Received: 1 October 2025 | Revised: 3 March 2026 | Accepted: 23 March 2026 | Published online: 30 April 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–9
DOI: 10.47852/bonviewJDSIS62027813

BON VIEW PUBLISHING

Contrastive Pretrain and Supervised Fine-Tune for Land-Cover Classification in Data-Scarce Countries

Gianfausto Bottini[1], Francesco Nicola Tubiello[1] and Pengyu Hao[2,*]

[1] *Environment Statistics, Food and Agriculture Organization of the United Nations, Italy*

[2] *Agro-informatics, Food and Agriculture Organization of the United Nations, Italy*

Abstract: Deep neural networks are widely used to extract patterns and features from satellite imagery in Earth Observation, but their reliance on large labeled datasets limits deployment in data-scarce regions where ground surveys are expensive or infeasible. This study introduces a novel operational pipeline that couples SimCLR contrastive pretraining on unlabeled Sentinel-2 tiles with lightweight supervised fine-tuning to deliver tile-level predictions that are aggregated into district-level land-cover statistics, explicitly targeting administrative reporting rather than pixel-wise mapping. The key methodological contribution is this tile-to-district design—together with domain-aware, conservative augmentations and RGB-only pretraining—that aims to learn transferable representations under seasonality and atmospheric variability while requiring only a small expert-labeled set. The approach is evaluated in challenging contexts (with a pilot case study in Lesotho) and can generate updated district summaries within hours once trained, supporting rapid deployment where labels and compute are constrained. Benchmarking against ESA WorldCover and Google Dynamic World shows strong coherence of aggregated class distributions using cosine similarity, complemented by standard classification metrics on labeled tiles, indicating that reliable administrative-level indicators can be produced even without per-pixel agreement. These results underscore the practical relevance of the method for national statistical offices, humanitarian programs, and monitoring initiatives that need scalable, affordable land-cover statistics while minimizing dependence on extensive field campaigns.

Keywords: contrastive learning, cost-efficient pipeline, Sentinel-2, low-income food-deficit countries, zonal statistics

1. Introduction

Deep learning has substantially expanded what can be extracted from Earth Observation (EO) imagery, often outperforming traditional machine learning baselines in land-cover classification when large labeled datasets are available. However, in many data-scarce countries, reliable in situ labels are expensive or infeasible to collect, creating a persistent bottleneck for deploying high-performing models at the national scale [1–3].

In data-scarce countries, limited field data and heterogeneous landscapes make fine-grained land-cover legends difficult to train and validate reliably at the national scale [4–5]. We therefore adopt a compact, operational legend aligned with ESA WorldCover and focus on seven high-level classes (tree cover, shrubland, grassland, cropland, built-up, bare/sparse vegetation, and permanent water), which capture the dominant land-use/land-cover processes needed for reporting while reflecting the main sources of ambiguity in these settings (e.g., shrub–grass transitions, cropland–natural vegetation mosaics, and sparse cover in semi-arid areas) [6]. Cropland extent is a key reporting variable, and its global measurement remains an active research topic, motivating transparent and reproducible EO-based reporting workflows [7].

This study addresses that bottleneck with a semi-supervised pipeline that learns transferable representations from unlabeled Sentinel-2 tiles via contrastive pretraining and then adapts them using a small expert-labeled set through supervised fine-tuning. The output is intentionally designed for operational use: tile-level predictions are aggregated into district-level land-cover class distributions, enabling administrative reporting without requiring pixel-accurate mapping.

This design differs from common past approaches in two ways. First, compared to fully supervised land-cover pipelines, the proposed method reduces dependence on large, annotated datasets by shifting most learning to self-supervised pretraining on readily available imagery. Second, unlike pixel-wise land-cover mapping methods and global products that emphasize per-pixel agreement, the pipeline prioritizes administrative-scale indicators by aggregating tile predictions to districts and validating coherence of class distributions against ESA WorldCover and Google Dynamic World using cosine similarity alongside standard classification metrics [1, 2, 5]. These choices align with recent progress in self-supervised learning for EO and emerging large-scale pretraining resources (e.g., SSL4EO), while focusing the contribution

***Corresponding author:** Pengyu Hao, Agro-informatics, Food and Agriculture Organization of the United Nations, Italy. Email: pengyu.hao@fao.org

on a pragmatic reporting unit that matches national monitoring workflows in data-scarce settings. In addition, because the workflow relies on openly available Sentinel-2 imagery and can be executed on widely accessible, low-cost compute, recurring production of district-level indicators has a near-zero marginal cost in terms of software licensing and data acquisition.

Section 2 clarifies how this contribution should be interpreted relative to existing EO self-supervised learning literature and specifies what is (and is not) claimed as novel.

The remainder of the paper is structured as follows: Section 2 details the methodological novelty; Section 3 situates the work within the literature; Section 4 presents the methodology, data, and results; and Section 5 discusses advantages, limitations, and future directions.

2. Positioning of the Contribution

Contrastive self-supervised learning for EO has been studied extensively, including approaches that pretrain on remote sensing imagery and transfer to downstream classification tasks. Accordingly, we do not claim novelty in the general idea of "contrastive pretraining followed by supervised fine-tuning" on EO data; instead, we focus on how this learning paradigm is packaged and evaluated for administrative reporting in data-scarce settings [3, 8].

