land principally occupied by agriculture with significant areas of natural vegetation	land principally occupied by agriculture with significant areas of natural vegetation	mixed farmland
311	forest and semi-natural zones	forests	broad-leaved forest	broad-leaved forest	forests
312	forest and semi-natural zones	forests	coniferous forest	coniferous forest	forests
313	forest and semi-natural zones	forests	mixed forest	mixed forest	forests
321	forest and semi-natural zones	scrub and/or herbaceous vegetation associations	natural grasslands	natural grassland and sparsely vegetated areas	shrubland
322	forest and semi-natural zones	scrub and/or herbaceous vegetation associations	moors and heathland	moors heathland and sclerophyllous vegetation	shrubland

Table A2. *Cont.*

CLC	Caption 1 (CLC L1)	Caption 2 (CLC L2)	Caption 3 (CLC L3)	Caption 4 (BigEarthNet)	Caption 5 (Custom)
324	forest and semi-natural zones	scrub and/or herbaceous vegetation associations	transitional woodland–shrub	transitional woodland–shrub	shrubland
331	forest and semi-natural zones	open spaces with little or no vegetation	beaches dunes sands	beaches dunes sands	unlabeled
332	forest and semi-natural zones	open spaces with little or no vegetation	bare rocks	unlabeled	unlabeled
323	forest and semi-natural zones	scrub and/or herbaceous vegetation associations	sclerophyllous vegetation	moors heathland and sclerophyllous vegetation	shrubland
333	forest and semi-natural zones	open spaces with little or no vegetation	sparsely vegetated areas	natural grassland and sparsely vegetated areas	bare ground
411	wetlands	inland wetlands	inland marshes	inland wetlands	wetlands
421	wetlands	maritime wetlands	salt marshes	coastal wetlands	wetlands
511	water bodies	inland waters	water courses	inland waters	water bodies
512	water bodies	inland waters	water bodies	inland waters	water bodies
521	water bodies	marine waters	coastal lagoons	marine waters	water bodies
523	water bodies	marine waters	sea and ocean	marine waters	water bodies

Table A3. Number of test patches in which each CLC Level-3 class is present (test-set support), by patch size, under the fixed benchmark vocabulary. Rows sorted by minimum support; the 17 classes with a minimum of at least ten form the well-supported subset of Table 4.

CLC Level-3 Class	64	112	120	128	224	256	280	336
permanently irrigated land	0	1	1	1	1	1	2	2
green urban areas	1	1	1	1	4	6	6	7
port areas	1	3	3	3	4	4	4	4
airports	2	2	2	2	3	3	3	5
salt marshes	2	2	2	2	2	3	3	3
sparsely vegetated areas	2	2	2	3	3	3	3	3
continuous urban fabric	2	2	2	2	2	2	3	3
bare rocks	4	5	5	5	5	5	5	5
construction sites	4	6	7	7	7	9	10	11
dump sites	4	4	4	4	4	4	4	5
rice fields	5	7	8	8	9	9	10	11
moors and heathland	5	8	8	8	14	14	14	14
sport and leisure facilities	6	8	8	8	10	11	11	11
sea and ocean	6	6	7	7	7	7	8	8
coastal lagoons	7	7	7	9	9	9	9	10
road and rail networks	7	8	8	8	13	15	17	18
beaches dunes sands	8	11	11	11	15	15	15	15
mineral extraction sites	20	23	23	23	31	31	31	34
water bodies	41	51	53	54	71	77	79	89
fruit trees and berry plantations	42	80	84	87	136	152	164	195
vineyards	45	62	67	68	108	120	132	153
inland marshes	47	54	57	61	78	84	88	98
water courses	53	81	86	88	115	123	129	134
industrial or commercial units	56	72	73	75	98	105	106	122
natural grasslands	80	109	111	115	153	161	167	186
transitional woodland-shrub	84	118	127	133	221	241	265	303
mixed forest	97	124	126	128	169	176	182	193
coniferous forest	98	132	135	136	160	169	176	187
complex cultivation patterns	182	269	285	298	437	479	510	578
land principally occupied by agriculture	193	292	308	316	460	502	539	612
discontinuous urban fabric	314	407	417	434	574	619	663	742
pastures	401	571	601	620	836	879	904	983
broad-leaved forest	494	589	606	616	740	766	790	865
non-irrigated arable land	741	859	882	895	975	997	1008	1039

Sclerophyllous vegetation (323), present only at 336 pixels, is outside the fixed vocabulary and not listed. Support grows with scale for fragmented and mosaicked classes and is flat for compact ones.

