



CLIP-RSICD: A Fine-Tuned Vision Transformer for Urban Land Cover Classification

CLIP-RSICD: Um Vision Transformer com Ajuste Fino para Classificação da Cobertura do Solo Urbano

Felinto Junior da Costa¹ , Rodrigo Rossetto Pescim² , Mariana Ragassi Urbano²

Received: January 7, 2026

Received in revised form: June 5, 2026

Accepted: July 29, 2026

Available online: August 28, 2026

ABSTRACT

Reliable mapping of intra-urban land-cover categories at fine spatial and thematic resolutions is fundamental to evidence-based spatial planning. Nevertheless, transferring classification models across territories remains difficult because urban morphology varies regionally. This challenge is pronounced in South American urban centers, where rapid expansion alters metropolitan dynamics, demanding finegrained land monitoring to support zoning policies, infrastructure distribution, and ecological protection. Addressing this issue, we evaluate a remote-sensing (RS) adapted Contrastive Language–Image Pre-Training (CLIP) visual encoder for multi-class urban land cover (ULC) categorization. The proposed taxonomy incorporates ten classes, spanning high-, medium-, and low-density built environments, industrial complexes, vegetative covers, and exposed soil. The high-resolution (HR) ULC products offer a rigorous empirical framework for urban morphological assessment and planning applications in intermediate Latin American cities. Using HR, open-access satellite imagery, the framework was implemented across the central urban cluster of the Londrina Metropolitan Region, Southern Brazil. The domain-adapted CLIP model reached an overall accuracy of 93%, outperforming nine baseline deep learning architectures. Our findings highlight the capacity of vision-language backbones to capture spatial patterns and extract discriminative representations from RS data, providing a scalable solution for urban spatial analysis.

keywords urban morphology, decision-support, spatial mapping, vision-language models

RESUMO

O mapeamento preciso das categorias de cobertura do solo urbano (CSU), com elevado detalhamento espacial e temático, é fundamental para o planejamento territorial baseado em evidências. Contudo, a generalização dos modelos de classificação é limitada pelas variações regionais da morfologia urbana. Esse cenário é marcante nos centros urbanos brasileiros, onde a expansão contínua reconfigura dinâmicas metropolitanas e intraurbanas, exigindo o monitoramento detalhado do solo para subsidiar o zoneamento, a distribuição de infraestrutura e a proteção ambiental. Diante disso, avalia-se o uso do codificador visual do modelo *Contrastive Language–Image Pre-Training* (CLIP), adaptado com imagens de sensoriamento remoto (SR), para a categorização multiclassee da CSU. O esquema taxonômico adota dez classes, abrangendo setores edificado de alta, média e baixa densidade, áreas industriais, áreas vegetadas e solo exposto. Os produtos de CSU de alta resolução (AR) constituem uma base empírica sólida para a caracterização morfológica e a análise aplicada ao planejamento em cidades de porte médio. A partir de imagens de SR de AR e livre acesso, a metodologia foi aplicada ao núcleo urbano central da Região Metropolitana de Londrina, Sul do Brasil. O modelo CLIP alcançou acurácia global de 93%, superando o desempenho de nove arquiteturas convencionais de aprendizado profundo. Os resultados evidenciam que codificadores pré-treinados em linguagem-visão conseguem extrair representações discriminativas de imagens de SR, configurando-se como uma ferramenta escalável para a análise territorial.

palavras-chave mapeamento espacial, modelos visão-linguagem, morfologia urbana, suporte à decisão

¹Prof. MSc. Dept. of Statistics, UEL, PR, Brazil. fjcosta@uel.br

²Prof. Dr. Dept. of Statistics, UEL, PR, Brazil. rrpescim@uel.br, mrurbano@uel.br

Introduction

Contemporary socioeconomic transformations find one of their most visible manifestations in the ongoing expansion of urban areas. The replacement of natural and semi-natural landscapes by human-made structures deeply alters ecological balances, infrastructure demands, and the overall spatial arrangement of cities (Weng et al., 2018; Zhu & Liang, 2020). Between 1990 and 2015, global built environments grew by nearly 30%, predominantly through horizontal sprawl in developing nations (Jochem et al., 2021; Lall et al., 2021). From a spatial management perspective, the primary concern extends beyond the sheer scale of physical expansion to understanding the precise spatial patterns and morphological manifestations through which urban surfaces evolve.

Urban dynamics involve a dual movement: outward peripheral growth and internal densification of existing cores (Clark, 1982). Each process exerts distinct pressures on city administration. While peripheral sprawl typically accelerates environmental degradation, habitat fragmentation, and exposure to natural hazards (Das et al., 2023), internal intensification places heightened stress on transport networks, energy grids, water distribution, and sanitation systems (Lu et al., 2023). In the South American context, particularly across Brazil, these dynamics are intensified by irregular growth patterns and historical structural inequalities, rendering detailed spatial data indispensable for territorial governance and land-use policy (Cerbaro et al., 2020; Instituto Brasileiro de Geografia e Estatística [IBGE], 2023; Sparovek et al., 2015).

Given that remote sensing (RS) directly records physical surface cover rather than functional land use, high-resolution (HR) intra-urban land cover (LC) mapping provides the foundational baseline for infrastructure planning, environmental assessment, and evidence-based policy implementation (Burley, 1961; Esch et al., 2018; Jensen & Cowen, 2011).

However, categorizing urban LC from HR satellite imagery remains a complex computational task. Urban environments exhibit high spatial heterogeneity, sharp spectral transitions, and strong visual similarities among distinct functional classes. For instance, low-density residential sectors, medium-density residential fabrics, industrial complexes, and bare land may display overlapping spectral and textural characteristics despite serving vastly different planning functions. Consequently, broad, low-resolution land inventories fail to satisfy the needs of local governance, which requires fine-grained distinctions among density gradients, industrial sectors, vegetative strata, exposed soil, agricultural fields, and hydrological features.

Traditional machine learning algorithms, such as Random Forest, Support Vector Machines, and Gradient Boosting, have long been applied to surface classification problems (Breiman, 2001; Cortes & Vapnik, 1995; Friedman, 2001). Nevertheless, their effectiveness degrades as remote sensing data scale up in spatial resolution, volume, and semantic complexity (Deren et al., 2014; Zhu et al., 2017; Zikopoulos et al., 2011). Deep learning (DL) architectures overcome these bottlenecks by extracting hierarchical feature representations directly from raw pixels, eliminating manual feature engineering and boosting accuracy in complex mapping tasks (Storie & Henry, 2018; Zhao et al., 2023). As a result, DL frameworks have become the benchmark standard for producing HR LC products (Hou et al., 2010; Souza, et al., 2020; Srivastava et al., 2010; Zhang et al., 2016).

Among advanced DL approaches, Vision Transformers (ViTs) stand out for their capacity to capture long-range contextual relationships rather than focusing solely on local pixel neighborhoods. In fine-resolution urban remote sensing, context is paramount because urban surface categories are defined not merely by localized textures or colors, but by the spatial organization and geometric interplay of buildings, canopy cover, open soil, roadways, and water bodies.

Contrastive Language–Image Pre-Training (CLIP) offers an effective transformer-based paradigm by learning joint visual-textual feature spaces from massive image-text pairs (Radford et al., 2021). Although originally trained on general web content, CLIP visual backbones can be tailored to remote sensing tasks through domain-specific fine-tuning. For example, adapting the model with the Remote Sensing Image Captioning Dataset (RSICD) shifts the learned representation space toward the semantic nuances of aerial and satellite imagery (Lu et al., 2018). While recent literature has evaluated parameter-efficient adaptation techniques like LoRA for general remote sensing tasks in Latin American urban contexts (Costa & Urbano, 2026), the application of domain-shifted CLIP embeddings for fine-grained density mapping in South American urban cores warrants dedicated investigation. This combination offers a framework specifically aligned with the structural complexity of intermediate Brazilian urban centers.

To address these needs, this study evaluates fine-grained multi-class LC mapping using a linear classifier stacked on feature embeddings extracted from a domain-adapted CLIP-RSICD image encoder. The empirical evaluation is conducted across the continuous urban cluster formed by Londrina, Cambé, and Ibiporã, the core metropolitan axis of Northern Paraná, Brazil. The spatial taxonomy comprises ten distinct land categories: Developed High Density, Developed Medium Density, Developed Low Density, Developed Industrial Areas, Bare Soil, Grass/Pasture, Shrubs/Scrubs, Row Crops, Perennial Trees, and Perennial Water. The primary objective is to determine whether features derived from a fine-tuned vision-language backbone can effectively distinguish intricate urban land surface configurations, providing a scalable tool for data-driven municipal management.