Our main contribution is an operational pipeline whose primary deliverable is district-level land-cover class distributions derived by aggregating tile-level predictions, rather than pixel-accurate land-cover maps. This tile-to-district design aligns model outputs with the unit of analysis used in many national reporting and humanitarian monitoring workflows, where stable administrative indicators are often more actionable than detailed boundaries.

A second contribution is the evaluation protocol: we validate administrative-scale usability by comparing aggregated class distributions against widely used 10 m reference products (ESA WorldCover and Google Dynamic World) using cosine similarity, complemented by standard classification metrics on labeled tiles where available. This choice makes the validation consistent with the intended use (district indicators) and avoids treating pixel-level disagreement as the sole proxy for usefulness when the reporting unit is coarser than the pixel.

Finally, the pipeline is designed around practical deployment constraints typical of data-scarce contexts: limited expert labels, limited compute, and the need for rapid updates once a pretrained encoder is available. In our implementation, the workflow relies on openly available Sentinel-2 imagery and can be executed on widely accessible, low-cost compute (e.g., a single GPU session), so that recurring production of district-level indicators has a near-zero marginal cost in terms of software licensing and data acquisition.

Within this operational framing, we adopt conservative, semantics-preserving augmentations and restrict pretraining inputs to RGB to reduce sensitivity to seasonal and atmospheric variability while keeping the training recipe simple and reproducible.

3. Literature Review

Land-cover mapping is a long-standing and high-impact application of EO, but modern deep learning approaches often remain difficult to operationalize in data-scarce countries because they depend on large, curated labeled datasets. In many low-resource settings, field surveys and systematic labeling campaigns are expensive, logistically constrained, and sometimes infeasible, which creates a persistent bottleneck even when satellite imagery is openly available at high spatial and temporal resolution. In response, recent EO literature has increasingly shifted from purely supervised learning toward self-supervised and contrastive representation learning, where models learn transferable features from abundant unlabeled archives and are later adapted using small labeled sets.

Recent surveys document the rapid rise of self-supervised learning in EO as a practical answer to label scarcity and as a means to improve transfer across sensors, regions, and seasons [3, 9]. These reviews highlight that contrastive learning, masked-image modeling, and distillation-based pretraining often improve downstream performance under low-label regimes for classification, segmentation, and change detection tasks [10, 11]. A key theme is the need to adapt general self-supervised recipes to EO-specific variability—seasonality, atmospheric effects, and subtle spectral similarity—so that learned invariances do not erase land-cover semantics. Seasonal Contrast (SeCo), for example, explicitly leverages seasonal diversity during pretraining and reports improved generalization for land-cover tasks when models are exposed to multi-season imagery [12].

In parallel, the land-cover community increasingly recognizes that evaluation should reflect intended use. Many operational users (e.g., statistical agencies and monitoring programs) require stable, interpretable indicators at administrative scales rather than pixel-perfect boundaries. This has motivated evaluation approaches that consider agreement of aggregated class composition, especially when dense ground truth is unavailable and when external reference products disagree at the pixel level. The present work is situated at this intersection: it uses contrastive pretraining to mitigate label scarcity and frames outputs and validation around district-level reporting needs rather than purely pixel-wise mapping objectives.

3.1. Self-supervised trends

Self-supervised learning has become a dominant paradigm for EO pretraining because EO archives are large, diverse, and largely unlabeled, yet contain strong latent structure that models can exploit (e.g., repeated seasonal cycles, settlement morphology, hydrological signatures, and vegetation texture). In general, the goal of self-supervised pretraining is to learn a representation space that captures stable and transferable features, after which a lightweight supervised step can adapt the representation to a specific legend and task.

In EO, three families of self-supervised methods are particularly common. First, contrastive learning learns invariances by bringing embeddings of two augmented views of the same sample closer together and pushing embeddings of different samples apart; this class includes SimCLR-style frameworks, which have become a standard baseline [8]. Related Siamese approaches such as SimSiam learn invariances without explicit negative pairs, providing an alternative route to robust representations under limited labels [13]. Redundancy-reduction objectives, exemplified by Barlow Twins, also encourage informative and non-collapsed features without relying on negative sampling [14, 15]. Second, masked-image modeling learns to reconstruct missing patches or predict masked content, which can emphasize spatial context and multi-scale structure [16]. Third, distillation and clustering-based approaches learn representations by matching student outputs to a teacher signal or to prototype assignments, often improving stability and reducing the need for explicit negative sampling [17, 18]. Reviews emphasize that these methods consistently

outperform supervised-only baselines when labels are limited, and that they tend to produce features that transfer better across geographies and acquisition conditions than features learned from small supervised sets [3, 9, 19].

A practical implication for operational pipelines is that the value of self-supervised learning depends on how well the pretraining objective and augmentations capture "nuisance" variability without destroying semantics. In land-cover problems, many classes differ by subtle cues (texture, structure, and weak spectral differences), and these cues can shift with season and management practices. This makes augmentation design particularly important: overly aggressive transforms can remove information needed to separate classes such as shrubland, grassland, and cropland, whereas overly weak transforms can produce representations that do not generalize. EO-specific work therefore often favors conservative, semantics-preserving augmentations as a safe default when building an operational pipeline intended to work in multiple regions and under limited supervision.

