CIVICEMBED: FEATURE-SPECIFIC EMBEDDINGS FOR EFFICIENT GEOGRAPHIC REASONING AND RETRIEVAL

Josephine Wang
MIT
Cambridge, MA, USA
josiexw@mit.edu

Julien Coquet
ETH Zurich
Zurich, Switzerland
jcoquet@student.ethz.ch

Jeffrey Huang
EPFL
Lausanne, Switzerland
jeffrey.huang@epfl.ch

ABSTRACT

Diverse geography, from natural landscapes to urban areas, poses challenges for terrain-sensitive development. Existing geospatial embeddings are largely monolithic, limiting feature-specific comparison. We present CivicEmbed, a lightweight approach that learns separate embedding spaces for topography, water proximity, vegetation cover, and road density using self-supervised contrastive learning on thematic raster layers. These modular encoders support efficient analogical reasoning via retrieval: users can emphasize individual geographic features or weighted combinations of features to retrieve spatial analogs that match the selected constraints. Feature-specific encoders achieve $32\times$ compression (128-D vs. $4,096$-D raw patches) while improving retrieval metrics on certain features. We implemented a FAISS-backed retrieval system at the scale of Switzerland, providing a foundation for data-driven decisions in architecture, transit design, and land-use planning.

1 INTRODUCTION

Constrained by challenging geographies, urban planners often rely on analogical reasoning, examining other locations with similar geographic conditions to inform infrastructure design. Supporting this process requires a systematic way to quantify geographic similarity.

Recently, geospatial representation learning has made progress in capturing and predicting geographic features. Early work like Tile2Vec (Jean et al., 2018) demonstrated the potential of self-supervised approaches by learning spatial embeddings via triplet loss, showing that geographically proximate satellite tiles naturally map to similar representations. More recent efforts have expanded beyond image-based representations to incorporate structured geospatial data. S2Vec (Choudhury et al., 2026) introduced masked autoencoding on S2 cells, demonstrating that feature-based and image-based embeddings can be effectively fused. Similarly, the srai library (Gramacki et al., 2023) has worked to standardize geospatial representation learning, providing implementations of methods like Hex2Vec that encode OpenStreetMap features into region embeddings.

Existing geographic representations typically produce a single embedding that combines multiple factors. While effective for general-purpose representation learning, this design is less suited to geographic similarity search settings where users may wish to emphasize different spatial features. We treat geographic similarity search as a practical primitive for efficient spatial reasoning: given a query region and user-defined spatial constraints, the system should retrieve relevant spatial analogs while allowing flexible emphasis on features such as topography or road density.

In this paper, we propose a lightweight framework for quantifying geographic similarity through feature-specific embeddings that can be flexibly combined. We trained separate encoders for topography, water proximity, vegetation cover, and road density, each using a single thematic raster layer with self-supervised contrastive learning (Li et al., 2022).

This approach offers several advantages. **(i) Feature-aligned representations.** With each embedding corresponding to a distinct geographic layer, users can directly probe which aspects of two locations are similar. **(ii) Data-light training.** Our encoders learn from a single raster layer, making the method extensible to new features with minimal data collection and no labeled datasets.

(iii) Efficient compression. Our 128-D embeddings achieve $32\times$ compression over raw $4,096$-D patches, enabling scalable FAISS indexing while maintaining competitive retrieval performance. **(iv) Aligned with planning needs.** Separating geographic features matches how urban planners reason about sites (e.g., steep slope, forested, low accessibility), bridging modern representation learning with practical decision factors for urban development.

2 METHODS

2.1 CONSTRUCTING THE TOPOGRAPHY ENCODER

We train separate encoders for each thematic raster layer, so that each geographic feature is captured in its own embedding space. The process for collecting the GeoTIFF files for each feature is described in Appendix A.

We mapped each coordinate to an S2 cell at level 16 ($\sim 100 \times 100$ m) with s2sphere. For each cell, we extracted a 64×64 patch from the input GeoTIFF. Patches with missing values or read errors were discarded.

To train the topography encoder, we sampled triplets using precomputed great-circle distances: anchor-positive pairs were within $1 - 5$ km and negatives were > 10 km away with a patch cosine similarity ≤ 0.2 to ensure dissimilarity. The encoder used a ResNet-18 backbone with two heads: a 128-D L2-normalized projection for contrastive learning and a linear regressor predicting mean elevation. The encoder was trained for 50 epochs with a batch size of 32 using the Adam optimizer (learning rate = $1\mathrm{e} - 3$) and InfoNCE loss. At each epoch, we logged the average loss along with the mean cosine similarity between anchor-positive and anchor-negative embeddings.

2.2 CONSTRUCTING THE WATER, VEGETATION, AND ROAD DENSITY ENCODERS

The water proximity encoder was trained similarly on a water proximity raster scaled by $1/10$ (Pekel et al., 2016). Patches were extracted using the same S2 grid. A lightweight convolutional neural network (CNN) followed by global average pooling produced the 128-D embeddings. A regression head predicted the mean distance to water. The loss combined InfoNCE and MSE with $\lambda = 1.0$. More information on the different encoder architectures and ablation tests is in Appendix B. Triplets were sampled by binning cells into near (< 3 km), mid ($3 - 10$ km), and far (> 10 km). For these three encoders, anchors and positives came from the same bin, and negatives came from a different bin. Following similar approaches in self-supervised spatial learning (Jean et al., 2018), bin thresholds were chosen to balance sample distribution while ensuring semantic similarity within bins and distinctiveness across bins.

The vegetation cover encoder reused the same architecture and loss. Cells were binned by mean vegetation into low ($0 - 5$), mid ($5 - 40$), and high ($40 - 100$).

The road density encoder followed the same structure. Cells were binned into low (< 0.01), medium ($0.01 - 0.05$), and high (> 0.05) road density.

2.3 EVALUATION

We evaluated encoder performance on held-out S2 cells using three diagnostic categories: (i) linear probing regression, (ii) neighborhood retrieval, and (iii) Spearman correlation. We extracted patches using S2-aligned windows, holding out 20% of cells for testing via spatial block splits. The scalar target for each patch is the mean raster value. The linear probing fits a ridge regressor $\hat{y}_i = w^T e_i$ on training embeddings and then evaluates on test data.

For retrieval, we compute $K = 10$ nearest neighbors in embedding space (cosine similarity on ℓ_2-normalized embeddings) and measure (i) mean absolute target difference and (ii) rank-sensitive nDCG(inv)@10 using a continuous graded relevance (Järvelin & Kekäläinen, 2002). Spearman ρ measures monotonic alignment between embedding distances and $|y_i - y_j|$. Full metric definitions are in Appendix C.

We compared the encoder results against two baselines: (i) SatCLIP-ResNet18-L10, a pretrained location encoder that maps coordinates to embeddings via satellite imagery (Klemmer et al., 2024),

and (ii) Satellite RGB encoder, the same architecture as our topography encoder trained on Switzerland satellite imagery (Appendix C).

2.4 Visualizing the embedding space & top-K retrieval

For visualizing the embedding spaces, we projected embeddings sampled across Switzerland to 1-D via PCA (normalized to $[0, 1]$) and rendered them as color-coded grids. Figure 2 displays the general pipeline from input to embedding space visualization.

For top-K patch retrieval, we constructed 15 FAISS indexes using unit-normalized embeddings (Appendix H). At query time, top-K similar items were retrieved via inner product (i.e. cosine similarity).

3 Results

During training, all four encoders exhibited the expected learning dynamics: positive similarity increased, while loss and negative similarity decreased over time (Figure 3, Appendix D). The topography encoder had a less consistent decrease in negative similarity, possibly because geographically distant patches can share similar elevation profiles, making the negative samples in triplet training less distinct.

Table 1: Performance of feature-specific CivicEmbed encoders on linear probing and retrieval metrics. Confidence intervals are reported in Appendix E. The encoders Water, Vegetation, and Road refer to water proximity, vegetation cover, and road density, respectively.

Encoder	**Method**	R^2	**MAE**	**mean**$\lvert\Delta y\rvert$**@10**	**nDCG(inv)@10**	**Spearman** ρ
Topography	CivicEmbed	0.4001	238	2.19	0.9637	0.4784
	PCA-128	–	–	39.0	0.8537	0.8047
	Raw-4096	–	–	39.8	0.8503	0.7992
Water	CivicEmbed	0.6560	0.626	0.0078	0.9979	0.6514
	PCA-128	–	–	0.0471	0.9781	0.6681
	Raw-4096	–	–	0.0546	0.9742	0.6876
Vegetation	CivicEmbed	0.1954	17.6	2.43	0.8975	0.4925
	PCA-128	–	–	0.805	0.9238	0.9445
	Raw-4096	–	–	0.836	0.9240	0.9420
Road	CivicEmbed	0.7951	0.0594	0.0225	0.9835	0.2567
	PCA-128	–	–	0.0208	0.9846	0.7252
	Raw-4096	–	–	0.0244	0.9820	0.7218

To assess the value of learned compression, we compared our 128-D embeddings against two non-learned baselines: flattened 64×64 S2 patches (Raw-4096) and a 128-D PCA projection of those patches (PCA-128) (Appendix C). Raw-4096 and PCA-128 isolate the value of learned compression relative to direct patch matching, whereas the satellite RGB encoder and SatCLIP evaluate against more generic or monolithic learned representations.