Additionally, the operational framework is taxonomy-agnostic: while demonstrated using ten urban surface classes, the workflow can be adapted to alternative or expanded classification schemes, provided representative training samples are available and classes retain visual distinctiveness in fine-resolution imagery.

Material and methods

Deep Learning in Computer Vision

Convolutional Neural Networks (CNNs) have long achieved remarkable success in remote sensing applications, particularly for urban scene understanding and land surface mapping (Alem & Kumar, 2022; Castelluccio et al., 2015; Penatti et al., 2015; Storie & Henry, 2018; Ulmas & Liiv, 2020; Zhang et al., 2018). Nonetheless, their intrinsic reliance on localized feature extraction constrains their capability to capture long-range spatial context, which has driven the rapid adoption of attention-driven network designs.

Originally introduced by Vaswani et al. (2017), transformer architectures have established state-of-the-art performance across numerous domains, including earth observation (Scheibenreif et al., 2022; Wang et al., 2024). The ViT extends self-attention mechanisms to image data by decomposing input arrays into non-overlapping patches and computing global context across all spatial regions, effectively balancing localized visual cues with macro-level spatial structures (Dosovitskiy et al., 2021). Recent literature highlights the efficacy of ViT-based models in LC mapping, including post-fire landscape assessment (Gonçalves et al., 2023), footprint extraction (Wang et al., 2022), and thematic land-use classification (Rangel et al., 2024), achieving accuracies near 99% on standardized benchmarks.

Vision-language foundation models, such as CLIP, mark a transformative shift by leveraging massive multimodal pre-training. CLIP's visual encoder, frequently built upon a ViT backbone, learns versatile feature representations through contrastive objective functions that align textual descriptions with visual inputs.

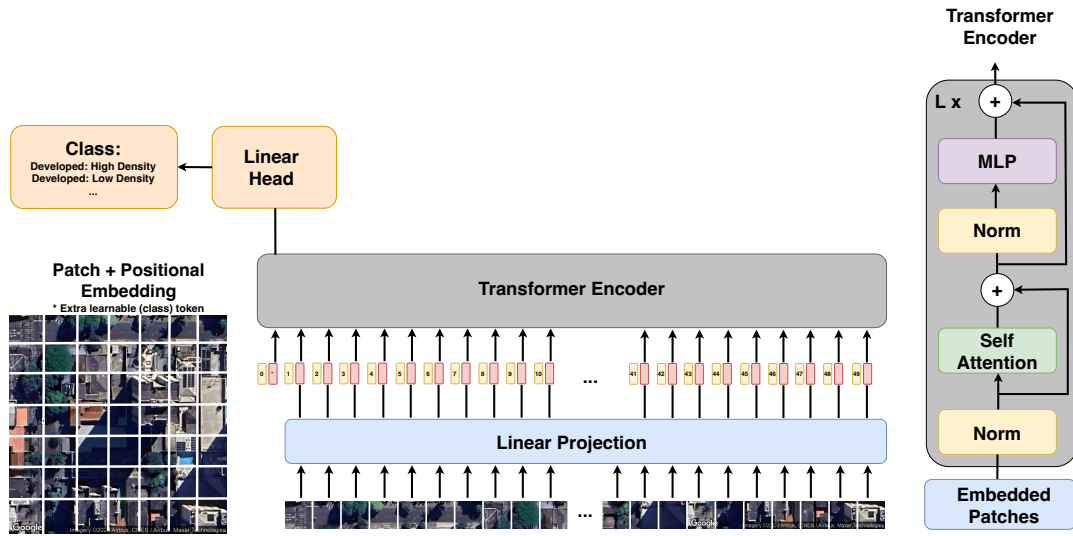
Adapting CLIP to specialized remote sensing datasets allows for effective cross-domain transfer learning (Qiu et al., 2020; Vali et al., 2020). However, evaluating domain-adapted CLIP visual encoders for HR, multi-class urban LC mapping in intermediate South American metropolitan settings remains insufficiently explored, a methodological gap this investigation seeks to bridge.

Vision Transformers

The ViT framework redefines computer vision processing by treating two-dimensional imagery as sequentially structured tokens (Figure 1). Given an input image array $X \in \mathbb{R}^{H \times W \times C}$, the spatial domain is decomposed into non-overlapping square patches of size $P \times P$. Each image patch is flattened and mapped via a linear projection layer into a D -dimensional feature space, forming an input sequence of $N = H \times W / P^2$ patch tokens.

To inject geometric context into the sequence, since attention operators are permutation-invariant, trainable one-dimensional positional encodings are summed with the patch representations. Furthermore, a learnable classification token (CLS) is prepended to the token sequence, serving as a global aggregator for image-level context (Devlin et al., 2019).

The resulting vector sequence is fed through a cascade of L identical Transformer encoder blocks. Each block incorporates Layer Normalization (LayerNorm), Multi-Head Self-Attention (MHSA), and a Multilayer Perceptron (MLP) module, configured with residual connections around each sub-layer.

Figure 1 - ViT architecture.

From "Efficient Specialization of Foundation Vision Models for Urban Land Cover Classification," by F. J. da Costa and M. R. Urbano, (2026), *Artificial Intelligence in Geosciences*.

Within the attention block, interaction weights across all spatial patches are derived by projecting input feature sequences into Query (**Q**), Key (**K**), and Value (**V**) matrices. Scaled dot-product attention is formulated as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V}, \quad (1)$$

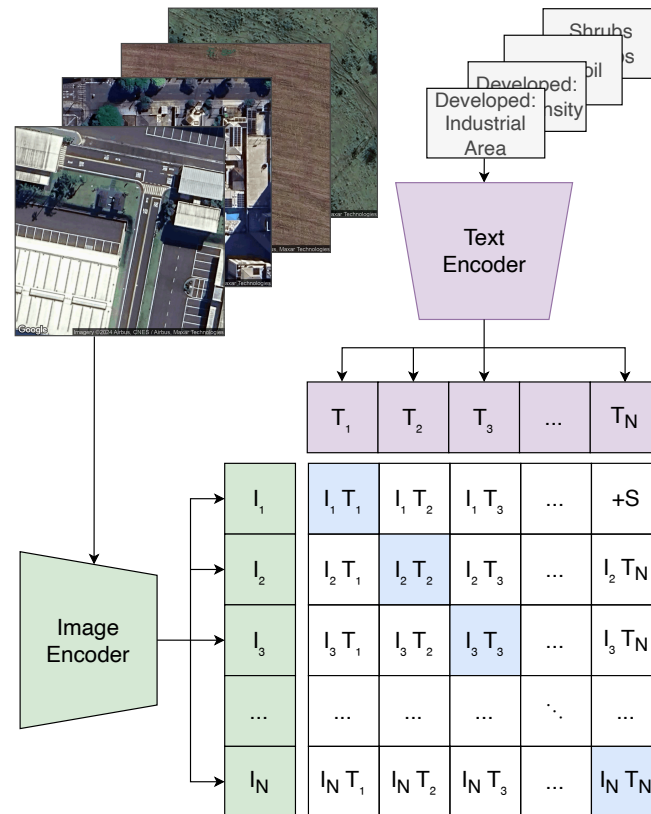
where d_k denotes the channel dimension of the keys. By distributing this operation across parallel attention heads, the network simultaneously models complex local interactions and long-range spatial dependencies. Ultimately, the contextualized representation corresponding to the *CLS* token is extracted from the final encoder block and passed to a linear classification head to output class probabilities.

Contrastive Language-Image Pre-training

The CLIP framework establishes a multimodal representation-learning paradigm by jointly training image and text encoders on approximately 400 million image-caption pairs to align their representations within a shared embedding space. Its dual-encoder architecture comprises a visual backbone, implemented as either a ViT or a Residual Network (ResNet), and a Transformer-based text encoder.