Large-scale EO resources have further accelerated these trends. SSL4EO-S12, for example, provides a large collection of multimodal Sentinel-1/2 tiles designed for EO self-supervised pretraining and benchmarking, supporting a range of downstream tasks under limited labels [11]. Complementary efforts extend the paradigm to other sensors and longer time horizons, enabling cross-sensor transfer when operational needs require long records or different spatial resolutions [11]. This shift toward "foundation-style" EO pretraining also motivates methods to handle redundancy in archives and reduce pretraining costs while preserving representation quality, which is particularly relevant when compute is constrained [10].

Within this landscape, our approach adapts SimCLR [8] to Sentinel-2 tiles using conservative RGB augmentations, prioritizing robustness and reproducibility under operational constraints. We emphasize downstream utility for administrative reporting by aggregating tile-level predictions to districts and validating coherence of class composition rather than focusing exclusively on pixel-level agreement [3].

3.2. Land-cover context

Land cover describes the biophysical characteristics of the Earth's surface shaped by natural processes and human activity, and EO enables systematic observation of these characteristics through repeated measurements from satellite sensors. In operational settings—particularly in data-scarce countries—land-cover products are often used for monitoring and reporting, where stakeholders require interpretable indicators that can be updated regularly and compared across time and administrative units. This practical context influences both legend design and evaluation.

At the product level, recent global 10 m land-cover datasets and near-real-time probabilistic products provide important reference points. ESA WorldCover provides global maps with documented class definitions and product guidance that facilitate comparisons at policy-relevant scales [2]. Dynamic World provides frequent 10 m class probabilities from Sentinel-2, enabling rapid updates and offering a practical benchmark for aggregated reporting [1]. However, comparative studies show that pixel-level disagreement among 10 m global products can be substantial, even when products report high overall accuracy, due to differences in legends, temporal support, training labels, and modeling choices [5]. As a result, validation strategies that focus only on pixel-wise agreement can overstate disagreement relative to what matters for reporting and can under-represent the utility of aggregated indicators.

Methodologically, land-cover classification has shifted from classical machine learning approaches (e.g., Random Forests and SVMs stands for Support Vector Machines (SVMs)) toward convolutional architectures (e.g., ResNet variants), which learn hierarchical spectral–spatial features and often transfer better across regions. Yet these gains frequently depend on large labeled datasets, which are precisely what data-scarce countries lack, motivating approaches that reduce label dependence and enable rapid deployment. In parallel, operational reporting frameworks often emphasize compact, high-level classes that are interpretable and comparable across contexts, reinforcing the practicality of high-level legends aligned with widely used products and reporting needs [6].

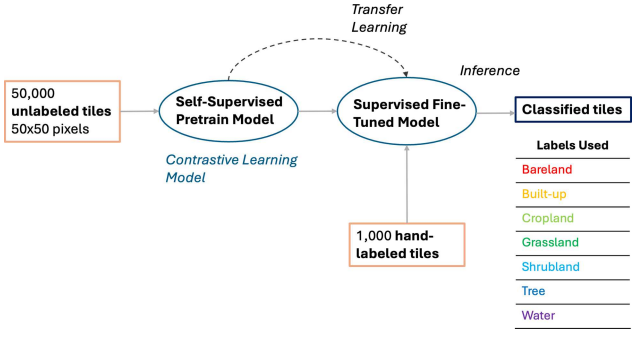
These considerations motivate the design choices in this paper: we adopt a compact operational legend aligned with widely used global products and focus on an output that supports reporting at the district scale. This does not replace pixel-wise products; rather, it offers a pragmatic alternative when pixel-accurate labeling and validation are infeasible and when administrative-level indicators are the primary deliverable.

4. Methodology, Data, and Results

We propose a cost-efficient operational pipeline that combines SimCLR-style contrastive pretraining on unlabeled Sentinel-2 tiles with lightweight supervised fine-tuning to predict land-cover classes at the tile level and aggregate them into district-level statistics for administrative reporting. The full pipeline—from imagery selection and quality screening to tiling, pretraining, fine-tuning, inference, and district aggregation—is summarized in Figure 1. While the broader program includes pilots in Yemen and Syria, the quantitative experiments reported here are for Lesotho; Yemen and Syria contribute only auxiliary unlabeled tiles used to increase diversity in appearance during the self-supervised stage and are not part of supervised training or evaluation.

A key design choice is that the primary deliverable is an administrative-scale indicator: a district-level land-cover class distribution derived by aggregating tile predictions. This shifts the emphasis from pixel-accurate boundary mapping to stable, interpretable reporting units, which is often the dominant requirement in data-scarce monitoring workflows. For transparency and reproducibility, we therefore describe below (i) the three-stage learning workflow, (ii) the data and controls used to reduce nuisance variability, and (iii) the evaluation strategy used to assess both tile-level predictive quality and district-level coherence against widely used reference products.

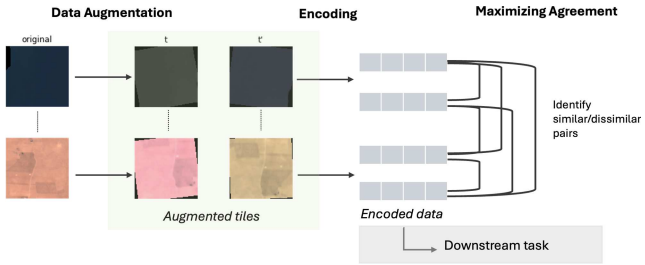
Figure 1
Explanatory diagram of the overall workflow



4.1. Methodology

The workflow has three stages: (i) contrastive pretraining with SimCLR using unlabeled tiles, (ii) supervised fine-tuning of a ResNet classifier using a compact labeled dataset, and (iii) inference over the full area of interest followed by district aggregation. The SimCLR training process is illustrated in Figure 2, and the end-to-end workflow is depicted in Figure 1.