Our compressed embeddings maintain similar nDCG(inv)@10 (Table 1). The mean$\lvert\Delta y\rvert$@10 is improved for topography, water proximity, and road density, but vegetation cover retrieval performs worse (2.43 vs. 0.836 raw-4096). One possible cause is improper triplet sampling, motivating alternative sampling strategies such as distance-based or learned similarity metrics.

The road density encoder's Spearman ρ is significantly lower than the Raw-4096 baseline (0.2567 vs. 0.7218) despite strong local retrieval performance. We attribute this to the fact that road networks can be sparse and structurally diverse: two patches with identical density can have different spatial arrangements (e.g., a grid vs. a single highway), so the embedding captures density but not more complex patterns. The underperformance of the vegetation cover and road density encoder in certain metrics motivates future work on structure-aware encoders. Overall, the learned embeddings better capture local neighborhood structure relevant for retrieval at the loss of exact feature values.

Compared with the satellite RGB encoder and SatCLIP, the feature-specific encoders generally achieve stronger task-aligned retrieval performance, particularly for topography and water proximity (Appendix F and G).

Table 2: Retrieval efficiency comparison between the 128-D CivicEmbed index and the 4096-D raw patch index over Switzerland. Full benchmarking details are in Appendix H.

	CivicEmbed	**Raw patch**
Dimensionality	128-D	4096-D
Index size	144 MB	855 MB
Mean latency	0.318 ms	5.030 ms

Beyond retrieval quality, the compressed embedding index yields a smaller FAISS index size and faster query time. On the full Switzerland-scale dataset, the 128-D CivicEmbed index is smaller on disk than the 4096-D raw patch index (144 MB vs. 855 MB) and supports faster nearest-neighbor search (0.318 ms vs. 5.030 ms mean latency) (Table 2). Thus, the 32$\times$ compression results in more efficient retrieval.

Finally, the color-coded 1-D PCA projections closely match the corresponding input GeoTIFFs (Figure 4), indicating that the single-feature embedding spaces preserve relevant information from each raster layer. These visualizations support feature-specific embeddings serving as a basis for geographic similarity search.

4 Discussion

Lower dimensionality predictably improves storage and query time, but our results also demonstrate that learned feature-specific compression can preserve or improve retrieval quality relative to raw-patch and PCA baselines for some modalities, specifically topography and water proximity. Thus, compact, modular embeddings can retain task-relevant structure while enabling efficient retrieval and comparison.

However, performance varies across features, highlighting limitations of the current approach. In particular, vegetation cover and road density encoders underperform the raw-patch baselines on some metrics, suggesting that the current models do not fully capture the structural complexity of these modalities. More structure-aware architectures, such as graph-based or multi-scale models, may better capture complex patterns. More broadly, our patch-based formulation operates at a fixed spatial scale, which limits its ability to represent larger geographic structures that extend across multiple patches.

Our experiments focus on Switzerland, whose geographic diversity makes it a useful testbed. However, the current study does not establish how well the learned embeddings transfer globally. Extending the framework to broader and more diverse geographies will be important for evaluating whether the learned feature spaces capture more general geographic patterns. Future work could also incorporate additional data sources, including outputs from large geospatial foundation models such as The AlphaEarth Foundations team (2025), as well as temporal inputs to capture dynamics such as seasonal change, snow cover, or urban expansion (Feng et al., 2025; Manas et al., 2021).

Beyond single patch retrieval, the learned representations also appear to support richer forms of geographic reasoning. In Appendix I, we demonstrate that a learned fusion MLP improves over simple averaging for combined multi-feature retrieval while maintaining near-zero residual correlations between features, suggesting that the shared fused space remains largely disentangled. In Appendix J, we further show that the same embeddings can support city-to-city similarity analysis. These results are preliminary, but they suggest that the modular representation may extend beyond patch-level nearest-neighbor retrieval toward more complex spatial reasoning tasks.

5 Conclusion

Our results demonstrate that our modular embedding approach captures geographic analogies. By supporting efficient similarity queries through compressed embeddings, this approach addresses a gap in geospatial AI. Such a tool enables planners and researchers to retrieve precedent locations, like regions with comparable topography for landslide mitigation or cities with similar layout and vegetation cover for design insights. Overall, this work moves toward geospatially aware search engines that compare complex terrains and support data-driven spatial planning.

REFERENCES

Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10161–10170. IEEE, October 2021. doi: 10.1109/iccv48922.2021.01002.

Shushman Choudhury, Chandrakumari Suvarna, Iveel Tsogsuren, Abdul Rahman Kreidieh, Elad Aharoni, Chun-Ta Lu, and Neha Arora. S2vec: Self-supervised geospatial embeddings for the built environment. *ACM Trans. Spatial Algorithms Syst.*, January 2026. ISSN 2374-0353. doi: 10.1145/3787217. URL https://doi.org/10.1145/3787217.

European Commission. Global surface water - data access. https://global-surface-water.appspot.com/download, 2024.

European Space Agency. Copernicus open access hub. https://browser.dataspace.copernicus.eu/, 2025.

Federal Office of Topography swisstopo. Swisstlm3d. https://www.swisstopo.admin.ch/de/landschaftsmodell-swisstlm3d#swissTLM3D---Download, March 2024.

Zhengpeng Feng, Zhecheng Xiong, Xin Wang, Wenwen Huang, Connor W. Coley, and Chenlin Chen. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. *arXiv preprint arXiv:2506.20380*, August 2025. URL https://arxiv.org/abs/2506.20380.

Piotr Gramacki, Kacper Leśniara, Kamil Raczycki, Szymon Woźniak, Marcin Przymus, and Piotr Szymański. Srai: Towards standardization of geospatial ai. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '23, pp. 43–52, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703485. doi: 10.1145/3615886.3627740. URL https://doi.org/10.1145/3615886.3627740.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data, 2018. URL https://arxiv.org/abs/1805.02855.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery, 2024. URL https://arxiv.org/abs/2311.17179.

Nathan Külling and Antoine Adde. Sweco25: Vegetation (vege). https://zenodo.org/records/10635551, February 2024.

Haifeng Li, Yi Li, Guo Zhang, Ruoyun Liu, Haozhe Huang, Qing Zhu, and Chao Tao. Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. ISSN 1558-0644. doi: 10.1109/tgrs.2022.3147513. URL http://dx.doi.org/10.1109/TGRS.2022.3147513.

Lukasmartinelli. Swissdem: Digital elevation model for switzerland from srtm (1 arc second / 25m) as download. https://github.com/lukasmartinelli/swissdem, 2024.

Oscar Manas, Alexandre Lacoste, Xavier Giro-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9394–9403. IEEE, October 2021. doi: 10.1109/iccv48922.2021.00928.

Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S. Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540:418–422, December 2016. doi: 10.1038/nature20584. URL https://doi.org/10.1038/nature20584.

D. Robinson, C. Baru, L. Band, C. Crosby, R. Devarakonda, V. Nandigam, T. Phan, and J. Schlief. Opentopography: A cyberinfrastructure facility for access and analysis of high-resolution topographic data. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, 2017. doi: 10.1145/3093338.3093375. URL https://doi.org/10.1145/3093338.3093375.

The AlphaEarth Foundations team. Alphaearth foundations helps map our planet in unprecedented detail. https://deepmind.google/discover/blog/alphaearth-foundations-helps-map-our-planet-in-unprecedented-detail/, July 2025.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. URL http://arxiv.org/abs/1311.2901.

APPENDIX

A DATA COLLECTION AND PROCESSING

Topographic data was obtained from a digital elevation map (DEM) of Switzerland without additional preprocessing (Lukasmartinelli, 2024).

Water coverage data was obtained from the European Commission (2024) Global Surface Water dataset. To calculate water proximity, we applied a tiled Euclidean distance transform (1024×1024px tiles with 128px overlap).

Vegetation cover data was derived by combining binary coniferous and deciduous tree cover maps for Switzerland into a single GeoTIFF without further processing (Külling & Adde, 2024).

Road density data was sourced from the Federal Office of Topography swisstopo (2024) dataset, converted from a shapefile to a 0.0001 degree-resolution GeoTIFF to preserve geometry. All final GeoTIFF files are shown in Figure 1 in Appendix L.

Satellite data was downloaded from the European Space Agency (2025) across 18 days in June to August 2025. True-color and scene classification maps were merged to minimize cloud cover, reducing the cloud cover percentage to 7.70%.

B ENCODER ARCHITECTURE

Architecturally, we made different design choices for the four geographic features. Topography benefited from a ResNet-18 backbone and distance-based triplet sampling, consistent with prior work on geographically structured contrastive learning (Ayush et al., 2021). In contrast, the water proximity, vegetation cover, and road density encoders performed better with a simple 3-layer CNN with InfoNCE+MSE training, although the strength of this choice varied by modality.