During training, the feature representations produced by both encoders are projected into the common embedding space and compared across each batch using cosine similarity. For a batch containing (N) matched image-text pairs, these pairwise comparisons form an ($N \times N$) similarity matrix, as illustrated in Figure 2. The diagonal elements correspond to the (N) correct image-text associations, whereas the off-diagonal elements represent mismatched combinations that act as in-batch negatives. A symmetric contrastive objective is then used to increase the similarity of corresponding image-text representations while decreasing their similarity relative to non-matching pairs. Through this learning mechanism, CLIP acquires a joint vision-language representation space in which semantically related visual and textual information is brought into close alignment.

Given a mini-batch containing N image-text tuples, the contrastive loss jointly optimizes bidirectional retrieval (image-to-text and text-to-image). This is achieved by maximizing the normalized cosine similarity $\text{sim}(I_i, T_i)$ of matched pairs while penalizing incorrect pairings, scaled by a learnable temperature parameter τ . This pre-training mechanism produces versatile and zero-shot transferable visual features. When fine-tuned on specialized domain corpora such as the RSICD, the visual encoder adapts its latent space to earth observation semantics, preserving high-level visual representation power while sharpening sensitivity to aerial and orbital imagery features.

Figure 2 - Architecture pipeline of the CLIP model.

Adapted from "Learning Transferable Visual Models From Natural Language Supervision," by A. Radford et al., (2021), *Proceedings of the 38th International Conference on Machine Learning*.

Fine-tuning for domain adaptation

Fine-tuning represents a core transfer learning strategy whereby models pre-trained on massive datasets are tailored to specialized downstream tasks. By initializing model parameters with general visual representations and optimizing them on target domain samples using attenuated learning rates, this procedure enables effective model adaptation while requiring significantly less task-specific training data (Vali et al., 2020).

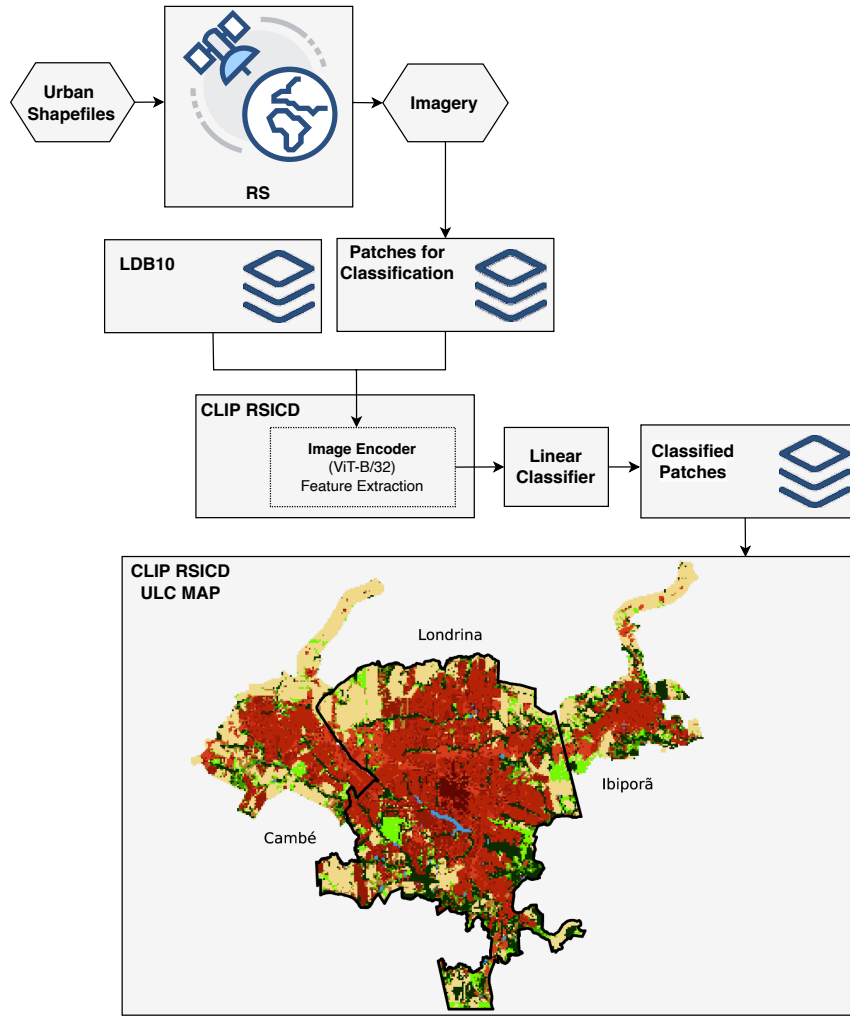
In earth observation, domain adaptation mitigates the challenge of constrained annotation availability by recycling feature spaces acquired from broad-scale image corpora, thereby reducing overfitting risks and boosting out-of-distribution performance (Vali et al., 2020). A notable instance is CLIP-RSICD, developed by Arutiunian et al. (2021), which adapts the original CLIP visual backbone specifically to remote sensing imagery semantics.

Building upon conventional fine-tuning paradigms, parameter-efficient adaptation strategies (PEFT), such as Low-Rank Adaptation (LoRA) (Hu et al., 2021), have recently gained prominence. These techniques drastically lower computational requirements and memory overhead by freezing pre-trained weights and inserting low-rank trainable matrices, achieving competitive performance relative to full-parameter optimization.

Experimental setup

General workflow

The operational pipeline developed for this research is depicted in Figure 3, outlining the sequence of processing steps from raw satellite data acquisition to the final thematic LC products.

Figure 3 - Workflow illustrating the proposed approach for imagery acquisition and ULC map production.

Imagery specification

Source HR optical data were retrieved using the GM Static API platform (Google LLC, 2024), which dynamically directs queries to appropriate imagery providers based on target spatial resolution and regional availability. This guarantees uniform visual properties, spatial fidelity, and complete workflow replicability across all sampled tiles. Imagery metadata corresponds to 2024 (Map data: Google; Imagery: Airbus, CNES/Airbus, Maxar Technologies), matching the download period executed in February 2024. Municipal boundaries for extracting prediction tiles were defined using official urban perimeter shapefiles sourced from the Paraná Interativo open spatial data portal.

Zoom level The spatial level of detail in tile retrieval is determined by the map tile zoom parameter. Zoom level (z) controls the rendering scale of geographic features within quadtree tile systems. Imagery tiles are structured hierarchically as 256×256 pixel arrays: at $z = 0$, a single tile spans the globe, with each integer increment partitioning the preceding tile into four sub-quadrants, yielding a $2^z \times 2^z$ global grid. The physical ground distance subtended by a tile at a given latitude is computed by

$$S_{tile} = C \cdot \cos(\text{lat}) / 2^z, \quad (2)$$

while the ground sample distance per pixel corresponds to

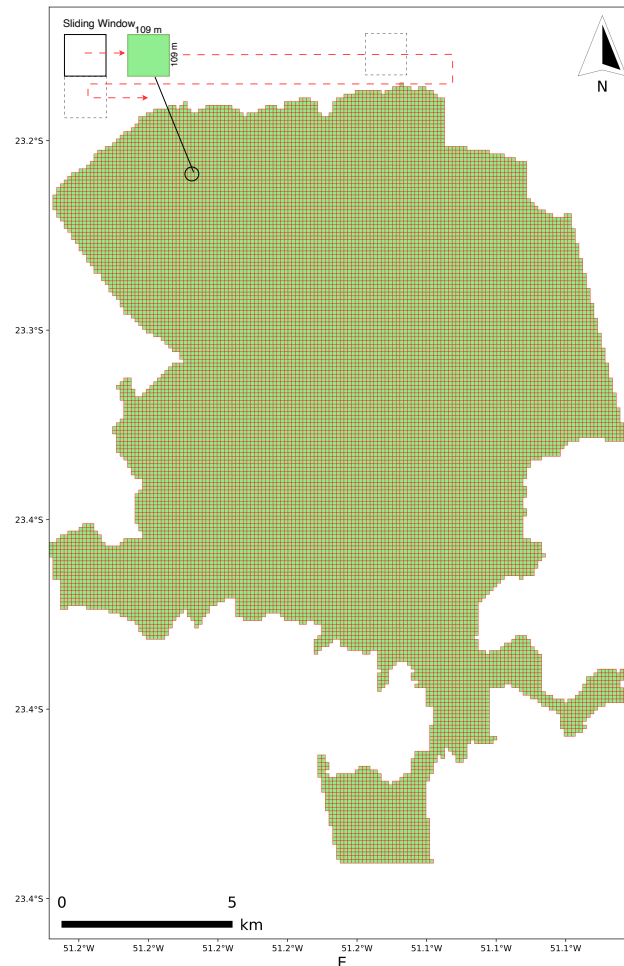
$$S_{pixel} = C \cdot \cos(\text{lat}) / 2^{z+8}, \quad (3)$$

where C represents Earth's equatorial circumference.