Figure 2
SimCLR phases



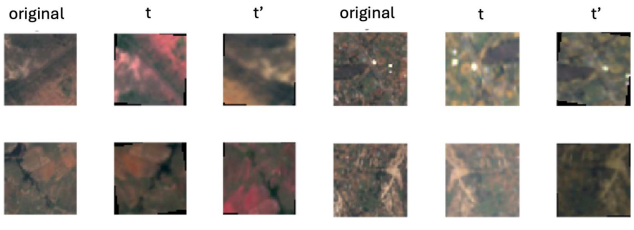
4.1.1. Contrastive pretraining

Input and sampling. Sentinel-2 Level-2A imagery is tiled into 50 × 50 RGB patches (RGB-only inputs). A tile-based representation provides a practical unit for both self-supervised representation learning and later supervised labeling under scarce ground truth. Tiles can contain heterogeneous land-cover signals (e.g., mosaics of cropland and natural vegetation), and the contrastive objective is intended to learn stable spectral–spatial cues that remain informative even when local boundaries are not perfectly resolved.

Sampling of unlabeled tiles. We use 50,000 unlabeled tiles for contrastive pretraining, sampled from the overall imagery range used in the Lesotho experiment (late December 2021 to February 2022). A subset of additional unlabeled tiles may originate from the Yemen and Syria pilots to increase diversity in unlabeled appearance; however, no labeled training or evaluation on those pilots is included in this paper. This design increases variability in the unlabeled corpus while preserving the experimental focus and reported results on Lesotho.

Augmentations and semantics preservation. For each tile, we generate two stochastically augmented views using conservative transformations designed to preserve land-cover semantics. The applied augmentations are shown in Figure 3 and include random cropping (up to 20%), mild brightness/contrast perturbations (up to 10%), and Gaussian blur. In EO, augmentation design is particularly sensitive because many land-cover classes differ through subtle spectral–textural cues and may vary seasonally. We therefore avoid aggressive perturbations that could remove discriminative information (e.g., strong color jitter or heavy geometric distortion) and instead apply a conservative policy that targets nuisance variability (illumination and mild blur) while keeping class semantics intact.

Figure 3
Data augmentation applied to tiles



Loss, positives, and negatives. We optimize the SimCLR NT-Xent objective, where each augmented pair of the same tile forms a positive pair and all other samples in the batch provide negatives. Negatives are sampled within the batch (in-batch negatives), and no memory bank is used, consistent with SimCLR's large-batch formulation. This training scheme encourages the encoder to learn representations that are invariant to the selected augmentations while remaining discriminative across different tiles.

Pretraining hyperparameters. We follow the original SimCLR parameterization: batch size 4096, temperature τ, and 100 epochs. Optimization uses Layer-wise Adaptive Rate Scaling (LARS) [20], a 10-epoch linear warmup, and cosine learning-rate decay. These choices prioritize a stable and reproducible training recipe that can be executed under practical compute constraints.

Hardware and runtime. Experiments are executed on Google Colab with an NVIDIA Tesla T4 GPU (16 GB). In our operational framing, contrastive pretraining is treated as an occasional offline step: once a pretrained encoder is available, supervised fine-tuning and inference/aggregation can be rerun rapidly to refresh district-level reporting outputs.

4.1.2. Supervised fine-tuning

Operational legend. The supervised model predicts a seven-class operational legend aligned with ESA WorldCover high-level categories: tree cover, shrubland, grassland, cropland, built-up, bare/sparse vegetation, and permanent water. This compact legend is selected to balance interpretability and learnability under limited labels and to support administrative reporting where high-level categories are often more useful than fine-grained classes that are difficult to validate.

Transfer protocol. Encoder weights obtained from contrastive pretraining are transferred to a supervised ResNet classifier [21] and fine-tuned using a small expert-labeled tile dataset. The encoder is fully trainable during fine-tuning (no staged freezing) to allow adaptation of representations to the target legend under limited labels.

Labeled set construction and class balance. The labeled tile set is assembled as a balanced subset across classes, which improves training stability in the low-label regime by reducing domination by majority classes during optimization. Importantly, this class balance affects only the supervised adaptation stage; the final district-level class proportions are derived from inference over the full area of interest and therefore reflect real spatial prevalence rather than the labeled sample distribution.

Further training details and regularization:

1) Training schedule: fine-tuning follows a fixed epoch budget consistent across backbone comparisons.
2) Class balancing/weights: the labeled tile set is assembled as a balanced subset across classes, so no additional class weighting is used.
3) Early stopping: early stopping is applied based on validation performance to limit overfitting in the low-label regime.
4) MixUp experiment: the ResNet34 + MixUp configuration reported in Table 1.
5) Uses MixUp [22] as the only change relative to plain ResNet34.

Table 1 reports three evaluation regimes: (i) an in-distribution random tile holdout (ID), (ii) a spatially disjoint tile holdout created via spatial blocking (Spatial Out-of-distribution

Table 1
Validation results across ResNet architectures. MU stands for MixUp regularization

Model	ID	Spatial OOD	Temporal OOD
RESNET18	0.93	0.56	0.36
RESNET34 + MU	0.83	0.18	0.22
RESNET50	0.95	0.96	0.96
RESNET152	0.84	0.45	0.6

(OOD)), and (iii) a date-disjoint holdout within the December 2021–February 2022 imagery window (Temporal OOD). These splits are designed to probe generalization under realistic deployment conditions while remaining feasible under limited labeled data.