Table A4. Descriptive effect of the labelling rule: CLC Level-3 label sets regenerated from the reference rasters at four area thresholds, all 10,000 patches.

Patch Size	Threshold (%)	Absolute Area (ha)	Mean Labels/Patch	Single-Label (%)	Classes Present
64	0	0.00	2.24	28.8	34
64	1	0.41	2.12	31.6	34
64	2	0.82	2.06	33.3	34
64	5	2.05	1.91	37.6	34
128	0	0.00	3.16	15.8	34
128	1	1.64	2.93	17.6	34
128	2	3.28	2.81	19.1	34
128	5	8.19	2.53	22.5	34
224	0	0.00	4.24	8.8	34
224	1	5.02	3.87	10.5	34
224	2	10.04	3.66	11.9	34
224	5	25.09	3.16	15.2	34
336	0	0.00	5.32	4.5	35
336	1	11.29	4.77	6.4	35
336	2	22.58	4.44	7.6	35
336	5	56.45	3.39	12.4	34

The 2% rows reproduce the published label sets exactly. At 5% and 336 pixels the absolute threshold exceeds the 25 ha CLC minimum mapping unit, and one class present in the map is no longer labelled. No patch is left without a label at any threshold.

Table A5. Nine-group legend used to compare CLC with ESA WorldCover 2020 (Section 3.8).

Group	CLC Level-3 Codes	WorldCover Classes
artificial	111–142	50 built-up
cropland	211–223, 241–244	40 cropland
grassland	231, 321	30 grassland
shrubland	322, 323, 324	20 shrubland
tree	311–313	10 tree cover, 95 mangroves
bare	331–334	60 bare/sparse vegetation, 100 moss and lichen
snow	335	70 snow and ice
wetland	411–423	90 herbaceous wetland
water	511–523	80 permanent water bodies

Green urban areas (141) and sport and leisure facilities (142) are kept in the artificial group, following CLC Level 1, although their physical cover is largely vegetated; the consequence is visible in Table A6.

Table A6. Pixel-level confusion between the CLC map (rows) and WorldCover 2020 (columns) over the 1500 test patches at 128 pixels, under the nine-group legend; rows normalised to one. The shrubland column is empty because WorldCover 2020 maps virtually no shrubland over Romania.

CLC/WorldCover	Artif.	Crop	Grass	Shrub	Tree	Bare	Snow	Wetl.	Water
artificial	**0.27**	0.17	0.23	0.00	0.25	0.07	0.00	0.00	0.01
cropland	0.01	**0.71**	0.19	0.00	0.08	0.01	0.00	0.00	0.00
grassland	0.00	0.13	**0.67**	0.00	0.18	0.01	0.00	0.01	0.00
shrubland	0.00	0.01	0.28	**0.00**	0.69	0.01	0.00	0.00	0.00
tree	0.00	0.00	0.03	0.00	**0.96**	0.00	0.00	0.00	0.00
bare	0.02	0.11	0.38	0.00	0.14	**0.12**	0.00	0.16	0.08
wetland	0.00	0.02	0.06	0.00	0.06	0.01	0.00	**0.84**	0.01
water	0.01	0.04	0.06	0.00	0.07	0.02	0.00	0.05	**0.75**

The snow row (CLC 335) has no pixels in the test patches and is omitted.

Appendix B

Table A7. Mean per-location average precision at CLC Level 3 (Experiment 1 models, mean over three seeds) inside quartiles of the distance from each test patch to its nearest training patch.

Patch Size	Q1: 3.39–3.75 km	Q2: 3.75–4.01 km	Q3: 4.01–4.36 km	Q4: 4.36–10.93 km
64	0.810	0.803	0.780	0.784
112	0.849	0.848	0.828	0.839
120	0.833	0.827	0.815	0.822
128	0.836	0.832	0.818	0.827
224	0.879	0.873	0.872	0.874
256	0.878	0.888	0.879	0.877
280	0.872	0.885	0.880	0.879
336	0.877	0.882	0.883	0.885

375 test patches per quartile. Distances are between patch centres; footprints never overlap (Section 2.8).

Table A8. Leave-scene-out cross-validation at CLC Level 3 (five folds of 7–8 Sentinel-2 scenes, one seed), compared with the canonical partition.