Equations (2) and (3) assume a spherical Earth approximation under the Web Mercator projection. Increasing zoom levels enhance spatial granularity at the expense of footprint area: at $z = 19$, ground resolution reaches approximately 0.30 m/pixel under standard web tile requests, which is sufficient to discriminate fine intra-urban structures. Consequently, $z = 19$ was selected for all image harvesting in this study.

Sliding Window Image patches were systematically sampled across the defined urban boundaries using a spatial grid traversal algorithm. To maintain constant physical window dimensions regardless of geographical position, spatial sampling was conducted in Universal Transverse Mercator (UTM) projected coordinates. The sliding mechanism initiated at the northwestern corner of each municipal perimeter, progressing horizontally before stepping downward row-by-row until full territorial coverage was achieved (Figure 4). To prevent sampling artifacts along municipal borders, a tile was retained for model inference only if the intersection between the window area and the official urban perimeter exceeded 75%. Adjacent windows were configured with zero overlap. Center coordinates for each valid patch were converted back to the WGS84 geographic coordinate system to execute API image queries at zoom level 19, returning 800×800 pixel image tiles covering an area of roughly 109×109 m at a nominal Ground Sampling Distance (GSD) of 0.13 m.

Figure 4 - Schematic representation of the spatial sliding window deployed over the urban boundary of Londrina, Paraná, Brazil.



Datasets

Training dataset Model optimization and evaluation were conducted using the LDB10 dataset, an annotated reference benchmark comprising 2,687 HR RGB image tiles distributed across ten urban land-cover classes within Londrina's urban perimeter.

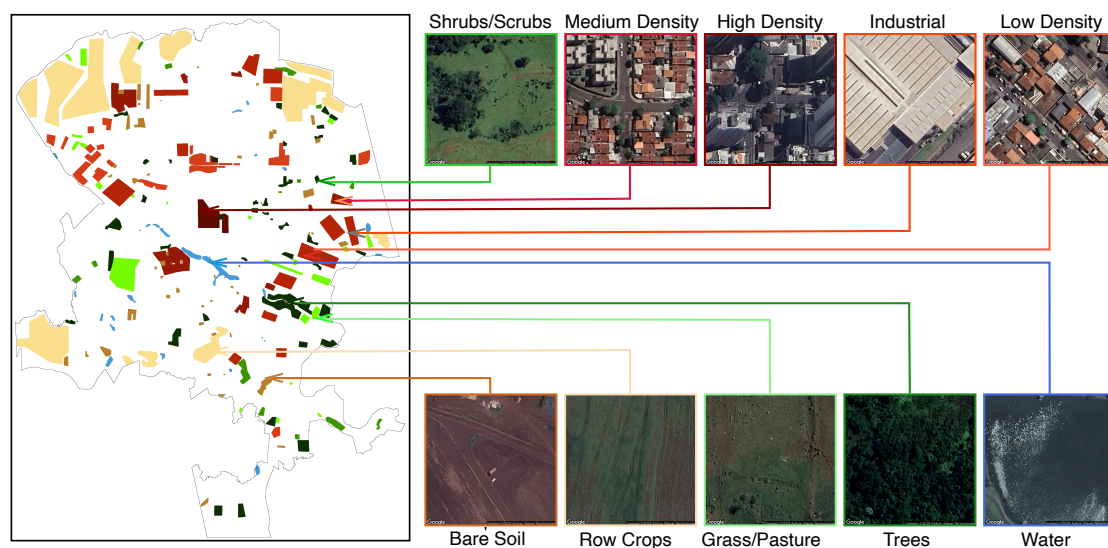
Reference polygon generation was implemented in JavaScript on the Google Earth Engine (GEE) cloud-computing platform (Gorelick et al., 2017), whereas the HR imagery used to generate the image patches was

obtained from Google Earth (GE). Ground-truth reference areas were established by manually delineating homogeneous land-cover polygons within Londrina's urban perimeter. Delineation was guided by visual interpretation of USGS Landsat 9 Level 2, Collection 2, Tier 1 composites acquired between January 2023 and June 2024 and displayed in GEE as the reference basemap.

Polygons were delineated at representative locations for each of the ten target land-cover classes, with their spatial distribution shown in Figure 5. Because thematic homogeneity is inherently scale- and class-dependent, polygon boundaries were defined according to the spatial configuration and structural characteristics of each class to maximize internal thematic consistency.

Image patches were subsequently extracted from these polygons under two spatial constraints: patches were required to be non-overlapping, and at least 75% of each patch footprint had to fall within the corresponding reference polygon. These criteria were imposed to limit class mixing along polygon boundaries while retaining representative within-class spatial and visual variability.

Figure 5 - Spatial distribution and examples of the reference polygons delineated within Londrina's urban perimeter. The left panel shows the spatial distribution of the polygons across the study area, whereas the right panel presents representative 800 × 800 pixel image patches from the LDB10 training dataset for the ten land-cover classes.



From “Efficient Specialization of Foundation Vision Models for Urban Land Cover Classification,” by F. J. da Costa and M. R. Urbano (2026), *Artificial Intelligence in Geosciences*.

The resulting LDB10 benchmark comprises 2,687 georeferenced HR RGB image patches covering a total area of 31.92 km², equivalent to approximately 14.59% of Londrina's urban perimeter as detailed in Table 1.

Table 1 - Land-cover classes and corresponding numbers of image samples in the LDB10 dataset.

Class	No. Samples	Class	No. Samples
Developed: High Density	78	Developed: Medium Density	195
Developed: Low Density	374	Developed: Industrial Areas	206
Bare Soil	74	Grass/Pasture	158
Shrubs/Scrubs	48	Row Crops	1,306
Perennial Trees	213	Perennial Water	35

Inference dataset Following the spatial sampling procedure previously outlined, image patches were continuously retrieved using a non-overlapping sliding window bounded by municipal urban perimeters. The resulting target inference corpus encompasses a total of 28,110 georeferenced HR RGB patches across the study area, distributed as follows: 18,302 tiles in Londrina, 5,905 in Cambé, and 3,903 in Ibiporã. Collectively, these tiles span an aggregate spatial surface of approximately 333.97 km², covering the continuous urban

limits of the three municipalities. This set of unlabeled imagery serves as the direct input for downstream spatial prediction.

Both spatial datasets comprise 800×800 pixel RGB image tiles, corresponding to a geographic footprint of roughly 109×109 m at a ground sampling distance (GSD) of 0.13 m, retrieved through the GM Static API at zoom level 19 (refer to Section Imagery specification). At the spatial resolution represented by the image patches, within-tile thematic homogeneity is consistent with the adopted classification scheme, while sufficient intra-class visual variability is retained to support feature learning across deep neural architectures (Chuvieco & Huete, 2009).

Feature extraction

Feature extraction was performed using the fine-tuned CLIP-RSICD vision backbone (Lu et al., 2018), which employs a ViT-B/32 architecture to convert image patches into 512-dimensional latent feature vectors. This embedding strategy was systematically applied to both the annotated LDB10 training dataset and the unlabelled inference imagery. By aggregating global spatial and semantic relationships through the *CLS* token, the encoder maps each input tile into a high-dimensional representation space where LC classes display enhanced separability. These extracted feature vectors subsequently served as inputs to train and deploy a linear classifier, assigning a discrete thematic label to every spatial patch.

Classifier training

A polytomous One-vs-Rest (OvR) logistic regression model was fitted on the 512-dimensional feature embeddings derived from the LDB10 benchmark dataset (refer to Section Datasets). For any given patch embedding x_i ($i = 1, \dots, N$), the classifier computes a class-conditional probability distribution $\mathbf{y}_{i,j} \in [0, 1]$ across all $j = 1, \dots, k$ target categories.

To mitigate performance degradation caused by inter-class sample disparities, a class-balancing undersampling threshold was implemented based on the mean class frequency across the training set. Specifically, overrepresented categories with sample volumes exceeding the dataset mean were randomly truncated to match this value, whereas underrepresented classes were retained in full. This strategy yields a more balanced target distribution, curbing majority-class bias and improving structural generalization.