4.1.3. Inference and district aggregation

After fine-tuning, the classifier predicts a land-cover class for each tile across the area of interest. Predictions are then aggregated into district-level class distributions for administrative reporting, consistent with the intended operational output depicted in Figure 1.

Aggregation transforms a set of tile-level categorical predictions into a compositional indicator per district by counting predicted classes within each administrative unit and normalizing to obtain proportions. This produces an interpretable summary of land-cover composition (e.g., percent cropland, percent grassland) and improves robustness by averaging out isolated tile errors and mixed-tile ambiguity. The resulting district-level class vectors also support coherence-style comparisons with reference products that are naturally defined at the pixel level but can be summarized into the same administrative units.

4.2. Data and experimental design

4.2.1. Geography and study period

The reported case study is the Kingdom of Lesotho (10 districts, with finer analysis possible at 80 constituencies). Sentinel-2 Level-2A [20, 23] imagery is taken from late December 2021 through February 2022 to maintain seasonal consistency within the experiment.

Restricting imagery to a narrow seasonal window serves two purposes. First, it reduces confounding variation from phenology and season-dependent appearance, which is particularly important for vegetation-related classes that can shift quickly over time. Second, it supports interpretability when comparing against external reference products, since temporal mismatch is a known source of disagreement across global land-cover datasets [5, 6]. In operational deployments, the same principle supports "like-with-like" reporting across years by encouraging seasonal comparability for repeated runs.

4.2.2. Cloud and seasonal controls

Cloud and quality screening uses Sentinel-2 Level-2A auxiliary layers (notably the Scene Classification Layer, SCL) [23] to filter observations before tiling. The restricted time range (Dec–Feb) reduces seasonal mismatch between the imagery used for pretraining, fine-tuning, and evaluation.

While masking substantially reduces cloud contamination, thin cirrus, haze, and shadows may persist and introduce noise in both contrastive pretraining and supervised fine-tuning. In practice, the district aggregation step mitigates the influence of isolated noisy tiles, since reporting indicators are computed over many tiles and are therefore less sensitive to sporadic artifacts unless contamination is systematic. For very cloudy regions, an operational extension is to broaden the acquisition window slightly to increase the number of clear observations while maintaining seasonal comparability.

4.2.3. Validation splits

The manuscript reports three validation regimes (ID, Spatial OOD, Temporal OOD) in Table 1. The ID regime is a random tile holdout from the same sampling distribution. "Spatial OOD" is implemented as a spatially disjoint holdout of tiles created via spatial blocking [24] within the same overall study area (i.e., tiles are separated geographically to reduce spatial autocorrelation between train and validation), and no district is excluded. "Temporal OOD" is implemented as a date-disjoint holdout within the same overall December 2021–February 2022 imagery window, testing robustness to acquisition-date variability rather than cross-season generalization.

4.3. Results

4.3.1. Backbone benchmarking

We benchmark four ResNet variants [21] (ResNet18, ResNet34 + MixUp, ResNet50, ResNet152) across the three validation regimes described above. The quantitative outcomes are reported in Table 1, and ResNet50 is selected because it shows the strongest and most consistent generalization across embedded, spatially disjoint, and temporally disjoint evaluations. This backbone choice is motivated by operational robustness: a model that maintains performance under both spatial disjointness and acquisition-date shifts is more likely to behave predictably when applied to new tiles within the same country or to refreshed imagery within the same seasonal window.

4.3.2. Tile-level classification metrics

We report standard classification metrics on held-out labeled tiles, including overall accuracy and macro-averaged precision/recall/F1. The reported performance (accuracy 0.91; macro-precision 0.89; macro-recall 0.87; macro-F1 0.88) indicates that the pretrained-then-fine-tuned model performs competitively under a small labeled budget. As expected for land-cover classification at 10 m resolution and with tile-based labels, most confusion concentrates in spectrally similar or transitional classes and in heterogeneous areas where multiple land-cover types can occur within a single tile. This behavior aligns with the broader motivation for administrative aggregation: when the objective is reporting at the district scale, local mixed-tile ambiguity can be averaged out to produce stable district indicators.

4.3.3. District-level coherence against reference products

To validate the administrative reporting objective, we compute a Coherence Index using cosine similarity between district-level class distributions from our pipeline and those derived from ESA WorldCover and Google Dynamic World [1, 2]. District-level coherence values are provided in Table 2, and class-wise coherence values are provided in Table 3.