B.1 ABLATION RESULTS

To determine a suitable backbone for each feature-specific encoder, we compared a 3-layer CNN against a ResNet-18 backbone. For topography, a ResNet-18 backbone yielded consistent improvements across all evaluation metrics. For the water proximity and vegetation cover encoders, the ResNet-18 backbone improved some metrics but underperformed the 3-layer CNN on nDCG(inv)@10, our primary retrieval metric, so we selected the 3-layer CNN for these encoders. For the road density encoder, the ResNet-18 training exhibited signs of poor convergence. After 20 epochs, negative pair similarity remained mostly stagnant at 0.996, indicating insufficient separation between dissimilar patches and suggesting the model was unlikely to converge to a discriminative embedding space within the full training budget. Given this evidence of poor learning dynamics, we did not run the ResNet-18 road density encoder to completion and instead adopted the CNN backbone (Table 3).

Backbone performance appears to depend on the feature being modeled. One possible explanation is that elevation exhibits broader spatial gradients, which may have benefited from the deeper ResNet-18 in our setup. In contrast, water proximity, vegetation cover, and road density may be more sensitive to local patterns, which can be captured adequately by a shallow CNN (Zeiler & Fergus, 2013). We emphasize that this interpretation is a hypothesis rather than a controlled finding.

Additionally, we evaluated training a topography encoder with positives and negatives selected via binning rather than distance, as well as the InfoNCE + MSE loss, consistent with the approach used for the other three features (Table 3). The topography encoder with binned triplet sampling performed substantially worse than the distance-based sampling, particularly on retrieval metrics, suggesting that geographic distance provides a more informative notion of similarity for elevation. The InfoNCE + MSE loss also underperformed the InfoNCE loss alone, indicating that pure contrastive training with distance-based triplets was the more effective choice for topography.

Table 3: Selected ablation results for backbone and training design choices relative to the final encoders in Table 1. For topography, the 3-layer CNN substantially underperforms the selected ResNet-18 backbone (R^2: -0.031 vs. 0.400, nDCG(inv)@10: 0.098 vs. 0.964), supporting the use of a deeper architecture. For water proximity, ResNet-18 improves some metrics but underperforms the selected 3-layer CNN on nDCG(inv)@10 (0.995 vs. 0.998), our primary retrieval metric. For vegetation cover, ResNet-18 also degrades retrieval quality relative to the selected CNN (nDCG(inv)@10: 0.619 vs. 0.898). Alternative topography training strategies, including bin-based triplet sampling and InfoNCE + MSE, likewise underperform the selected distance-based InfoNCE configuration (nDCG(inv)@10: 0.411 and 0.826 vs. 0.964). We do not report a full ResNet-18 road density result because training showed poor convergence and was stopped before 50 epochs.

Encoder	R^2	**MAE**	**mean**$\lvert\Delta y\rvert$**@10**	**nDCG(inv)@10**	**Spearman** ρ
Topography (3-layer CNN)	-0.0311	358.9896	374.8761	0.0982	0.1486
Water proximity (ResNet-18)	0.9014	0.2833	0.0121	0.9949	0.5801
Vegetation cover (ResNet-18)	-0.3649	20.4318	8.6074	0.6192	0.4582
Topography (ResNet-18 + binning)	0.1051	317.3294	90.2430	0.4108	0.7402
Topography (ResNet-18 + MSE loss)	0.1122	352.0080	29.9433	0.8261	0.3672

C Evaluation metrics

C.1 Linear probing

For linear probing, each patch P_i is assigned a scalar target

$$y_i = \frac{1}{64 \times 64} \sum_{u,v} P_i(u, v), \tag{1}$$

which represents the average value of the feature over the patch (e.g., mean elevation, water proximity, vegetation cover, or road density). We then evaluate how well this information can be recovered from the learned embedding e_i by fitting a ridge regressor on the training set,

$$\hat{y}_i = w^\top e_i, \tag{2}$$

and reporting test-set R^2 and MAE:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}, \qquad \text{MAE} = \frac{1}{n} \sum_i |y_i - \hat{y}_i|. \tag{3}$$

Here, R^2 measures how well the embedding preserves feature information in a linearly recoverable form, while MAE measures the typical absolute prediction error in the original units of the feature.

C.2 Retrieval metrics

Because targets y_i are continuous, we additionally evaluate rank-sensitive retrieval with graded relevance using an inverse target-difference relevance function. For a query i and retrieved neighbor j, we define graded relevance as

$$\text{rel}(i, j) = \frac{1}{1 + |y_i - y_j|}. \tag{4}$$

This inverse relevance definition is used for our nDCG(inv)@10 metric. Given the ranked neighbor list $N_K(i) = \{j_1, \ldots, j_K\}$ obtained by cosine similarity on ℓ_2-normalized embeddings, we compute

$$\text{DCG@K}(i) = \sum_{r=1}^{K} \frac{\text{rel}(i, j_r)}{\log_2(r + 1)}. \tag{5}$$

We normalize by the ideal ordering (highest relevance first) over all candidates in the test set:

$$\mathrm{nDCG@K}(i) = \frac{\mathrm{DCG@K}(i)}{\mathrm{IDCG@K}(i)}, \tag{6}$$

where IDCG@K(i) is the DCG@K obtained by sorting all test candidates by decreasing $\mathrm{rel}(i, \cdot)$.

We also report mean$|\Delta y|$@K over test queries:

$$\text{mean}|\Delta y|\text{@K} = \frac{1}{n} \sum_i \frac{1}{K} \sum_{j \in N_K(i)} |y_i - y_j|. \tag{7}$$

Mean$|\Delta$y$|$@10 captures how close the retrieved neighbors are in actual feature value, while nDCG(inv)@10 measures whether the most relevant matches appear near the top of the ranked list.

C.3 Spearman rank correlation

Spearman rank correlation measures the correlation between embedding distances and the absolute difference in the target attribute. On a random subset of up to $4,000$ test points, we form the cosine distance matrix $D = 1 - \tilde{E}\tilde{E}^T$ with row-normalized embeddings $\tilde{E}$, and the absolute target-difference matrix $T_{i,j} = |y_i - y_j|$. We compute Spearman correlation ρ between the upper triangles of D and T. A high correlation indicates that larger embedding distances reliably correspond to larger target differences.

Spearman ρ measures monotonic alignment between embedding distances and $|y_i - y_j|$, indicating whether the geometry of the embedding space reflects meaningful differences in the underlying geographic attribute.

C.4 Baseline setups

C.4.1 Raw patch baseline

To assess the value of learned compression, we evaluate raw pixel-based retrieval as an additional baseline. Each 64×64 S2 patch is flattened into a $4,096$-D vector, standardized using training-set statistics, and L2-normalized before cosine similarity search. We report results for both the full $4,096$-D raw patch and a 128-D PCA projection, allowing comparison to CivicEmbed at both the original dimensionality and a matched compressed dimensionality. This baseline tests whether learned compression to 128-D provides benefits over direct pixel matching while quantifying the compression-performance tradeoff.

C.4.2 Satellite RGB encoder evaluation

We assessed the encoder trained on satellite imagery as a monolithic comparison. For each geographic feature, we retrieved the corresponding feature-specific patches and computed scalar targets as the mean raster value per patch (Equation 1). We applied the same spatial block split, linear probing regression (Equation 2), retrieval metrics (Appendix C), and Spearman correlation. This evaluation revealed the extent to which a general-purpose satellite encoder encodes task-relevant information, compared to feature-specific encoders trained on the thematic rasters.

C.4.3 SatCLIP evaluation

The SatCLIP embeddings are extracted via the ResNet18-L10 checkpoint. For each database point i with coordinates (λ_i, ϕ_i) (longitude, latitude), we compute a SatCLIP location embedding $s_i \in \mathbb{R}^d$ and ℓ_2-normalize it:

$$\tilde{s}_i = \frac{s_i}{||s_i||_2} \tag{8}$$

We applied the same spatial block split, linear probing regression (Equation 2), retrieval metrics (Appendix C), and Spearman correlation as for our feature-specific encoders and the satellite RGB encoder. For each feature, targets y_i are computed from the corresponding thematic raster patches,

and only the embedding function differs. The same embedding per coordinate is reused across all feature target evaluations.

We note that SatCLIP was pretrained on global imagery for general location encoding, while our encoders are task-specific and trained on Switzerland. This comparison demonstrates the value of feature-specific training.

D Encoder training curves

At their best epochs (Table 4), the topography, water proximity, vegetation cover, and road density encoders reached final losses of 0.0133, 0.0015, 0.0207, and 6.4116, with a significant difference between positive and negative similarity. The results show that the models learned discriminative, feature-specific spaces. The road density encoder's higher loss likely reflects difficulty minimizing the MSE term, as sparse and irregular road patterns are harder to reconstruct than smoother features like elevation or vegetation. However, the 1-D PCA visualization indicates the encoder can differentiate between high and low road density areas (Figure 4 in Appendix L).

Table 4: Training metrics at the best epoch for each encoder. All encoders achieve strong separation between positive and negative pair similarities, indicating that each encoder successfully learned a discriminative embedding space for its respective feature. The monolithic encoder refers to the satellite RGB encoder trained as a monolithic comparison.