Patch Size	Per-Fold mAP, Common Classes	Pooled mAP, 34 Classes [95% CI]	Pooled Micro mAP	Canonical mAP (Table 3)
128	0.421 ± 0.022 (24 classes)	0.316 [0.309, 0.336]	0.689	0.484 ± 0.008
336	0.521 ± 0.023 (25 classes)	0.410 [0.401, 0.427]	0.793	0.607 ± 0.014
Δ	+0.100	+0.094	+0.104	+0.123

Per-fold values are mean ± SD over folds; "common classes" are those with at least one positive in every fold's test set. Pooled values score each of the 10,000 patches once, by the model whose training folds exclude its scene; the interval is a bootstrap over patches.

Appendix C

Table A9. Mean per-location average precision at CLC Level 3 (Experiment 1 models, mean over three seeds) inside strata of pixel agreement between the CLC map of each test patch and ESA WorldCover 2020 under the nine-group legend of Table A5 (agreement computed at 128 pixels and applied as a patch property to all scales).

Patch Size	Agreement < 0.6 ($n = 484$)	0.6–0.8 ($n = 340$)	0.8–0.9 ($n = 205$)	≥ 0.9 ($n = 471$)
64	0.696	0.749	0.828	0.914
112	0.789	0.809	0.838	0.919
120	0.761	0.796	0.831	0.907
128	0.775	0.797	0.827	0.906
224	0.863	0.854	0.865	0.905
256	0.878	0.867	0.860	0.902
280	0.880	0.871	0.855	0.894
336	0.890	0.872	0.860	0.889

Appendix D

SENSERO uses stratified sampling to ensure coverage of land-cover classes across the benchmark partitions, including classes that occupy only a small fraction of the Romanian territory. This design supports evaluation across a broad CLC vocabulary, but it also means that the benchmark class distribution should not be interpreted as an estimate of the national land-cover distribution. Table A10 illustrates this difference by comparing the national area share of each class with its share of label occurrences in SENSERO. The two quantities are not directly equivalent, however, because class presence within a patch does not measure the area occupied by that class; fragmented or linear classes may therefore occur in many patches while covering a comparatively small area.

Table A10. Comparison between national area share and SENSERO label-occurrence share for CLC Level-3 classes at the 128-pixel scale.

Code	CLC Level-3 Class	National (%)	SENSERO (%)	Ratio	Patches
121	Industrial or commercial units	0.448	1.577	3.52	443
511	Water courses	0.729	1.918	2.63	539
112	Discontinuous urban fabric	4.578	10.406	2.27	2924
242	Complex cultivation patterns	3.426	7.068	2.06	1986
243	Land principally occupied by agric.	3.764	7.406	1.97	2081
512	Water bodies	0.651	1.235	1.90	347
221	Vineyards	0.923	1.591	1.72	447
222	Fruit trees and berry plantations	1.145	1.947	1.70	547
324	Transitional woodland-shrub	1.878	3.068	1.63	862
231	Pastures	10.785	14.727	1.37	4138
411	Inland marshes	1.183	1.488	1.26	418
321	Natural grasslands	2.378	2.744	1.15	771
313	Mixed forest	4.291	3.192	0.74	897
311	Broad-leaved forest	20.112	14.695	0.73	4129
521	Coastal lagoons	0.285	0.199	0.70	56
312	Coniferous forest	4.887	3.182	0.65	894
213	Rice fields	0.240	0.146	0.61	41
211	Non-irrigated arable land	35.317	21.261	0.60	5974
322	Moors and heathland	0.287	0.157	0.55	44
523	Sea and ocean	2.105	0.167	0.08	47

Both shares are normalised over the classes included in the comparison. The ratio is defined as the SENSERO label-occurrence share divided by the national area share; a value of 1 indicates equal shares under the two measures. Class occurrence and areal coverage are distinct quantities, and the ratio should therefore not be interpreted as a direct measure of sampling over- or under-representation.