Cosine similarity is used because it summarizes agreement in compositional structure: it measures whether two sources assign similar relative class proportions to the same district, which directly matches the reporting objective. This is especially useful when reference products differ in temporal support, labeling methodology, and legend implementation,

Table 2
Coherence index per district vs ESA and Google dynamic world

District	ESA	Google
Berea	0.994	0.568
Butha	0.999	0.742
Leribe	0.997	0.632
Mafeteng	0.911	0.297
Maseru	0.995	0.713
Mohale's Hoek	0.992	0.541
Mokhotlong	0.999	0.948
Quacha's Nek	0.998	0.971
Quthing	0.996	0.648
Thaba-Tseka	0.999	0.901

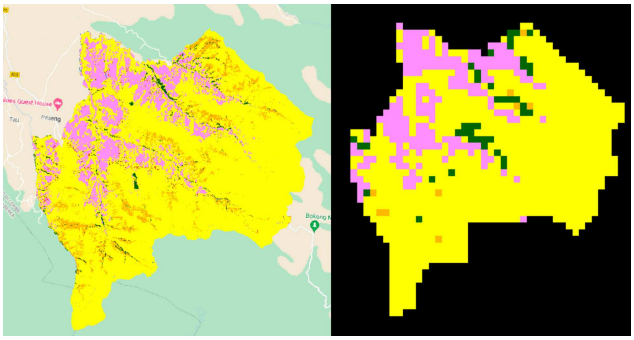
Table 3
Validation results across ResNet architectures

Class	ESA	Google
Built-up	0.975	0.947
Cropland	0.977	0.478
Tree	0.791	0.645
Shrubland	0.51	0.765
Grassland	0.996	0.951
Permanent water	0.801	0.485

because district-level comparisons reduce sensitivity to pixel-level boundary discrepancies.

Qualitative comparisons against ESA WorldCover for two example areas are shown in Figures 4 and 5, illustrating typical agreement patterns as well as the kinds of local discrepancies expected in mixed landscapes and along class boundaries.

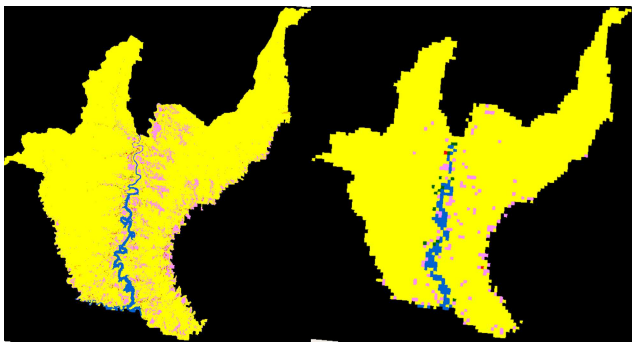
Figure 4
Bolahla, Lesotho. ESA v described method



4.3.4. Comments to results

Evidence from our experimental comparisons indicates that initializing the classifier with SimCLR-style contrastive pretraining improves downstream performance under the same labeled budget, supporting the use of conservative, semantics-preserving augmentations to learn transferable features under typical EO variability. Consistent with this, the selected backbone benefits from representation learning that is robust to the subtle inter-class boundaries and atmospheric/seasonal effects that commonly affect EO imagery.

Figure 5
Matsoku, Lesotho. ESA v described method



Across backbone variants, ResNet50 provides the most consistent generalization across the three evaluation regimes reported in Table 1, motivating its selection for subsequent inference. In our setting, "spatial" generalization refers to a spatially disjoint tile holdout created via spatial blocking within the same overall study area (not a district holdout), and "temporal" generalization refers to a date-disjoint holdout within the same overall late-December 2021 to February 2022 imagery window (not a different season). Shallower or more regularized configurations show less stable behavior across these regimes, suggesting that ResNet50 offers a favorable capacity–generalization trade-off for tile-level land-cover classification in this low-label context.

At the reporting scale of interest, aggregating tile predictions into district-level class distributions yields stable indicators that align closely with ESA WorldCover and provide an interpretable basis for comparison with Dynamic World. Remaining discrepancies are concentrated in mixed or transitional landscapes (e.g., shrubland/grassland boundaries) and may also reflect temporal asynchrony and probabilistic labeling in Dynamic World, which reinforces district-level coherence as a practical operational target for administrative reporting.

4.3.5. Operational considerations

All runs are executed on Google Colab using an NVIDIA Tesla T4 GPU (16 GB). In this setting, contrastive pretraining is the main offline cost, whereas fine-tuning, inference, and district aggregation provide a practical pathway to fast reporting updates once the encoder is pretrained. This separation is operationally relevant for data-scarce contexts: institutions can refresh the encoder occasionally (or reuse a previously trained encoder) and then update district-level reporting outputs more frequently with relatively low incremental compute and minimal additional labeling effort.

5. Discussion: Limitations and Conditions, Advantages, and Future Directions

This paper presents a practical deep learning approach for land-cover classification in data-scarce regions, using contrastive learning as an offline pretraining stage followed by supervised fine-tuning on a compact labeled set of Sentinel-2 tiles. The reported pilot case study in Lesotho demonstrates an end-to-end workflow that produces tile-level land-cover predictions and then aggregates them into district-level land-cover class distributions intended for administrative reporting. This framing emphasizes operational usability: the goal is not to compete with pixel-accurate global mapping products on boundary delineation but

to provide scalable, affordable, and quickly updateable land-cover indicators aligned with decision-making units [25].

A central implication of the tile-to-district design is that model utility is primarily assessed at the reporting scale. Tile-level metrics quantify whether the classifier can learn the compact legend under limited labels, while district-level coherence against external reference products evaluates whether aggregated class composition is plausible and consistent with widely used baselines. This dual evaluation is important in data-scarce settings where dense ground truth is rarely available and where pixel-level disagreement between products can reflect legend and temporal differences rather than purely model error.

Conditions and intended use. The approach is most appropriate when the decision unit is administrative reporting (district/constituency statistics) rather than pixel-accurate mapping. In such settings, stakeholders often require stable class proportions (e.g., cropland share, grassland share) that can be produced regularly and compared across time and regions, even if fine boundaries remain uncertain. The proposed workflow is therefore best interpreted as a reporting pipeline that generates compositional indicators, not as a segmentation system designed for parcel-level mapping or boundary extraction.