Encoder	**Loss**	**Positive similarity**	**Negative similarity**
Topography	0.013	1.000	-0.447
Water proximity	0.002	0.963	-0.269
Vegetation cover	0.021	0.962	-0.780
Road density	6.412	0.999	-0.298
Monolithic	0.038	0.985	0.363

E Encoder performance with confidence intervals

We report bootstrap standard deviations ($\pm$) for all evaluation metrics (Table 5). We used $n = 500$ resamples for mean$|\Delta y|$@10, nDCG(inv)@10, and Spearman ρ, and $n = 200$ resamples for R^2 and MAE as linear probing variance converges faster than rank-based metrics.

We note that our test set is constructed via spatial block splitting of S2 level-16 cells, meaning test patches within the same block are spatially autocorrelated. Standard bootstrap resampling treats patches as independent, which may result in narrower confidence intervals. A block bootstrap would be more conservative but does not change the qualitative conclusions given the magnitude of the differences observed.

Table 5: Linear probing performance for CivicEmbed embeddings. R^2 and MAE are not reported for PCA-128 and Raw-4096 as these embeddings directly encode raw pixel values, making linear probing recovery of the patch mean trivial.

Encoder	R^2	**MAE**
Topography	0.400 ± 0.002	238 ± 1.2
Water proximity	0.656 ± 0.001	0.626 ± 0.006
Vegetation cover	0.195 ± 0.015	17.6 ± 0.3
Road density	0.795 ± 0.016	0.059 ± 0.001

E.1 Linear probing bootstrapping

For each bootstrap iteration, we resample the training set with replacement, refit a ridge regression probe with a fixed regularization strength α (selected via 5-fold cross-validation on the full training set using `RidgeCV`), and evaluate on the fixed held-out test set. R^2 and MAE are not reported for PCA-128 and Raw-4096 baselines as these embeddings directly encode raw pixel values, making linear recovery of the patch mean trivial ($R^2 \approx 1.0$).

Table 6: Retrieval metrics for all encoders and baselines. Column mean$|\Delta y|$@10 denotes mean absolute target difference among the top-10 retrieved neighbors. Column nDCG(inv)@10 uses inverse relevance weighting $1/(1 + |\Delta|)$.

Encoder	**Method**	**mean$\|\Delta y\|$@10**	**nDCG(inv)@10**	**Spearman ρ**
Topography	CivicEmbed	2.19 ± 0.27	0.964 ± 0.004	0.478 ± 0.009
	PCA-128	39.0 ± 2.2	0.854 ± 0.005	0.805 ± 0.005
	Raw-4096	39.8 ± 2.2	0.850 ± 0.005	0.799 ± 0.005
Water proximity	CivicEmbed	0.0078 ± 0.0006	0.998 ± 0.000	0.651 ± 0.010
	PCA-128	0.047 ± 0.003	0.978 ± 0.001	0.668 ± 0.012
	Raw-4096	0.055 ± 0.004	0.974 ± 0.001	0.688 ± 0.012
Vegetation cover	CivicEmbed	2.43 ± 0.10	0.898 ± 0.004	0.493 ± 0.015
	PCA-128	0.81 ± 0.04	0.924 ± 0.003	0.945 ± 0.003
	Raw-4096	0.84 ± 0.05	0.924 ± 0.003	0.942 ± 0.003
Road density	CivicEmbed	0.023 ± 0.001	0.984 ± 0.001	0.257 ± 0.010
	PCA-128	0.021 ± 0.001	0.985 ± 0.001	0.725 ± 0.010
	Raw-4096	0.024 ± 0.001	0.982 ± 0.001	0.722 ± 0.010

E.2 Retrieval metrics bootstrapping

For each bootstrap iteration, we resample test queries with replacement and recompute the mean nDCG and mean absolute target difference over the resampled queries, using precomputed nearest-neighbor indices from the full test set. The output estimates variance in the reported retrieval metrics across different test populations of the same size.

E.3 Spearman correlation bootstrapping

For each bootstrap iteration, we draw a random subset of test patches with replacement (up to $n = 4000$ to keep the $O(n^2)$ pairwise distance matrix tractable), compute cosine distances between their embeddings, and compute Spearman rank correlation between those distances and the absolute target differences $|y_i - y_j|$. We note that resampling with replacement for pairwise metrics is slightly conservative due to duplicate pairs appearing in the upper triangle.

F Satellite RGB encoder results

When evaluated on the same features (Table 7), the monolithic satellite RGB encoder shows mixed performance. It achieves moderately competitive retrieval on some features, particularly road density (nDCG(inv)@10 of 0.9602) and water proximity (0.9243), but performs worse on topography and vegetation cover. Its linear probing performance is consistently weak: only road density achieves a positive R^2 (0.0786), while topography, water proximity, and vegetation cover all yield negative R^2 values. In addition, Spearman correlation is near zero or negative for topography and vegetation cover, indicating that the embedding space does not consistently organize patches by the corresponding scalar feature.

These results suggest that while satellite RGB captures some broad spatial structure useful for local retrieval, it does not preserve feature-specific information in a linearly recoverable or globally aligned form. The findings support the advantage of feature-specific encoders when the goal is to predict or retrieve thematic raster information.

Table 7: Performance of the satellite RGB encoder evaluated on each geographic feature, using the same spatial block split and metrics as Table 1.

Feature	R^2	**MAE**	**mean$\|\Delta y\|$@10**	**nDCG(inv)@10**	**Spearman ρ**
Topography	-0.1618	372.3615	80.7233	0.8739	-0.0145
Water proximity	-0.5431	1.2096	0.2824	0.9243	0.5114
Vegetation cover	-0.1248	22.0859	5.4070	0.8253	-0.0010
Road density	0.0786	0.1198	0.0569	0.9602	0.1667

G SATCLIP RESULTS

SatCLIP shows uneven performance across features (Table 8). It is strongest on road density retrieval, where it achieves nDCG(inv)@10 of 0.9647, and it also attains relatively strong linear probing performance on topography ($R^2 = 0.4149$). However, it is significantly weaker on water proximity and vegetation cover, with near-zero R^2 on water and strongly negative R^2 on vegetation. Its low Spearman ρ on water proximity and road density features further indicates poor global alignment between embedding distances and target differences, even when retrieval quality is competitive.

These results suggest that SatCLIP captures broad geographic context useful for certain retrieval tasks, but does not consistently encode the feature-specific structure needed to predict or retrieve individual thematic raster information. This supports the advantage of feature-specific training for geographically-aware retrieval.

Table 8: Baseline comparison using SatCLIP location embeddings (ResNet18-L10). We evaluate the SatCLIP embeddings using the same held-out split and metrics as the feature-specific encoders.

Feature	R^2	**MAE**	**mean**$\lvert\Delta y\rvert$**@10**	**nDCG(inv)@10**	**Spearman** ρ
Topography	0.4149	269.6250	39.3860	0.8445	0.2657
Water proximity	0.0180	1.1361	0.2866	0.9072	0.0747
Vegetation cover	-0.8942	24.0653	3.5350	0.8473	0.4637
Road density	0.0298	0.1447	0.0460	0.9647	0.0298

H TOP-K RETRIEVAL COMPUTATION

We built 15 CivicEmbed retrieval indexes (4 single, 6 pairs, 4 triplets, 1 all-modality). Each CivicEmbed index uses IVF-PQ (`IndexIVFPQ` with `IndexFlatIP` quantizer) with $M = 8$, $\texttt{NBITS} = 8$, and $\texttt{nlist} = 4\sqrt{N}$ (capped at 32,768). Indexes were trained on $\sim$300,000 rows per modality with unit-normalized embeddings. Multi-modal indexes use equal-weight averaging of ℓ_2-normalized feature embeddings before training and querying. Retrieval uses inner product to implement cosine similarity. For efficiency comparison, we also built a separate raw-patch baseline index over flattened 4096-D patches covering the same geographic extent. Figure 5 in Appendix L visualizes the combined 1-D PCA projections with adjustable weights for user-defined feature importance.

H.1 RETRIEVAL EFFICIENCY

To quantify the retrieval efficiency gains from learned compression, we benchmarked query latency and index size for the 128-D CivicEmbed embedding index against the 4096-D raw patch index. Both indexes use FAISS IVF-PQ with inner product similarity and cover the same Switzerland-wide dataset at approximately 8.9M patches, with a small row-count difference due to patches skipped during raw raster extraction.

Query latency was measured using the benchmark script with $K = 10$ nearest neighbors and $n = 1000$ query vectors per index. Query vectors were reconstructed from each index's stored PQ codes to ensure realistic queries. Before measurement, we issued five untimed warm-up queries to each index. Single-query latency was measured by issuing one query at a time and recording wall-clock time via `time.perf_counter`. Batch throughput was measured with a batch size of 64. The `nprobe` parameter was set to $\sqrt{\text{nlist}}$ for both indexes. All benchmarks were run on a single CPU.