Following inference, the classified spatial tiles were geographically reassembled using their centroid coordinate metadata, generating the continuous CLIP-RSICD urban LC product. This output delivers a seamless spatial mapping of surface categories across the contiguous urban limits of Londrina, Cambé, and Ibiporã.

Evaluated models

This investigation utilizes the visual encoder of CLIP-RSICD, a domain-adapted variant of the Contrastive Language, Image Pre-Training (CLIP) architecture optimized for earth observation tasks. The underlying network backbone relies on a Vision Transformer (ViT-B/32) configuration.

Developed by Arutiunian et al. (2021), CLIP-RSICD was fine-tuned on the Remote Sensing Image Captioning Dataset (RSICD) (Lu et al., 2018). The RSICD corpus consists of 10,921 remote sensing scenes (224×224 pixels) harvested from platforms including Google Earth, *Baidu Map*, MapABC, and Tianditu. Each scene is annotated with five descriptive captions across 30 distinct land categories characterized by high intra-class variance and subtle inter-class boundaries. The domain-adapted weights are accessible via the Hugging Face hub.

To rigorously assess the discriminative strength of the embeddings extracted by CLIP-RSICD, its performance was benchmarked against nine baseline deep learning architectures, as outlined in Table 2.

Restricting training sample acquisition strictly to Londrina's urban boundary serves as an experimental setup to test cross-territorial model generalization. This design evaluates whether learned visual representations can be transferred across adjacent municipalities sharing analogous morphological structures and growth dynamics, preserving spatial integrity across distinct urban jurisdictions.

Table 2 - Deep learning benchmark architectures evaluated for feature extraction capability.

Model Family	Architecture	Pre-training	Feature Vector Dimension
CLIP	ViT-B/32, ViT-B/16, ViT-L/14	ImageNet	512–768
RemoteCLIP	ViT-B/32, ViT-L/14, ResNet-50	RS image-text	512–1024
ResNet	ResNet-18, ResNet-50, ResNet-101	ImageNet	512–2048

Evaluation metrics

Evaluating model classification performance relies on four primary outcomes: True Positives (TP_i), False Positives (FP_i), True Negatives (TN_i), and False Negatives (FN_i). These outcomes serve as the building blocks for calculating core quantitative metrics, summarized in Table 3, including Overall Accuracy, Precision, Recall, and the F1-Score.

Table 3 - Quantitative performance metrics used to evaluate model classification outputs.

Metric	Description	Equation
Accuracy (A_i)	Ratio of correctly classified instances relative to total evaluations	$\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$
Accuracy (P_i)	Proportion of positive predictions that correctly correspond to class i	$\frac{TP_i}{(TP_i + FP_i)}$
Accuracy (R_i)	Ratio of true class i instances successfully identified by the model	$\frac{TP_i}{(TP_i + FN_i)}$
Accuracy (F1 _{i})	Harmonic mean providing a joint measure of precision and recall	$\frac{(2 \cdot P_i \cdot R_i)}{(P_i + R_i)}$

In multi-class prediction, overall accuracy can yield overly optimistic results when evaluating skewed class distributions, as performance on dominant categories can mask misclassifications in minority classes. Class-wise precision, recall, and F1-Score mitigate this bias by evaluating individual category performance independently. Precision assesses prediction fidelity, recall measures class-specific retrieval completeness, and the F1-Score harmonic balance reconciles trade-offs between both metrics.

To summarize performance across all categories, two aggregation mechanisms are employed. Macro averaging attributes uniform weight to each target class regardless of sample count:

$$[\text{metric}]_{\text{MACRO}} = \frac{1}{N} \sum_{i=1}^N [\text{metric}]_i, \quad (4)$$

whereas weighted averaging adjusts class contributions according to their respective sample frequencies:

$$[\text{metric}]_{\text{WEIGHTED}} = \frac{\sum_{i=1}^N (\text{support}_i \cdot [\text{metric}]_i)}{\sum_{i=1}^N \text{support}_i}, \quad (5)$$

where support_i represents the number of ground-truth observations belonging to category i . Macro averaging provides an unweighted measure of per-class fairness, while weighted averaging reflects operational mapping accuracy proportional to physical class prevalence (Sokolova & Lapalme, 2009).

Results and discussion

The experimental findings presented herein serve as a proof-of-concept for the proposed workflow, demonstrating an operational pipeline that functions independently of any specific LC classification scheme.

The ten spatial categories delineated within the LDB10 benchmark are employed strictly as an illustrative implementation; the underlying framework can be seamlessly extended to alternative, reduced, or expanded land surface taxonomies as required by specific spatial planning contexts. Consequently, the performance

metrics reported in the subsequent sub-sections validate the methodological robustness and transferability of the pipeline, rather than providing a exhaustive diagnosis of urban land dynamics across the evaluated territories.

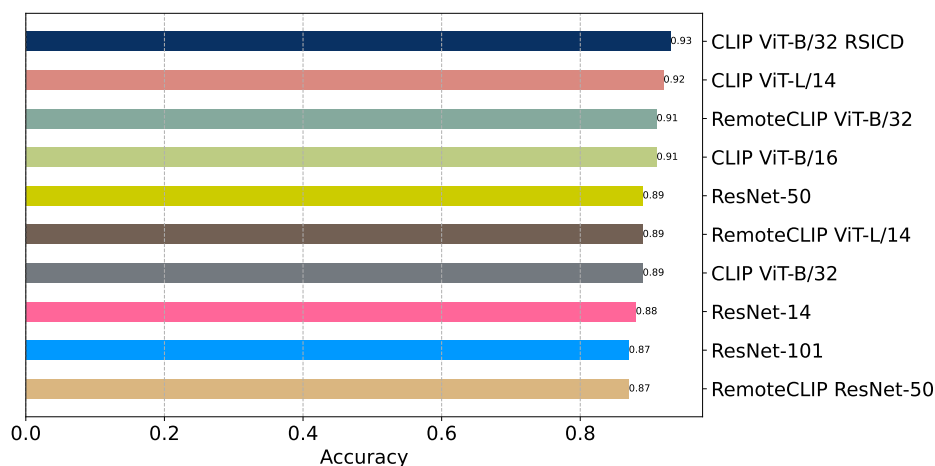
Classification performance

As emphasized in the systematic review by Truong et al. (2019), taxonomy cardinality plays a decisive role in shaping model predictive performance. Expanding the number of target classes generally increases intra-class spectral variance and semantic overlap, which frequently exerts a negative effect on global classification accuracy. This trend was documented across a meta-analysis of 64 remote sensing studies, which reported an average overall accuracy of 83.7% for systems operating across approximately ten land surface categories.

In line with the evaluation protocol, per-class performance was quantified using Precision, Recall, and F1-Score, computed independently relative to the sample support (i.e., total ground-truth patches per category) of each class.

The fine-tuned CLIP-RSICD model achieved a global accuracy of 93%, demonstrating strong capacity in discriminating urban LC classes and outperforming all nine baseline deep architectures evaluated. This top score surpasses the runner-up models, specifically CLIP ViT-L/14 and RemoteCLIP ViT-B/32, which both attained an overall accuracy of 92%, as illustrated in Figure 6.

Figure 6 - Overall classification accuracy across all evaluated deep learning models.



Nevertheless, while global accuracy offers a high-level benchmark of total correct predictions, it can obscure underlying misclassifications across individual land categories, particularly in imbalanced dataset settings. Highly frequent classes exert a disproportionate influence on global accuracy, potentially masking systematic errors within underrepresented categories.

To overcome this limitation, the F1-Score, the harmonic mean of Precision and Recall, serves as the primary evaluative metric for class-imbalanced scenarios. Unlike global accuracy, the F1-Score provides a balanced assessment that reconciles the trade, off between Precision and Recall, enabling a precise evaluation of model efficacy on low-frequency classes without statistical distortion from majority categories.

Aggregate performance across the target taxonomy was summarized using both macro and weighted averages. Macro averaging treats every class with equal priority regardless of sample size, whereas weighted averaging scales category metrics according to class support, reflecting real-world operational accuracy under class imbalance. Combined, these summary statistics provide a balanced analytical perspective for evaluating minority and majority surface types alike.

Integrating these granular class-wise metrics with overall accuracy, Table 4 provides a comprehensive synthesis of model performance, highlighting strong visual representations as well as specific categories requiring future refinement.