Practical deployment assumes (i) sufficient cloud-filtered Sentinel-2 observations over the target area during a comparable seasonal window and (ii) a compact but locally representative expert-labeled tile set aligned with the operational legend. Because the current implementation uses RGB-only inputs for contrastive pretraining and is trained/evaluated within a limited seasonal window, users should be cautious when extrapolating performance to different seasons or to areas where key classes are strongly phenology-dependent. In addition, since our "spatial OOD" validation is a spatially disjoint tile holdout within the same overall study area (spatial blocking; no district excluded), it should be interpreted as robustness to spatial autocorrelation rather than as a test on unseen administrative units.

Advantages. The main advantage of the method is that it reduces dependence on large labeled datasets by shifting most representation learning to contrastive pretraining on unlabeled imagery. Once a pretrained encoder is available, supervised fine-tuning requires only a small expert-labeled set and can be repeated with minimal incremental cost when the operational legend is updated or when limited new labels become available. This is aligned with typical constraints in data-scarce contexts, where labeling capacity is the primary bottleneck rather than data availability.

A second advantage is the speed and reproducibility of the workflow under constrained compute. In our implementation, runs are feasible on widely accessible resources (e.g., a single GPU session), which supports a realistic institutional workflow where pretraining is performed occasionally and reporting updates are generated more frequently. The tile-to-district aggregation further improves operational relevance by producing interpretable indicators at the same unit used for many national monitoring and humanitarian workflows, facilitating communication with non-technical stakeholders.

Finally, evaluating coherence against widely used reference products (ESA WorldCover and Google Dynamic World) provides an additional plausibility check at the reporting scale. This does not imply that reference products are ground truth; rather, it offers a practical triangulation strategy when independent validation data are limited, and it helps identify systematic discrepancies (e.g., in transitional vegetation classes) that may warrant targeted label collection or legend refinement.

Limitations. The tile-based design intentionally trades spatial precision for scalability and reporting alignment. Because a tile can include mixed land-cover signals, single-label supervision necessarily simplifies within-tile heterogeneity, and misclassifications are expected to concentrate along boundaries and in mosaicked landscapes. At 10 m resolution, fine structures (small settlements, narrow rivers, linear features) can be under-resolved within a 50 × 50 tile, which can propagate to confusion between built-up, bare/sparse vegetation, and vegetation classes in heterogeneous areas.

Agreement with global products should be interpreted carefully. Differences relative to Dynamic World and WorldCover can arise from temporal asynchrony (different acquisition periods), probabilistic labeling (Dynamic World probabilities versus categorical outputs), and legend/methodological differences. For this reason, district-level cosine similarity is best interpreted as coherence of aggregated indicators rather than as a measure of pixel-level correctness. Similarly, the reported "temporal OOD" regime is a date-disjoint holdout within the same December 2021–February 2022 window and does not directly quantify cross-season generalization, which remains an important challenge for robust operational deployment.

The current choice of RGB-only pretraining is also a limitation. While it supports conservative augmentation and a simple, reproducible recipe, it does not exploit the full multispectral richness of Sentinel-2 that is often useful for separating vegetation types, water, and bare soil under challenging atmospheric and seasonal conditions. This trade-off was intentional for operational simplicity in the pilot, but multispectral or index-augmented variants are likely to improve separability for difficult classes, especially shrubland/grassland transitions and sparse vegetation.

Future directions. Several extensions can strengthen the approach while maintaining the operational framing. A first direction is multi-season or multi-temporal exposure, where pretraining (and possibly fine-tuning) is performed on a stratified sample across seasons to improve robustness to phenology and to reduce sensitivity to the chosen seasonal window. Relatedly, incorporating time-series representations (e.g., short sequences of Sentinel-2 observations) could improve discrimination between classes that are spectrally similar on a single date but differ in seasonal dynamics.

A second direction is richer inputs during supervised adaptation, including additional Sentinel-2 bands and simple indices such as NDVI/EVI and water indices. This can be done while preserving the core idea of conservative augmentations and the tile-to-district operational output, but it requires careful augmentation design to avoid creating artifacts across bands. A third direction is active learning and label-efficiency improvements, where uncertain or high-disagreement tiles (e.g., shrub–grass boundaries, cropland mosaics) are prioritized for expert labeling to maximize gains per label. In addition, crop-specific ancillary datasets (e.g., CROPGRIDS [26, 27]) could be used to refine and sanity-check the cropland component of district-level indicators, especially in mosaic landscapes.

Finally, for institutional deployment, cloud-native execution (e.g., Earth Engine-based preprocessing with scalable inference) and standardized reporting outputs (district tables with versioned metadata) can make the pipeline easier to adopt and maintain. As these systems inform decisions, it is also important to strengthen communication and responsible-use practices, including clear statements about intended scale, temporal comparability constraints, and uncertainty summaries at the district level.

Acknowledgment

The authors are grateful to the FAO Data Lab and FAO Fisheries and Aquaculture Division (NFI) for the support provided to this work. They also thank ICAS (International Conference on Agriculture Statistics) and ESA (European Space Agency) EO for Africa Symposium for hosting related poster sessions in their conferences.

The views expressed in this publication are those of the authors only and do not necessarily reflect the views or policies of the Food and Agriculture Organization of the United Nations.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Gianfausto Bottini: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Francesco Nicola Tubiello:** Resources, Writing – review & editing, Supervision, Project administration. **Pengyu Hao:** Writing – review & editing.