The 128-D embedding index is $5.9\times$ smaller on disk, $15.8\times$ faster per query, and supports $25\times$ higher batch throughput than the raw patch index. The latency advantage is primarily driven by the lower PQ scan cost at retrieval time: with 128-D embeddings, IVF-PQ distance computations involve 8 codebook lookups vs. 64 for 4096-D patches. These results demonstrate that the $32\times$ compression from learned embeddings translates directly into retrieval efficiency gains beyond the reduction in storage.

While the current retrieval method demonstrates the feasibility of feature-specific geographic search, it represents an initial implementation. Future work will investigate alternative methods for querying

Table 9: Retrieval efficiency comparison between the 128-D CivicEmbed embedding index and the 4096-D raw patch index. Both use FAISS IVF-PQ with identical `nprobe` $= \sqrt{\text{nlist}}$ and $K = 10$. The row count difference of 2.84% reflects patches skipped during rasterio extraction.

	CivicEmbed (128-D)	**Raw patch (4096-D)**
Vectors indexed	8,917,125	8,663,835
Index size	144 MB	855 MB
Bytes per vector	16	64
Mean latency	0.318 ms	5.030 ms
Batch throughput	14,431 qps	577 qps
Latency ratio	15.8× faster (CivicEmbed)	
Throughput ratio	25.0× higher (CivicEmbed)	
Index size ratio	5.9× smaller (CivicEmbed)	

areas that span multiple patches (i.e. cities) to better support diverse urban planning workflows. Additionally, consultation with urban development experts will provide more insight on how to define city-to-city similarity.

I FUSED EMBEDDINGS FOR NON-LINEAR RETRIEVAL

While the modular encoders support flexible weighted averaging of feature-specific embeddings at query time, simple averaging cannot model interactions between geographic features. We therefore trained a lightweight fusion MLP that learns a non-linear 128-D embedding from the concatenation of all four feature-specific encoders.

I.1 ARCHITECTURE

The fusion model takes as input the four 128-D feature embeddings (topography, water proximity, vegetation cover, road density) concatenated into a 512-D vector. A shared trunk maps this to a 128-D fused embedding via two fully-connected layers with layer normalization, ReLU activations, and dropout ($p = 0.1$).

The contrastive head ℓ_2-normalizes $\mathbf{h}$ for use in InfoNCE. Four independent linear regression heads predict the scalar target for each feature from the shared trunk output.

For notational compactness, we abbreviate the four features as topography (topo), water proximity (water), vegetation cover (vege), and road density (road).

$$\mathbf{h} = \mathrm{MLP}([\mathbf{e}_{\text{topo}} \| \mathbf{e}_{\text{water}} \| \mathbf{e}_{\text{vege}} \| \mathbf{e}_{\text{road}}]) \in \mathbb{R}^{128}. \tag{9}$$

$$\hat{y}_k = \mathbf{w}_k^\top \mathbf{h}, \quad k \in \{\text{topo, water, vege, road}\}. \tag{10}$$

I.2 TRAINING

The model was trained with a combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}}(\mathbf{e}_a, \mathbf{e}_p, \mathbf{e}_n) + \lambda \cdot \mathcal{L}_{\text{reg}}, \tag{11}$$

where $\lambda = 0.5$ and $\mathcal{L}_{\text{reg}}$ is a scale-invariant MSE that normalizes each feature's loss by its batch variance to prevent high-magnitude features (e.g., elevation in metres) from dominating low-magnitude ones (e.g., road density):

$$\mathcal{L}_{\text{reg}} = \frac{1}{4} \sum_{k=1}^{4} \frac{\mathrm{MSE}(\hat{y}_k, y_k)}{\mathrm{Var}(y_k)}. \tag{12}$$

Triplets were constructed using cosine similarity on the averaged baseline embedding (equal-weight average of the four ℓ_2-normalized feature embeddings) such that the MLP learns to improve upon simple averaging rather than replicate it. Anchors and positives were sampled from the $k = 50$ most

similar cells, and negatives from the $k = 50$ least similar. The model was trained for 100 epochs with Adam (learning rate $= 1\mathrm{e}{-3}$) and a cosine annealing schedule, using a batch size of 64. The best validation loss of 0.0325 was reached at epoch 89, with positive and negative pair similarities of 0.989 and 0.152 respectively, indicating strong contrastive separation (Figure 6).

I.3 Evaluation

We evaluated the fused embedding against the averaging baseline using the same held-out spatial block split and metrics as Table 1. Results are shown in Table 10. The fusion MLP outperforms the averaging baseline on linear probing R^2 across all features, with the largest gains on vegetation cover (0.7951 vs. -0.0165) and topography (0.9364 vs. 0.5214), demonstrating that non-linear fusion captures feature-specific structure that simple averaging discards. Retrieval quality (nDCG(inv)@10) also improves or remains competitive across all features.

We note that the R^2 values are not directly comparable to the feature-specific encoder results in Table 1. For each target, the fusion MLP receives the corresponding feature-specific embedding as part of its 512-D input, along with the other three embeddings, so it operates with strictly more input information than an encoder trained from a single raster modality. The results therefore show that non-linear fusion is more effective than simple averaging given access to all four learned feature spaces.

Table 10: Per-feature retrieval performance of the fusion MLP versus the equal-weight averaging baseline. The fusion MLP achieves higher R^2 across all features, with the largest gains on vegetation cover and topography. Retrieval quality (nDCG(inv)@10) is also improved or matched, and mean$|\Delta y|$@10 is lower for topography, vegetation cover, and road density.

Feature	**Method**	R^2	**MAE**	**mean$\lvert\Delta y\rvert$@10**	**nDCG(inv)@10**	**Spearman ρ**
Topography	Fusion	0.9364	0.0620	0.0304	0.9825	0.6350
	Avg	0.5214	0.2824	0.0415	0.9801	0.4318
Water proximity	Fusion	0.9763	0.0131	0.0124	0.9913	0.4422
	Avg	0.7413	0.0485	0.0097	0.9935	0.4443
Vegetation cover	Fusion	0.7951	6.0090	4.1805	0.8254	0.2995
	Avg	-0.0165	16.8122	5.3335	0.8163	0.0487
Road density	Fusion	0.9252	0.0304	0.0258	0.9812	0.1354
	Avg	0.6511	0.0756	0.0322	0.9772	0.1675

To assess feature disentanglement, we computed the Pearson correlation matrix between the linear probing regression residuals for all four features on the fused embedding (Table 11). All off-diagonal correlations are near zero (range: -0.072 to 0.049), indicating that the shared 128-D space encodes the four features independently rather than conflating them.

Table 11: Pearson correlation matrix between regression residuals of the four features on the fused embedding. Near-zero off-diagonal values indicate that there is minimal cross-feature entanglement.

	Topography	**Water proximity**	**Vegetation cover**	**Road density**
Topography	1.0000	-0.0717	-0.0163	0.0067
Water proximity	-0.0717	1.0000	-0.0157	0.0491
Vegetation cover	-0.0163	-0.0157	1.0000	0.0101
Road density	0.0067	0.0491	0.0101	1.0000

J Qualitative geographic similarity results

To demonstrate geographic similarity search at city scale, we compared 8 Swiss cities using Maximum Mean Discrepancy (MMD) with RBF kernels to measure distributional similarity across embedding spaces (Gretton et al., 2012). We sampled patches at 150 m intervals within eight Swiss city boundaries: Bern, Grindelwald, Lausanne, Lauterbrunnen, Montreux, Thun, Zermatt, and Zurich.

We used two complementary city-to-city similarity methods. The first method applied MMD separately to the feature-specific embedding spaces, along with a simple combined representation, to

show which geographic features make two cities appear similar. The second method applied MMD in the learned fused embedding space, which provided a single overall similarity score after non-linear integration of topography, water proximity, vegetation cover, and road density.

For the feature-specific analysis, we encoded patches with the trained models and computed pairwise RBF-kernel MMD directly in each embedding space with $\gamma = 20$. The MMD distance is converted to a similarity score via $\text{sim}(X, Y) = \exp(-\text{MMD}(X, Y))$. The combined score is computed as an equal-weight average across the four feature-specific similarities. Table 12 shows the top-5 matches for representative query cities.

Table 12: City-to-city similarity (top-5) using RBF-kernel MMD on embedding distributions ($\gamma = 20$). Higher values indicate greater similarity (scale 0-1). The combined score is an equal-weight average of the four feature-specific scores. The column headers Water, Vegetation, and Road refer to water proximity, vegetation cover, and road density, respectively.