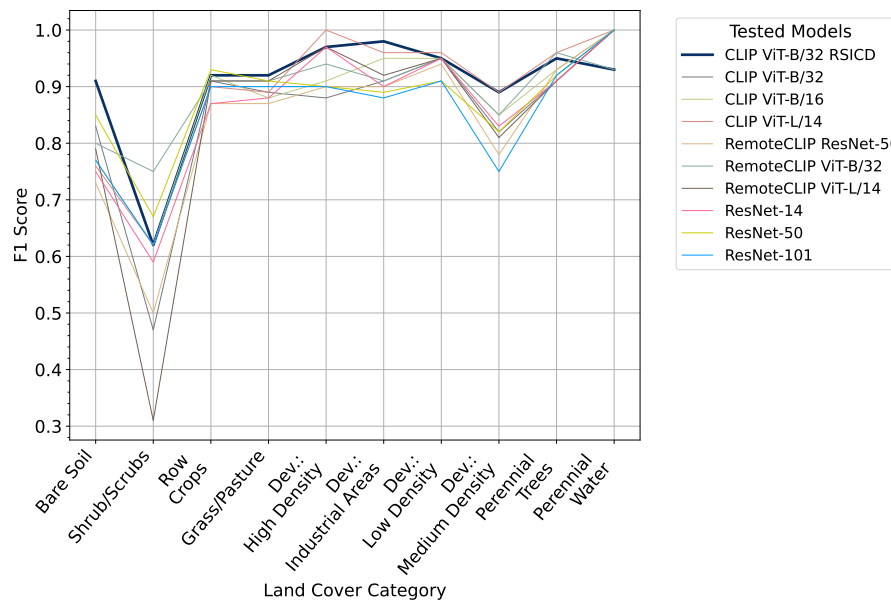
Table 4 - Detailed classification performance metrics for the domain-adapted CLIP ViT-B/32 RSICD visual backbone across the LDB10 benchmark.

LC category	Precision	Recall	F1-score	Support
Developed High Density	1.00	0.94	0.97	16
Developed Medium Density	0.92	0.87	0.89	39
Developed Low Density	0.95	0.96	0.95	54
Developed Industrial Areas	0.98	0.98	0.98	42
Bare Soil	0.83	1.00	0.91	15
Grass/Pasture	0.91	0.94	0.92	32
Shrubs/Scrubs	0.83	0.50	0.62	10
Row Crops	0.96	0.89	0.92	54
Perennial Trees	0.90	1.00	0.95	43
Perennial Water	0.88	1.00	0.93	7
Macro Average	0.91	0.91	0.91	312
Weighted Average	0.93	0.93	0.93	312

Comparative analysis of DL models

To evaluate the predictive capacity of CLIP-RSICD, its performance was benchmarked against nine representative baseline models, including standard CLIP variants, RemoteCLIP backbones (Liu et al., 2024), and ImageNet-pretrained ResNet architectures, as detailed in Section Evaluated models. Overall, CLIP-RSICD achieved the highest classification performance among all evaluated models.

More broadly, contrastive multimodal architectures consistently outperformed conventional supervised backbones in terms of overall accuracy, with domain-specific fine-tuning on the remote-sensing RSICD dataset yielding the largest performance gains. The class-wise F1-scores reported in Figure 7 further demonstrate that the advantage of CLIP-RSICD extends across the full land-cover taxonomy. Despite some variation in class-specific performance, partly associated with the imbalanced class distribution in LDB10, consistently high F1-scores were obtained for both majority and minority classes.

Figure 7 - Class-wise F1-Score breakdown for the evaluated benchmark architectures across all LC categories.

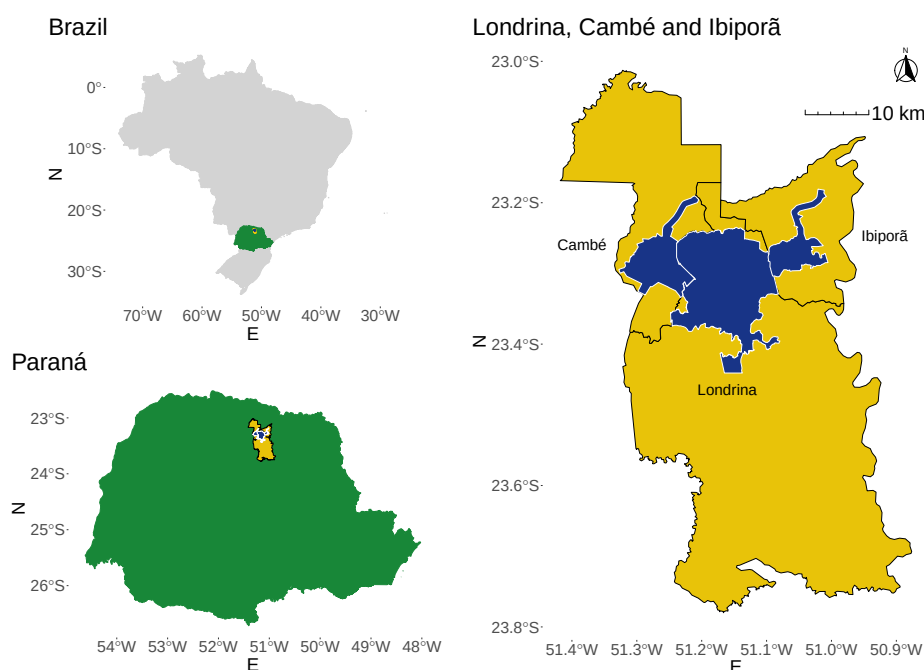
These results indicate that adapting contrastive language-image models to remote sensing semantics enhances feature representation for multi-class ULC mapping, providing an effective visual encoder for automated LC analysis and urban planning workflows.

Application region, ULC mapping

The state of Paraná, located in southern Brazil, holds substantial economic weight nationally. Official economic reporting places it fourth among Brazil's 26 states (Instituto Brasileiro de Geografia e Estatística [IBGE], 2024) in Gross Domestic Product (GDP), leading overall economic output across the southern region. Furthermore, eight of its 399 municipalities feature among the top 100 municipal economies nationwide. The state capital, Curitiba, occupies fifth place nationally, whereas Londrina stands in 36th position among Brazil's 5,570 municipalities (Instituto Brasileiro de Geografia e Estatística [IBGE], 2022).

To demonstrate model deployment and HR ULC map generation, the urban boundaries of Londrina, Cambé, and Ibiporã, situated in northern Paraná, were designated as the study area (Figure 8). Londrina, the core urban municipality, was established in the late 1920s as a planned agricultural settlement node. Following decades of rapid urbanization, it emerged by the 1970s as a key regional center for services, trade, and industry. Today, the municipality encompasses a total administrative territory of 1,652.57 km², including an official urban perimeter of 218.72 km² and an estimated population of approximately 556,000 residents as of 2022.

Figure 8 - Geographic context of the study site presented across three nested spatial scales: national boundary of Brazil highlighting Paraná state (top-left); state-level view showing the administrative boundaries of Londrina, Cambé, and Ibiporã (bottom-left); and detailed extent of the analyzed urban perimeters.