This distinction is important when interpreting aggregate performance metrics. Macro-averaged metrics are particularly informative for SENSERO because they assign equal weight to each class and therefore prevent the most frequent classes from dominating the evaluation. At CLC Level 3, however, 16 of the 34 classes represented in the test partition have fewer than ten positive samples at 128 pixels, 17 at 64 pixels and still 10 at 336 (Table A3). The sampling design guarantees that every class occurs in every partition, but it cannot make a class that covers 130 ha of the country common, and it does not attempt areal representativeness. The consequence for the metrics is quantified in Table 4: an average precision computed on one to five positives takes almost discrete values, so the bootstrap interval of the 34-class macro mAP is about ± 0.05 wide and skewed, several times the seed variability, whereas the interval over the 17 classes with adequate support is ± 0.02 and symmetric. Per-class trajectories of the rare classes across scales, which earlier drafts of this work were tempted to interpret, are not statistically resolvable; the class-level statements of Section 4.2 are accordingly restricted to well-supported classes. Users of the benchmark are advised to report the well-supported macro average alongside the full one, to prefer paired per-location comparisons for differences between configurations, and to read absolute macro values with the full-vocabulary interval in mind.

Micro-averaged metrics provide a complementary view, weighted towards the more frequent classes. Neither macro- nor micro-averaged performance, however, should be interpreted as an estimate of national mapping accuracy. The reported metrics are intended primarily for comparing models and experimental configurations under the common SENSERO sampling design.

Appendix E

Although SENSERO is a Romanian dataset, nothing in its construction is specific to Romania. The pipeline consists of fixed patch centres, extraction at several nested extents around each centre, and rule-based generation of multi-label sets from a rasterised reference map with an area threshold; each of these steps takes a country as a parameter. Four consequences follow, and each deserves stating with its caveat.

Methodology. Within the 39 countries covered by CLC, the pipeline can be re-run without modification: the same nomenclature, the same 25 ha MMU, the same 2018 reference year, and the same label-generation rule, whose sole free parameter (the area threshold) is shown in Section 4.5 not to drive the results. Outside Europe, the reference layer must be substituted, ESA WorldCover or Dynamic World [44] for a global 10 m physical-cover map, or national inventories such as NLCD or an LCCS-based product, and the rule applies unchanged, with one difference that users should expect: a pixel-level cover map has no MMU, so the label sets will be denser at every scale and the "literal" agreement stratum of Section 4.6 will dominate. Section 3.8 shows the crosswalk needed to relate such a product to CLC.

Cross-country transfer and domain adaptation. Because the BigEarthNet nomenclature is one of SENSERO's five layers and 120 pixels is one of its scales, a model trained on BigEarthNet can be evaluated on SENSERO without any change of label space or input size, over a territory that BigEarthNet's ten countries do not touch. This makes the dataset a ready-made target domain for cross-country transfer and unsupervised domain adaptation, with the additional and unusual property that the transfer can be studied as a function of the source model's patch size against a target available at eight. We do not pursue this here, because the controls it needs (source-side seeds, matched preprocessing) are orthogonal to the ones established in this paper, and treat it in a companion study.

Low-cost benchmarks for data-scarce regions. The dataset required one analyst-season of manual centre placement, 38 freely available Sentinel-2 scenes, a freely available reference inventory and open-source software (QGIS, SNAP, GDAL, Python). No field campaign and no annotation crowd were involved, and the quality controls of Section 2.10 are all programmatic or visual. For countries or regions absent from the large international benchmarks, which, as Figure 1 shows, includes an EU member state, this is a reproducible methodology for a national benchmark whose labels are traceable to an authoritative map. Its weakness is inherited from the reference layer, and Section 4.6 shows how to make that inheritance visible with an independent product rather than hide it.

References

1. Tuia, D.; Schindler, K.; Demir, B.; Zhu, X.X.; Kochupillai, M.; Džeroski, S.; van Rijn, J.N.; Hoos, H.H.; Frate, F.D.; Datcu, M.; et al. Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward. *IEEE Geosci. Remote. Sens. Mag.* **2025**, *13*, 119–141. https://doi.org/10.1109/MGRS.2024.3425961.
2. Gevaert, C.M. Explainable AI for earth observation: A review including societal and regulatory perspectives. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102869. https://doi.org/https://doi.org/10.1016/j.jag.2022.102869.
3. Haq, M.A.; Ahmed, A.; Khan, I.; Gyani, J.; Mohame, A.; Attia, E.-A.; Mangan, P.; Pandi, D. Analysis of environmental factors using AI and ML methods. *Sci. Rep.* **2022**, *12*, 13267. https://doi.org/10.1038/s41598-022-16665-7.
4. Janga, B.; Asamani, G.P.; Sun, Z.; Cristea, N. A Review of Practical AI for Remote Sensing in Earth Sciences. *Remote. Sens.* **2023**, *15*, 4112. https://doi.org/10.3390/rs15164112.
5. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2019**, *12*, 2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242.
6. Sumbul, G.; Charfuelan, M.; Demir, B.; Markl, V. BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In Proceeding of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 5901–5904. https://doi.org/10.1109/IGARSS.2019.8900532.