Query	Similar City	Combined	Topography	Water	Vegetation	Road
Zurich	Lausanne	0.8506	0.9087	0.9076	0.7949	0.7911
	Thun	0.8296	0.8086	0.9095	0.7814	0.8191
	Bern	0.8241	0.9124	0.7611	0.9245	0.6985
	Montreux	0.8149	0.8661	0.9022	0.8392	0.6521
	Lauterbrunnen	0.6791	0.7571	0.9089	0.5842	0.4662
Lausanne	Thun	0.9192	0.8695	0.8931	1.0000	0.9140
	Montreux	0.9027	0.9564	0.9769	0.8922	0.7850
	Zurich	0.8506	0.9087	0.9076	0.7949	0.7911
	Grindelwald	0.7639	0.7876	0.8466	0.7442	0.6773
	Lauterbrunnen	0.7548	0.7846	0.8548	0.7974	0.5825
Zermatt	Grindelwald	0.8961	0.8902	0.9449	0.7826	0.9666
	Lauterbrunnen	0.8687	0.8802	0.9395	0.6969	0.9582
	Thun	0.7648	0.9783	0.8949	0.5777	0.6082
	Montreux	0.7206	0.8303	0.7967	0.5255	0.7298
	Lausanne	0.7046	0.8609	0.7968	0.5197	0.6408

These feature-specific scores reveal several interpretable patterns, while also highlighting some limitations of the current city-level similarity measure. Alpine cities such as Zermatt, Grindelwald, and Lauterbrunnen are grouped together, consistent with their shared steep terrain and relatively sparse road density. For example, Zermatt has especially high road similarity with Grindelwald (0.9666) and Lauterbrunnen (0.9582), suggesting that the road embedding captures the low road density. Lakeside cities such as Lausanne and Montreux also show high water similarity (0.9769), consistent with their shared position on Lake Geneva. Some scores are less intuitive, for instance, Zermatt and Lausanne receive a relatively high topography similarity score (0.8609) despite clear differences in overall landscape. This finding suggests that the learned feature spaces capture local patch-level regularities, but do not yet fully match human judgments of city-scale geographic similarity. Urban areas such as Zurich, Bern, and Thun also obtain relatively high combined scores, likely reflecting overlap in road density, water access, and terrain.

For the fused analysis, each patch is encoded by the four feature-specific encoders, the resulting 128-D vectors are concatenated into a 512-D representation, and the fusion MLP maps this input into a shared 128-D fused embedding. We then compare cities as distributions of fused patch embeddings using RBF-kernel MMD with $\gamma = 5$. This produces a single similarity score that reflects all four modalities jointly.

The fused rankings are geographically plausible. Zurich is most similar to Thun, Lausanne, and Bern, consistent with their shared character as relatively dense Swiss cities with high road density and water access. Lausanne is closest to Thun, Zurich, and Montreux, reflecting a mix of urban and lakeside structure. Zermatt is closest to Grindelwald and Lauterbrunnen, as expected for alpine mountain towns.

The two methods serve different purposes. The feature-specific analysis provides interpretability by revealing which geographic features contribute to similarity between two cities. The fused analysis supports retrieval by producing a single overall similarity score integrating all modalities.

Table 13: City-to-city similarity (top-5) using RBF-kernel MMD on fused embedding distributions ($\gamma = 5$). Higher values indicate greater similarity (scale 0-1).

Query	Similar City	Fusion
Zurich	Thun	0.8684
	Lausanne	0.8289
	Bern	0.8226
	Montreux	0.6393
	Zermatt	0.5377
Lausanne	Thun	0.9252
	Zurich	0.8289
	Montreux	0.8114
	Bern	0.6563
	Lauterbrunnen	0.6078
Zermatt	Grindelwald	0.8766
	Lauterbrunnen	0.8602
	Montreux	0.7640
	Thun	0.6175
	Lausanne	0.5872

The similarities do not perfectly match human intuition in all cases, highlighting a limitation of geographically constrained training data where the encoder may learn region-specific patterns rather than universally transferable representations. Future work could expand training to more globally diverse datasets and investigate the sources of counterintuitive similarity scores.

K Rio de Janeiro embedding space

To assess the generalizability of our approach to diverse geographic contexts, we applied the same encoding methodology to Rio de Janeiro, Brazil. We trained separate topography and water proximity encoders on GeoTIFFs of Rio de Janeiro.

Topographic data was obtained from OpenTopography (Robinson et al., 2017). Water proximity was computed using the same Global Surface Water dataset (European Commission, 2024) and tiled Euclidean distance transform methodology as for Switzerland.

The encoders followed identical architectures and training procedures as described in Section 2. Water proximity bins were adjusted to maintain balanced sampling given Rio's different geographic distribution, with more patches being close to water. Cells were binned into near (< 0.2 km), mid (0.2-1.2 km), and far (> 1.2 km) categories, compared to the original thresholds of < 3 km, 3-10 km, and > 10 km used for Switzerland.

Figure 9 in Appendix L shows the 1-D PCA visualizations for both topography and water proximity embeddings alongside their corresponding input GeoTIFFs. The embeddings capture interpretable geographic patterns: the topography encoder distinguishes Rio's coastal lowlands from the mountainous inland terrain, while the water proximity encoder distinguishes coastal areas and rivers from inland regions. These results demonstrate that our feature-specific encoding approach generalizes beyond Switzerland, though the need to adjust water proximity bins for adapting to different geographic distributions highlights a requirement of region-specific configurations for optimal performance.

L FIGURES

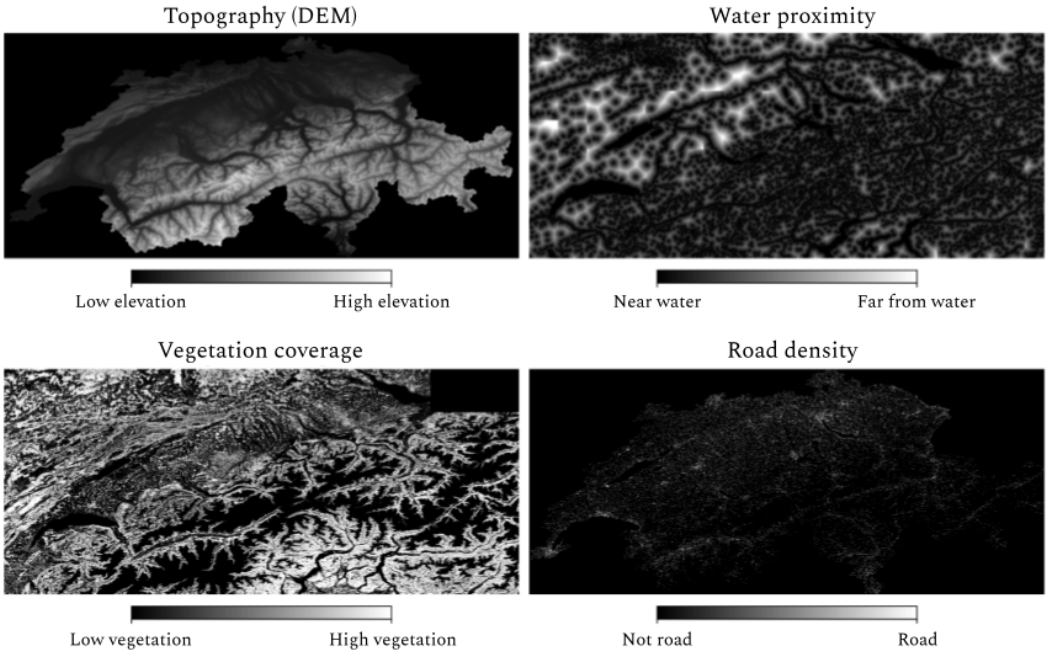

Figure 1: Input GeoTIFF files used to train the four feature-specific encoders. The digital elevation model (DEM) and water proximity rasters contain continuous-valued gradients, while the vegetation cover and road density rasters are binary masks.

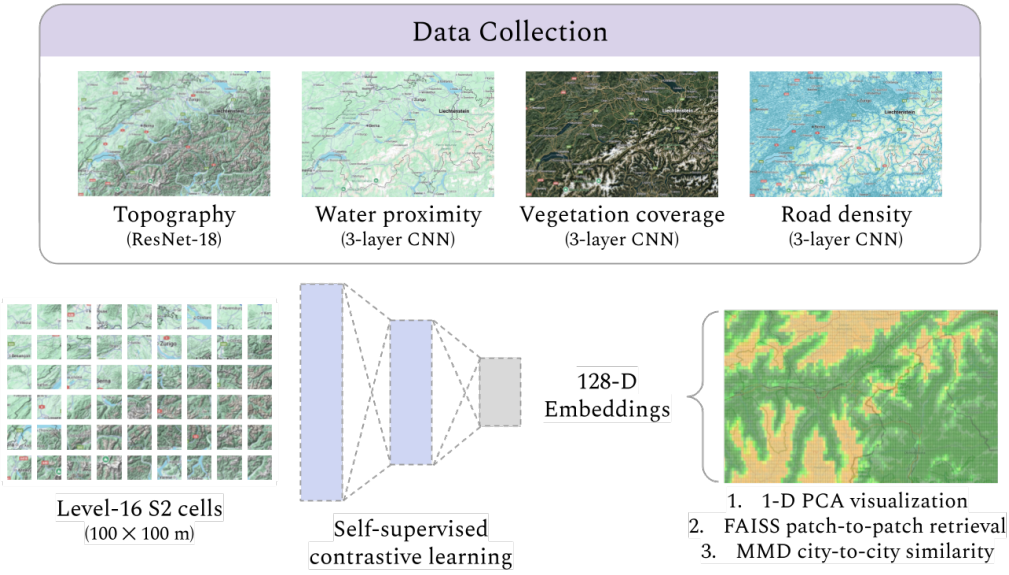

Figure 2: Pipeline illustration of data collection for four geographic features, S2 patch extraction for each GeoTIFF, encoder training, and the embedding space visualization. The learned embeddings are then used for similarity retrieval. Map data © 2026 Google.