References

[1] Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., . . . , & Tait, A. M. (2022). Dynamic World, Near real-time global 10 m land use land cover mapping. *Scientific Data*, *9*(1), 251. https://doi.org/10.1038/s41597-022-01307-4

[2] Zanaga, D., van de Kerchove, R., Daems, D., de Keersmaecker, W., Brockmann, C., Kirches, G., . . . , & Arino, O. (2022). *ESA WorldCover 10 m 2021 v200* [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.7254221

[3] Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., & Zhu, X. X. (2022). Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, *10*(4), 213–247. https://doi.org/10.1109/MGRS.2022.3198244

[4] Zhao, S., Tu, K., Ye, S., Tang, H., Hu, Y., & Xie, C. (2023). Land use and land cover classification meets deep learning: A review. *Sensors*, *23*(21), 8966. https://doi.org/10.3390/s23218966

[5] Xu, P., Tsendbazar, N.-E., Herold, M., de Bruin, S., Koopmans, M., Birch, T., . . . , & Zanaga, D. (2024). Comparative validation of recent 10 m-resolution global land cover maps. *Remote Sensing of Environment*, *311*, 114316. https://doi.org/10.1016/j.rse.2024.114316

[6] Food and Agriculture Organization of the United Nations (FAO). (n.d.). *Land Cover Classification System (LCCS)*. https://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1036361/

[7] Tubiello, F. N., Conchedda, G., Casse, L., Hao, P., Chen, Z., de Santis, G., . . . , & Muchoney, D. (2023). Measuring the world's cropland area. *Nature Food*, *4*(1), 30–32. https://doi.org/10.1038/s43016-022-00667-9

[8] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, 119*, 1597–1607.

[9] Tao, C., Qi, J., Guo, M., Zhu, Q., & Li, H. (2023). Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 5610426. https://doi.org/10.1109/TGRS.2023.3276853

[10] Kerdreux, T., Tuel, A., Febvre, Q., Mouche, A., & Chapron, B. (2025). Efficient self-supervised learning for earth observation via dynamic dataset curation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3008–3018. https://doi.org/10.1109/CVPRW67362.2025.00284

[11] Wang, Y., Braham, N. A. A., Xiong, Z., Liu, C., Albrecht, C. M., & Zhu, X. X. (2023). SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, *11*(3), 98–106. https://doi.org/10.1109/MGRS.2023.3281651

[12] Manas, O., Lacoste, A., Giró-i-Nieto, X., Vazquez, D., & Rodriguez, P. (2021). Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *2021 IEEE/CVF International Conference on Computer Vision*, 9394–9403. https://doi.org/10.1109/ICCV48922.2021.00928

[13] Chen, X., & He, K. (2021). Exploring simple Siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15745–15753. https://doi.org/10.1109/CVPR46437.2021.01549

[14] Zbontar, J., Li, J., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning, 139*, 12310–12320.

[15] Bardes, A., Ponce, J., & LeCun, Y. (2021). VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

[16] He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

[17] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision*, 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951

[18] Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T. (2022). Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*.

[19] Velazquez, D., López, P. R., Alonso, S., Gonfaus, J. M., Gonzalez, J., Richarte, G., . . . , & Lacoste, A. (2025). EarthView: A large scale remote sensing dataset for self-supervision. In *2025 IEEE/CVF Winter Conference on Applications of*

Computer Vision Workshops, 1138–1147. https://doi.org/10.1109/WACVW65960.2025.00136

[20] Ginsburg, B., Gitman, I., & You, Y. (2018). Large batch training of convolutional networks with layer-wise adaptive rate scaling. In *International Conference on Learning Representations*.

[21] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[22] Kim, J.-H., Choo, W., Jeong, H., & Song, H. O. (2021). Co-Mixup: Saliency guided joint mixup with supermodular diversity. In *International Conference on Learning Representations*.

[23] Drusch, M., del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., . . . , & Bargellini, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, *120*, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

[24] Eyster, H. N., & Beckage, B. (2024). Applying a deep learning pipeline to classify land cover from low-quality historical RGB imagery. *PeerJ Computer Science*, *10*, e2003. https://doi.org/10.7717/peerj-cs.2003

[25] Milà, C., Mateu, J., Pebesma, E., & Meyer, H. (2022). Nearest neighbour distance matching leave-one-out cross-validation for map validation. *Methods in Ecology and Evolution*, *13*(6), 1304–1316. https://doi.org/10.1111/2041-210X.13851

[26] Tang, F. H. M., Nguyen, T. H., Conchedda, G., Casse, L., Tubiello, F. N., & Maggi, F. (2024). CROPGRIDS: A global geo-referenced dataset of 173 crops. *Scientific Data*, *11*(1), 413. https://doi.org/10.1038/s41597-024-03247-7

[27] Bottini, G., Tubiello, F. N., Tang, F., Maggi, F., Conchedda, G., Casse, L., . . . , & Chen, Z. (2025). *Embedding CROPGRIDS into FAO geodata platforms* (FAO Statistics Working Paper Series: Issue 25–48). https://doi.org/10.4060/cd6325en

How to Cite: Bottini, G., Tubiello, F. N., & Hao, P. (2026). Contrastive Pretrain and Supervised Fine-Tune for Land-Cover Classification in Data-Scarce Countries. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62027813