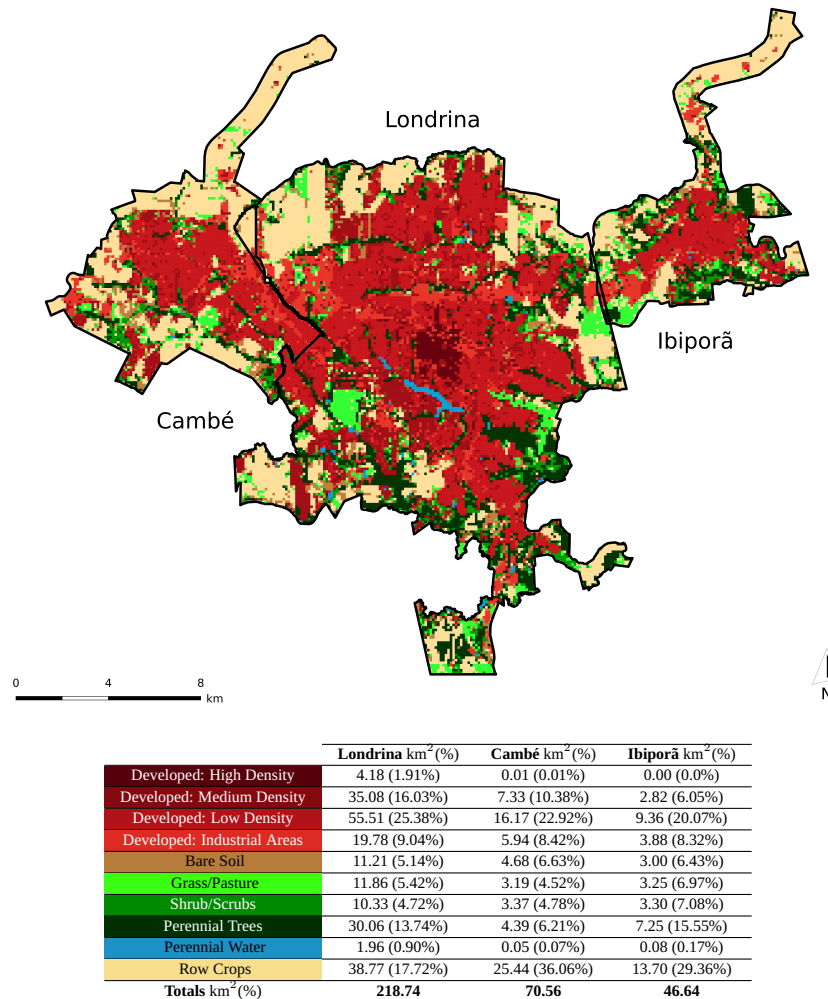


Adjacent to Londrina, the neighboring municipalities of Cambé and Ibiporã cover urban footprint areas of 70.57 km² and 46.62 km², with populations of roughly 107,000 and 52,000 inhabitants, respectively. Given their physical contiguity with Londrina, both municipalities maintain strong economic, functional, and demographic ties to the core city.

While these three administrative units do not constitute a single legally formalized metropolitan entity, their spatial continuity and socioeconomic integration justify evaluating them as an interdependent urban agglomeration. Combined, they comprise an urban territory of roughly 335.91 km² and an aggregate GDP of approximately US\$ 5.7 billion (as of 2021), illustrating the economic relevance of this urban cluster within northern Paraná.

Applying the CLIP-RSICD (ViT-B/32) visual backbone to the 28,110 inference patches yielded the HR ULC map displayed in Figure 9. This multi-class cartographic output for 2024 represents the primary product of the proposed pipeline, demonstrating its feasibility for generating continuous, municipal-scale thematic maps from optical satellite imagery.

Figure 9 - Spatially continuous multi-class ULC map generated for the study area (Londrina perimeter delineated in black) for the year 2024.



The ten-class taxonomy implemented here, particularly the fine-grained distinction between high, medium, and low-density built-up environments, delivers a higher spatial and thematic resolution than established global datasets. For instance, the CORINE Land Cover framework (Cole et al., 2022) aggregates built environments into just two broad classes (continuous and discontinuous urban fabric) with a minimum mapping unit (MMU) of 25 hectares, preventing fine intra-urban density assessments. Similarly, global products like Dynamic World (Brown et al., 2022) and MapBiomas (Souza et al., 2020) lack the localized density resolution required for granular urban planning. This underscores the advantage of custom taxonomies paired with localized adaptation strategies, which this workflow supports. Moreover, the extensive presence of agricultural row crops within official urban limits highlights the model's capacity to discriminate between visually similar cover types, such as row crops, bare soil, and pastures, under complex, heterogeneous land use conditions.

Conclusion

In contrast to traditional pixel- or spectral-based classification techniques, which often struggle under the pronounced intra-class heterogeneity and subtle inter-class boundaries characteristic of complex built environments, transformer-based vision encoders adapted to remote sensing semantics enable highly detailed multi-class urban LC mapping using standard RGB imagery alone.

This capability is supported by empirical validation: the fine-tuned CLIP-RSICD model reached a top overall accuracy of 93%, outperforming nine alternative deep learning baselines and markedly surpassing the benchmark average of 83.7% reported across comparable ten-class studies in the literature. These findings demonstrate that domain-specific fine-tuning plays a pivotal role in tailoring general-purpose vision foundation models to the spatial and morphological intricacies of urban landscapes.

The resulting framework demonstrated fine-grained mapping capabilities across built-up density gradients and industrial districts, moving beyond simplistic permeable/impermeable binary abstractions to deliver the spatial granularity required for urban planning applications.

Another key contribution is the taxonomy-agnostic design of the pipeline: while demonstrated using the ten categories of the LDB10 benchmark, the processing pipeline remains fully adaptable to custom, user-defined classification schemes. This flexibility directly overcomes a key limitation of macro-scale LC products, which aggregate urban surface features into coarse classes that fail to support intra-urban spatial assessment.

Furthermore, the methodology demonstrated strong cross-boundary spatial transferability: model parameters trained strictly on samples from Londrina yielded consistent and structurally sound LC maps across the adjacent municipalities of Cambé and Ibiporã. This transferability suggests that the learned feature representations capture underlying urban morphological regularities that generalize effectively across contiguous municipal systems.

Future work may extend this framework to multi-temporal LC change detection and evaluate its stability across alternate classification hierarchies. Additionally, coupling these spatial outputs with econometric models could offer valuable insights into how localized LC configurations influence property values and urban land economics.

Author Contributions

F. J. Costa conducted all aspects of this research. **R. R. Pescim** and **M. R. Urbano**: Conceptualization, Formal analysis, Methodology, Writing – review & editing.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability statement

The LDB10 dataset developed and evaluated in this work is publicly accessible via Figshare at <https://figshare.com/s/ed0e33906d0217f3f60d>. The acquisition and display of Google Maps imagery adhere strictly to the Google Maps Platform Terms of Service alongside the Google Brand Resource Center Geo Guidelines, which permit the reproduction of map tiles within academic publications provided appropriate source attribution is presented. Map data: Google. Imagery: Airbus, CNES / Airbus, Maxar Technologies.

References

- Alem, A., & Kumar, S. (2022). Transfer learning models for land cover and land use classification in remote sensing images. *Applied Artificial Intelligence*, 36(1), 2014192. <https://doi.org/10.1080/08839514.2021.2014192>
- Arutiunian, A., Vidhani, D., Venkatesh, G., Bhaskar, M., Ghosh, R., & Pal, S. (2021, October). Fine tuning CLIP with remote sensing (satellite) images and captions. *Hugging Face*. <https://huggingface.co/blog/fine-tune-clip-rsicc>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- 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., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., & Tait, A. M. (2022). Dynamic World, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9, 251. <https://doi.org/10.1038/s41597-022-01307-4>
- Burley, T. M. (1961). Land use and land utilization? *The Professional Geographer*, 13(6), 18–20. https://doi.org/10.1111/j.0033-0124.1961.136_18.x

- Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). Land use classification in remote sensing images by convolutional neural networks. [Preprint]. *Arxiv*. <https://doi.org/10.48550/arXiv.1508.00092>
- Cerbaro, M., Morse, S., Murphy, R., Lynch, J., & Griffiths, G. (2020). Information from earth observation for the management of sustainable land use and land cover in brazil: An analysis of user needs. *Sustainability*, 12(2), 489. <https://doi.org/10.3390/su12020489>
- Chuvieco, E., & Huete, A. (2009). *Fundamentals of satellite remote sensing*. CRC Press. <https://doi.org/10.1201/b18954>
- Clark, D. (1982). *Urban geography: An introductory guide*. Croom Helm.
- Cole, B., Smith, G., de la Barreda-Bautista, B., Hamer, A., Payne, M., Codd, T., Johnson, S. C. M., Chan, L. Y., & Balzter, H. (2022). Dynamic landscapes in the uk driven by pressures from energy production and forestry: Results of the corine land cover map 2018. *Land*, 11(2), 192. <https://doi.org/10.3390/land11020192>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Costa, F. J. d., & Urbano, M. R. (2026). Efficient specialization of foundation vision models for urban land cover classification. *Artificial Intelligence in Geosciences*, 7(2), 100225.
- Das, B., Khan, F., & Pir, M. (2023). Impact of urban sprawl on change of environment and consequences. *Environmental Science and Pollution Research*, 30(49), 106894–106897. <https://doi.org/10.1007/s11356-023-29192-3>
- Deren, L., Zhang, L., & Xia, G.-S. (2014). Automatic analysis and mining of remote sensing big data. *Acta Geodaetica et Cartographica Sinica*, 43(12), 1211. <https://doi.org/10.13485/j.cnki.11-2089.2014.0187>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019 [Proceedings]*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. [Preprint]. *ArXiv*. <https://doi.org/10.48550/arXiv.2010.11929>
- Esch, T., Heldens, W., & Hirner, A. (2018). The global urban footprint. In Q. Weng, D. Quattrochi, & P. E. Gamba (Eds.), *Urban remote sensing* (2nd, pp. 3–14). CRC Press, Taylor & Francis Group. <https://doi.org/10.1201/9781138586642>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Gonçalves, D. N., Junior, J. M., Carrilho, A. C., Acosta, P. R., Ramos, A. P. M., Gomes, F. D. G., Osco, L. P., da Rosa Oliveira, M., Martins, J. A. C., Júnior, G. A. D., de Araújo, M. S., Li, J., Roque, F., de Faria Peres, L., Gonçalves, W. N., & Libonati, R. (2023). Transformers for mapping burned areas in brazilian pantanal and amazon with planetscope imagery. *International Journal of Applied Earth Observation and Geoinformation*, 116, 103151. <https://doi.org/10.1016/j.jag.2022.103151>
- Google LLC. (2024). *Google Maps Static API*. <https://developers.google.com/maps/documentation/maps-static>

- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. [Preprint]. *ArXiv*, 1–26. <https://doi.org/10.48550/arXiv.2106.09685>
- Instituto Brasileiro de Geografia e Estatística. (2022). *Londrina: População*. <https://cidades.ibge.gov.br/brasil/pr/londrina/panorama>
- Instituto Brasileiro de Geografia e Estatística. (2024). *Em 2022, PIB Cresce em 24 Unidades da Federação*. <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/41893-em-2022-pib-cresce-em-24-unidades-da-federacao>
- Instituto Brasileiro de Geografia e Estatística. (2023). *Censo 2022 Indica que o Brasil Totaliza 203 Milhões de Habitantes*. <https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/22005-censo-2022-o-retrato-atualizado-do-brasil.html>
- Jensen, J. R., & Cowen, D. C. (2011). Remote sensing of urban/suburban infrastructure and socio-economic attributes. In M. Dodge, R. Kitchin, & C. Perkins (Eds.), *The map reader* (pp. 153–163). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9780470979587.ch22>
- Jochem, W. C., Leasure, D. R., Pannell, O., Chamberlain, H. R., Jones, P., & Tatem, A. J. (2021). Classifying settlement types from multi-scale spatial patterns of building footprints. *Environment and Planning B: Urban Analytics and City Science*, 48(5), 1161–1179. <https://doi.org/10.1177/2399808320921208>
- Lall, S. V., Lebrand, M. S. M., & Soppelsa, M. E. (2021). *The Evolution of City Form: Evidence from Satellite Data* (Policy Research Working Paper No. 9618). The World Bank. <https://doi.org/10.1596/1813-9450-9618>
- Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., & Zhou, J. (2024). RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–16. <https://doi.org/10.1109/TGRS.2024.3390838>
- Lu, H., Shang, Z., Ruan, Y., & Jiang, L. (2023). Study on urban expansion and population density changes based on the inverse s-shaped function. *Sustainability*, 15(13), 10464. <https://doi.org/10.3390/su151310464>
- Lu, X., Wang, B., Zheng, X., & Li, X. (2018). Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4), 2183–2195. <https://doi.org/10.1109/TGRS.2017.2776321>
- Penatti, O. A. B., Nogueira, K., & dos Santos, J. A. (2015). Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains? *Conference on Computer Vision and Pattern Recognition. In Institute of Electrical and Electronics Engineers, Conference on Computer Vision and Pattern Recognition. [Proceedings]. Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*. Boston, MA, USA, 44–51. <https://doi.org/10.1109/CVPRW.2015.7301382>
- Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10), 1872–1897. <https://doi.org/10.48550/arXiv.2003.08271>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models from Natural Language Supervision. *In International Conference on Machine Learning. [Proceedings]. 38th International Conference on Machine Learning*, 139, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>

- Rangel, A., Terven, J., Cordova-Esparza, D. M., & Chavez-Urbiola, E. A. (2024). Land Cover Image Classification. [Preprint]. *ArXiv*. <https://doi.org/10.48550/arXiv.2401.09607>
- Scheibenreif, L., Hanna, J., Mommert, M., & Borth, D. (2022). Self-supervised vision transformers for land cover segmentation and classification. In R. Chellappa. *Conference on Computer Vision and Pattern Recognition*. [Proceedings]. IEEE/CVF Conference on Computer and Pattern Recognition (CVPR) Workshops. New Orleans, Louisiana, 279–288. <https://doi.org/10.1109/CVPRW56347.2022.00036>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Souza, C. M., Jr., Shimbo, J. Z., Rosa, M. R., Parente, L. L., Alencar, A. A., Rudorff, B. F. T., Hasenack, H., Matsumoto, M., Ferreira, L. G., Souza-Filho, P. W. M., de Oliveira, S. W., Rocha, W. F., Fonseca, A. V., Marques, C. B., Diniz, C. G., Costa, D., Monteiro, D., Rosa, E. R., Eduardo, ... Azevedo, T. (2020). Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. *Remote Sensing*, 12(17), 2735. <https://doi.org/10.3390/rs12172735>
- Sparovek, G., Barretto, A., Matsumoto, M., & Berndes, G. (2015). Effects of governance on availability of land for agriculture and conservation in Brazil. *Environmental Science & Technology*, 49.
- Srivastava, P., Mukherjee, S., & Gupta, M. (2010). Impact of urbanization on land use/land cover change using remote sensing and GIS: A case study. *International Journal of Ecological Economics and Statistics*, 18, 106–117. <http://ceser.in/ceserp/index.php/ijeas/article/view/1894>
- Storie, C. D., & Henry, C. J. (2018). Deep Learning Neural Networks for Land Use Land Cover Mapping. In *International Geoscience and Remote Sensing Symposium*. [Proceedings]. IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium. Valencia, Spain., 6305–6308. <https://doi.org/10.1109/IGARSS.2018.8518619>
- Truong, V. T., Phan, D. C., Nasahara, K., & Tadono, T. (2019). How does land use/land cover map accuracy depend on the number of classification classes? *SOLA*, 15. <https://doi.org/10.2151/sola.2019-006>
- Ulmas, P., & Liiv, I. (2020). Segmentation of Satellite Imagery Using U-Net Models for Land Cover Classification. [Preprint]. *ArXiv*, 1–11. <https://doi.org/10.48550/arXiv.2003.02899>
- Vali, A., Comai, S., & Matteucci, M. (2020). Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. *Remote Sensing*, 12(15), 2495. <https://doi.org/10.3390/rs12152495>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, and S. Bengio, *International Conference on Neural Information Processing Systems*. [Proceedings]. 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, CA, USA., 6000–6010. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- Wang, L., Fang, S., Meng, X., & Li, R. (2022). Building extraction with vision transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11. <https://doi.org/10.1109/TGRS.2022.3186634>
- Wang, R., Ma, L., He, G., Johnson, B. A., Yan, Z., Chang, M., & Liang, Y. (2024). Transformers for remote sensing: A systematic review and analysis. *Sensors*, 24(11), 3495. <https://doi.org/10.3390/s24113495>
- Weng, Q., Quattrochi, D., & Gamba, P. E. (Eds.). (2018). *Urban remote sensing* (2nd. ed.). CRC Press, Taylor & Francis Group. <https://doi.org/10.1201/9781315166612>
- Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 22–40. <https://doi.org/10.1109/MGRS.2016.2540798>

- Zhang, P., Ke, Y., Zhang, Z., Wang, M., Li, P., & Zhang, S. (2018). Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery. *Sensors*, 18(11), 3717. <https://doi.org/10.3390/s18113717>
- 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>
- Zhu, X., & Liang, S. (2020). Urbanization: Monitoring and impact assessment. In S. Liang & J. Wang (Eds.), *Advanced remote sensing* (2nd, pp. 833–870). Academic Press.
- Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36. <https://doi.org/10.1109/MGRS.2017.2762307>
- Zikopoulos, P., Eaton, C., & IBM. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data* (1st). McGraw-Hill Osborne Media.