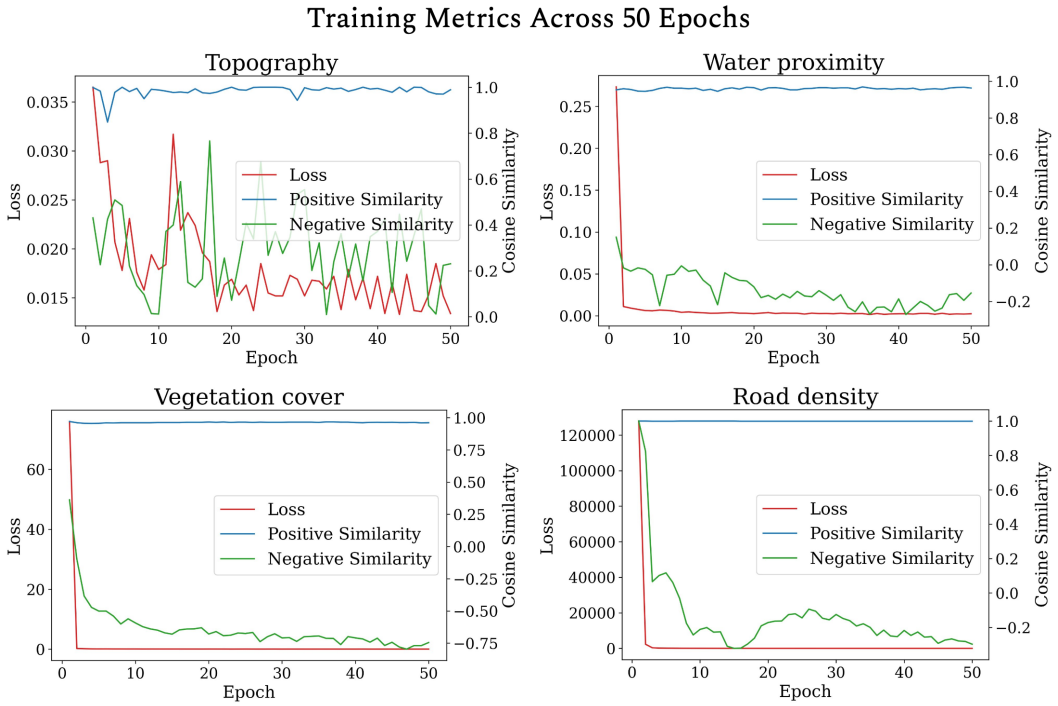

Figure 3: Plots of the loss (red), positive cosine similarity (blue), and negative cosine similarity (green) for each feature-specific encoder over 50 epochs.

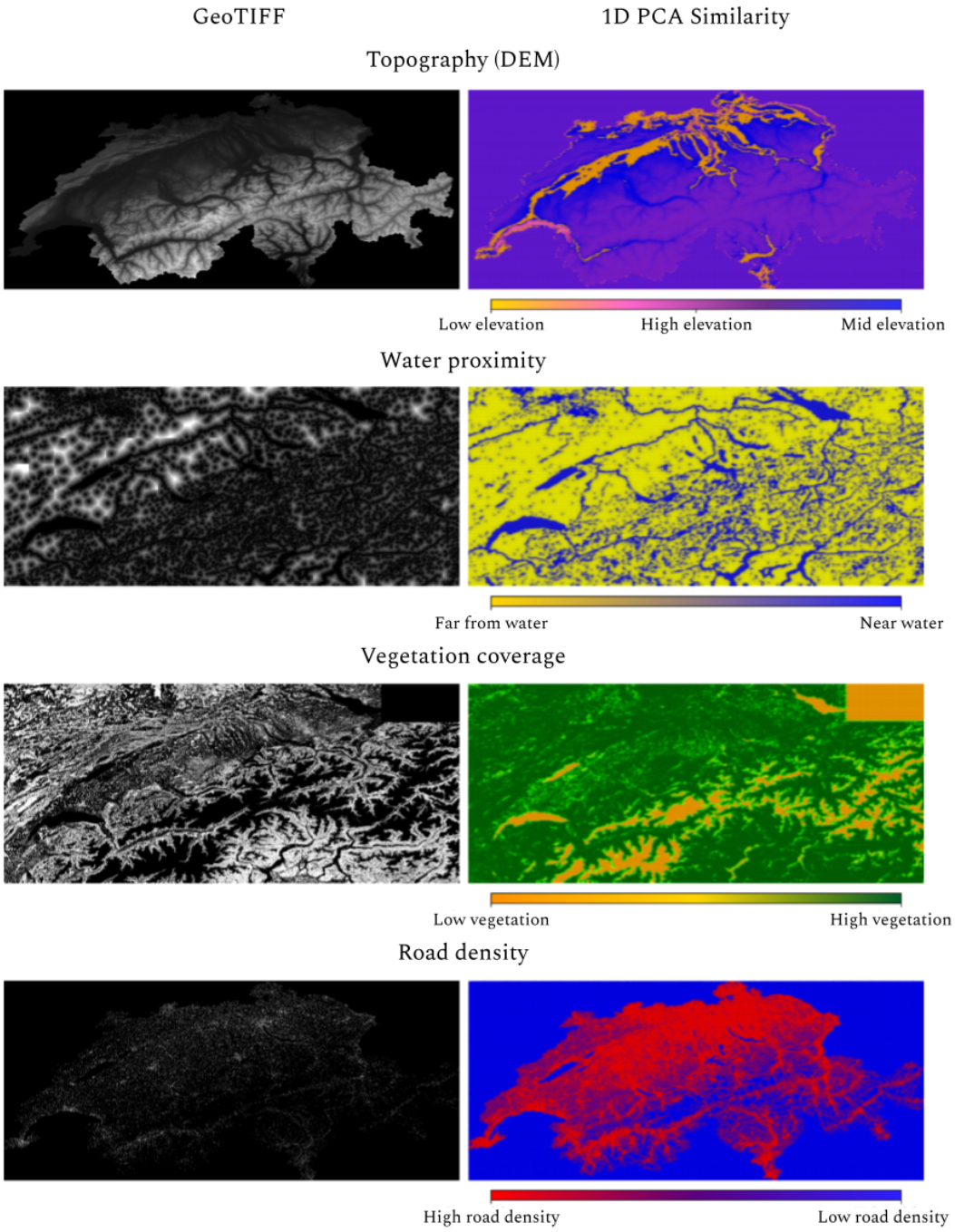

Figure 4: Side-by-side comparison of the input GeoTIFF and the learned embedding spaces for Switzerland. Each learned embedding space is projected to one dimension with PCA and rendered spatially, demonstrating that the encoders preserve broad geographic structure while compressing each modality into a compact representation. In the topography embedding space, the patches with low elevation are closer to the patches with high elevation than to the patches with mid-level elevation. This result could reflect the triplet training objective's emphasis on relative similarity, where both valleys and peaks may share structural features (e.g., steep gradients, surrounding terrain shapes) that differ from more uniform mid-elevation landscapes.

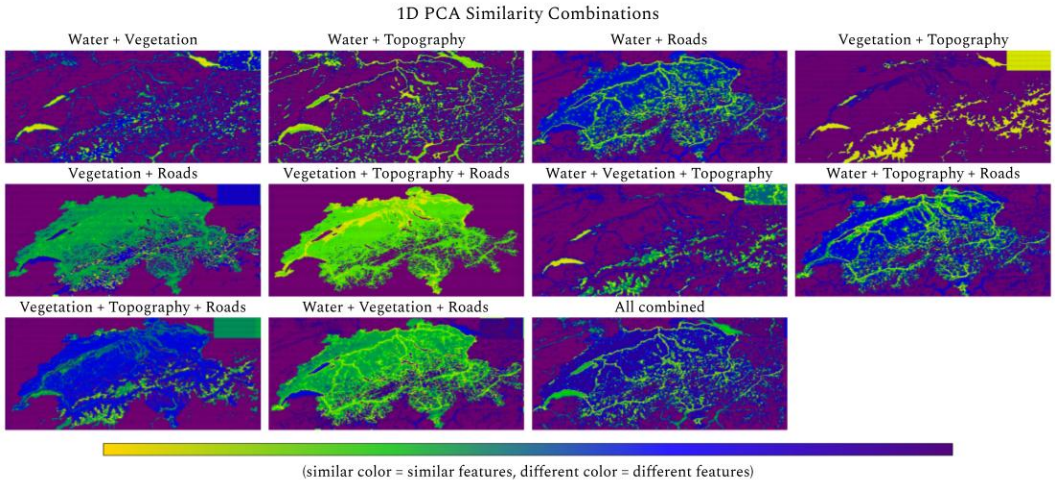

Figure 5: 1-D PCA visualizations of weighted combinations of multiple feature-specific embeddings.