7. Sumbul, G.; de Wall, A.; Kreuziger, T.; Marcelino, F.; Costa, H.; Benevides, P.; Caetano, M.; Demir, B.; Markl, V. BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]. *IEEE Geosci. Remote. Sens. Mag.* **2021**, *9*, 174–180. https://doi.org/10.1109/MGRS.2021.3089174.
8. Clasen, K.N.; Hackel, L.; Burgert, T.; Sumbul, G.; Demir, B.; Markl, V. reBen: Refined BigEarthNet Dataset for Remote Sensing Image Analysis. In Proceedings of the IGARSS 2025—2025 IEEE International Geoscience and Remote Sensing Symposium, Brisbane, Australia, 3–8 August 2025; pp. 1264–1268. https://doi.org/10.1109/IGARSS55030.2025.11242834.
9. Schmitt, M.; Hughes, L.H.; Qiu, C.; Zhu, X.X. SEN12MS—A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2019**, *IV-2/W7*, 153–160. https://doi.org/10.5194/isprs-annals-IV-2-W7-153-2019.
10. Zhu, X.X.; Hu, J.; Qiu, C.; Shi, Y.; Kang, J.; Mou, L.; Bagheri, H.; Häberle, M.; Hua, Y.; Huang, R.; et al. So2Sat LCZ42: A Benchmark Dataset for Global Local Climate Zones Classification. *IEEE Geosci. Remote. Sens. Mag.* **2019**. https://doi.org/10.14459/2018mp1483140.
11. Benhammou, Y.; Alcaraz-Segura, D.; Guirado, E.; Khaldi, R.; Achchab, B.; Herrera, F.; Tabik, S. Sentinel2GlobalLULC: A Sentinel-2 RGB image tile dataset for global land use/cover mapping with deep learning. *Sci. Data* **2022**, *9*, 681. https://doi.org/10.1038/s41597-022-01775-8.
12. AleMohammad, H.; Ballantyne, A.; Bromberg Gaber, Y.; Booth, K.; Nakanuku-Diggs, L.; Miglarese, A.; Richards, J.; Rahman, A.A.; Clay, R.; Boucher, M.; et al. LandCoverNet: A global land cover classification training dataset. *Remote. Sens.* **2022**, *14*, 3577. https://doi.org/10.3390/rs14153577.
13. Christie, G.; Fendley, N.; Wilson, J.; Mukherjee, R. Functional Map of the World. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 July 2018; pp. 6172–6180.
14. Wang, Y.; Braham, N.A.A.; Xiong, Z.; Liu, C.; Albrecht, C.M.; Zhu, X.X. SSL4EO-S12: A Large-Scale Multimodal, Multitemporal Dataset for Self-Supervised Learning in Earth Observation. *IEEE Geosci. Remote. Sens. Mag.* **2023**, *11*, 98–106.
15. Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; Kembhavi, A. SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 16772–16782.
16. Nedungadi, V.; Kariryaa, A.; Oehmcke, S.; Belongie, S.; Igel, C.; Lang, N. MMEarth: Exploring Multi-Modal Pretext Tasks for Geospatial Representation Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
17. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), San Jose, CA, USA, 2–5 November 2010; pp. 270–279. https://doi.org/10.1145/1869790.1869829.
18. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote. Sens.* **2017**, *55*, 3965–3981. https://doi.org/10.1109/TGRS.2017.2685945.
19. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998.
20. Fazzini, P.; De Felice Proia, G.; Adamo, M.; Blonda, P.; Petracchini, F.; Forte, L.; Tarantino, C. Sentinel-2 Remote Sensed Image Classification with Patchwise Trained ConvNets for Grassland Habitat Discrimination. *Remote. Sens.* **2021**, *13*, 2276. https://doi.org/10.3390/rs13122276.
21. Malczewska, A.; Malczewski, J.; Hejmanowska, B. Challenges in preparing datasets for super-resolution on the example of sentinel-2 and planet scope images. *Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2023**, *XLVIII-1/W3-2023*, 91–98. https://doi.org/10.5194/isprs-archives-XLVIII-1-W3-2023-91-2023.
22. Reed, C.J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; Darrell, T. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4065–4076.
23. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote. Sens.* **2017**, *55*, 5200–5211. https://doi.org/10.1109/TGRS.2017.2776321.
24. Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; Sun, X. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 4404119. https://doi.org/10.1109/TGRS.2021.3078451.
25. Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5642123. https://doi.org/10.1109/TGRS.2024.3449154.
26. Costachioiu, T.; Vasilescu, V.; Faur, D. On the Impact of Patch Size in LLM-Based Image Captioning of Sentinel-2 Imagery. In Proceedings of the IGARSS 2026—2026 IEEE International Geoscience and Remote Sensing Symposium, Washington, DC, USA, 9–4 August 2026.