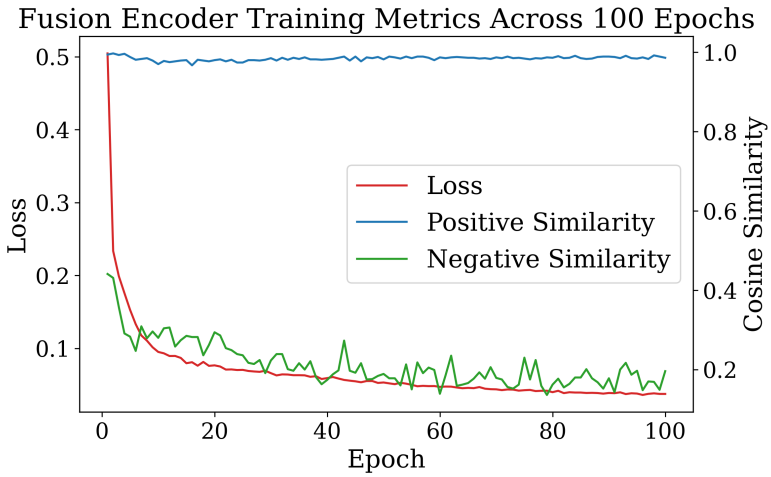

Figure 6: Plots of the loss (red), positive cosine similarity (blue), and negative cosine similarity (green) for the fused feature encoder over 100 epochs.

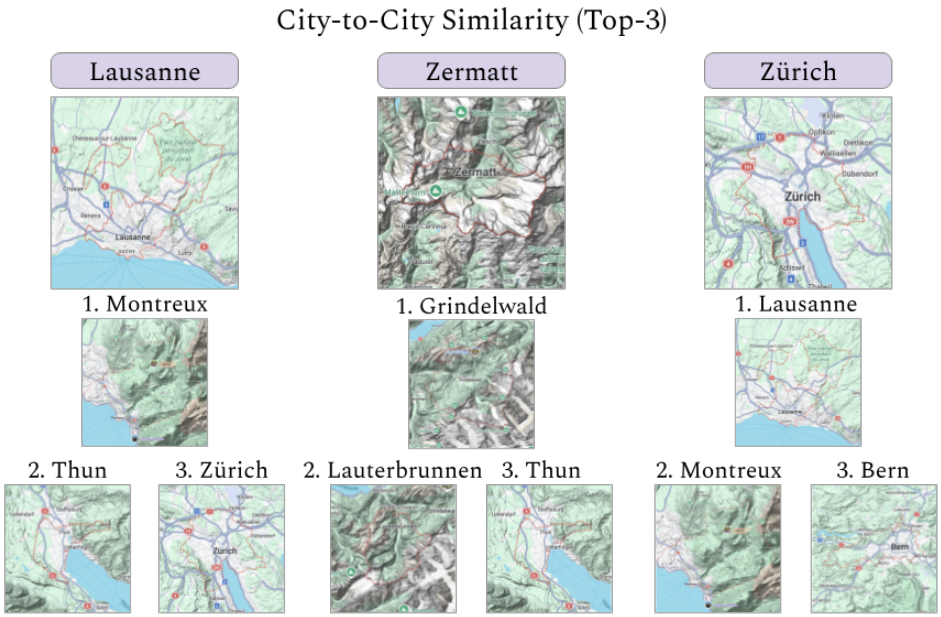

Figure 7: City-to-city similarity results using the feature-specific embeddings. Each column represents a query city (top), with the three most similar cities shown below based on 32-D PCA-projected MMD. Map data © 2026 Google.

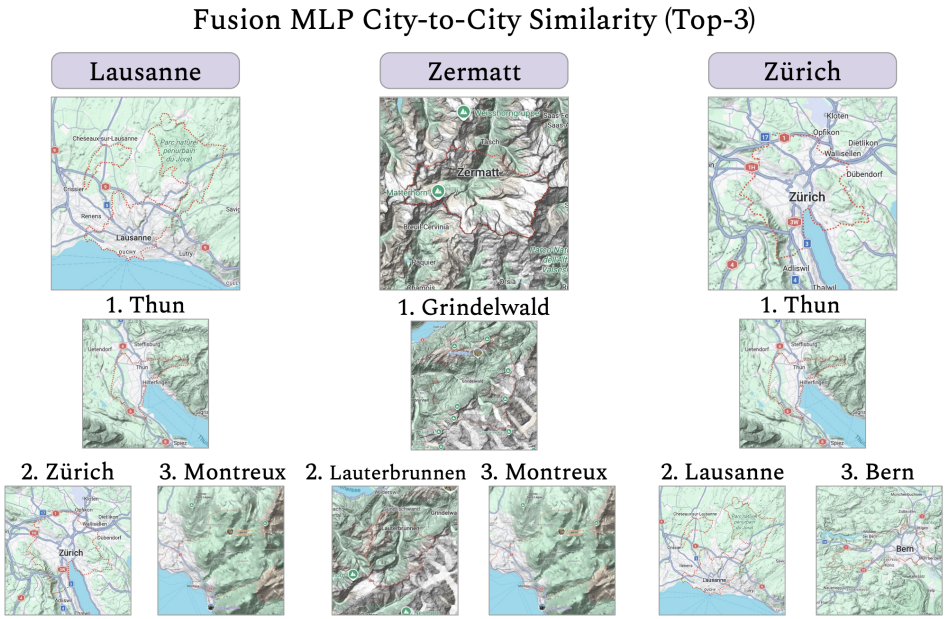

Figure 8: City-to-city similarity results using the learned fused embeddings. Each column shows a query city (top) and its three most similar cities below, ranked using PCA-projected MMD on distributions of fused patch embeddings. Compared with the feature-specific retrieval results in Figure 7, the fused model produces a single overall similarity ranking that better reflects joint geographic character across topography, water proximity, vegetation cover, and road density. Map data © 2026 Google.

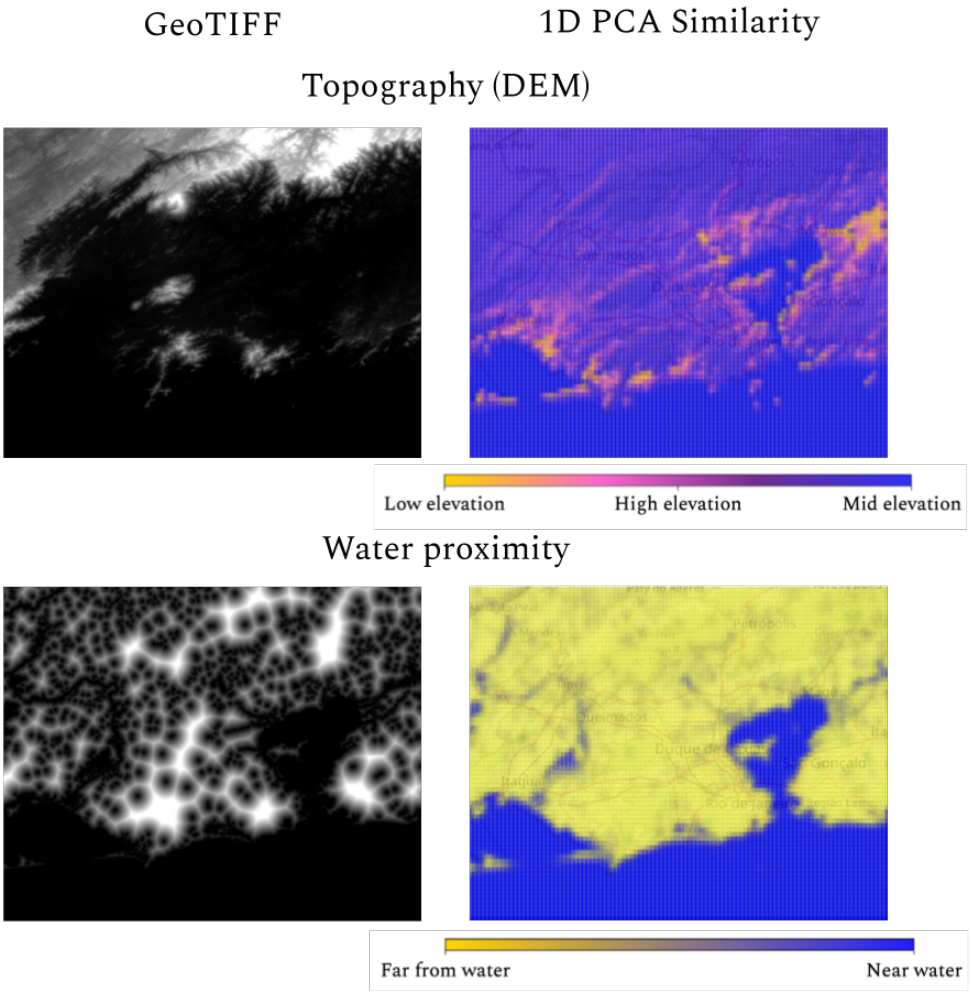

Figure 9: Side-by-side comparison of the input GeoTIFF and the learned embedding spaces for topography and water proximity, visualized via a 1-D PCA mapping for Rio de Janeiro.