27. Băicoianu, A.; Plajer, I.C.; Debu, M.; Ștefan, M.; Ivanovici, M.; Florea, C.; Cațaron, A.; Coliban, R.M.; Popa, Ș.; Oprișescu, Ș.; et al. DACIA5: A Sentinel-1 and Sentinel-2 Dataset for Agricultural Crop Identification Applications. *Big Earth Data* **2025**, *9*, 1226–1257. https://doi.org/10.1080/20964471.2025.2512685.
28. Rujoiu-Mare, M.R.; Olariu, B.; Mihai, B.A.; Nistor, C.; Săvulescu, I. Land Cover Classification in Romanian Carpathians and Subcarpathians Using Multi-Date Sentinel-2 Remote Sensing Imagery. *Eur. J. Remote. Sens.* **2017**, *50*, 496–508. https://doi.org/10.1080/22797254.2017.1365570.
29. CORINE Land Cover 2018 (Vector), Europe, 6-Yearly, Version 2020_20u1. European Union's Copernicus Land Monitoring Service information, European Environment Agency (EEA). 2018. Available online: https://land.copernicus.eu/en/products/corine-land-cover/clc2018 (accessed on 8 June 2026).
30. Copernicus Sentinel-2 (processed by ESA). *MSI Level-2A BOA Reflectance Product. Collection 1*; Dataset; European Space Agency: Paris, France, 2021. https://doi.org/10.5270/S2_-znk9xsj.
31. Copernicus Land Monitoring Service. *CLC2018 Technical Guidelines*; Technical Report; European Environment Agency (EEA): Copenhagen, Denmark, 2017.
32. European Space Agency. *Sentinel-2 Products Specification Document*; Technical Report S2-PDGS-CS-DI-PSD, Issue 15.0; ESA: Paris, France, 2024.
33. Sechidis, K.; Tsoumakas, G.; Vlahavas, I.P. On the Stratification of Multi-Label Data. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases—European Conference, ECML PKDD 2011 Part III*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6913, pp. 145–158. https://doi.org/10.1007/978-3-642-23808-6_10.
34. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234–240. https://doi.org/10.2307/143141.
35. Costachioiu, T.; Vasilescu, V.; Grivei, A.C.; Griparis, A.; Faur, D. *SENSERO: Semantic Earth-Observation and Natural-Language Scene Explanation for Romania (2018)*; Dataset; Zenodo: Geneva, Switzerland, 2026. https://doi.org/10.5281/zenodo.18847390.
36. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. https://doi.org/10.2307/3001968.
37. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70. https://doi.org/10.2307/4615733.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
39. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
41. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: Boca Raton, Florida, USA, 1994.
42. Zanaga, D.; Van De Kerchove, R.; De Keersmaecker, W.; Souverijns, N.; Brockmann, C.; Quast, R.; Wevers, J.; Grosu, A.; Paccini, A.; Vergnaud, S.; et al. *ESA WorldCover 10 m 2020 v100*; Zenodo: Geneva, Switzerland, 2021. https://doi.org/10.5281/zenodo.5571936.
43. Copernicus Land Monitoring Service. Data Policy. Available online: https://land.copernicus.eu/en/data-policy (accessed on 27 July 2026).
44. Brown, C.F.; Brumby, S.P.; Guzder-Williams, B.; Birch, T.; Hyde, S.B.; Mazzariello, J.; Czerwinski, W.; Pasquarella, V.J.; Haertel, R.; Ilyushchenko, S.; et al. Dynamic World, Near Real-Time Global 10 m Land Use Land Cover Mapping. *Sci. Data* **2022**, *9*, 251. https://doi.org/10.1038/s41597-022-01307-4.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.