SatOSM: Training geospatial foundation models with the Earth's largest open ground truth

Chenhui Zhang **Adam Yang** **Ruizhe Huang** **Jordi Laguarta Soler**

Xinyi Tong **Jonathan Giezendanner** **Sherrie Wang**

MIT

Abstract

Breakthroughs in large language models suggest that scaling model pretraining on large-scale, semantically diverse data can unlock generalizable understanding. In Earth observation (EO), a comparable "internet-scale" dataset has been missing. We introduce `SatOSM`, a large-scale dataset designed to fill this gap by leveraging OpenStreetMap (OSM) as a source of rich, open-vocabulary supervision. `SatOSM` contains 34 million high-resolution aerial images paired with 122 million OSM object instances across Europe, annotated by their footprints (polygons or line strings) and semantic tags (e.g., building=residential). These annotations, spanning buildings, roads, and land use polygons, are rasterized to the image extent to provide pixel-level instance masks suitable for large-scale pretraining with semantic and instance segmentation supervision. We then propose a novel cross-modal pretraining architecture that aligns mask-conditioned image embeddings — obtained via cross-attention between image tokens and OSM footprints — with text embeddings derived from OSM tags using a CLIP-style contrastive loss. A model pretrained on `SatOSM` outperforms a specialized U-Net trained from scratch and matches or surpasses existing foundation models on downstream tasks involving high-resolution EO imagery.

1 Introduction

Foundation models have revolutionized natural language processing [1, 17, 44] and computer vision [19, 24, 35, 36, 37, 49, 56] by pretraining on Internet-scale datasets with rich semantics, enabling the emergence of generalizable representations. Translating these advances to Earth observation (EO) data has become an active frontier [3, 5, 9, 20, 25, 38, 40, 41, 43, 50, 51, 52, 53, 54]. Geospatial foundation models (GFMs) promise to reduce the need for labeled data [27, 50] — which are often expensive or infeasible to collect at scale — and to lower the technical barrier for remote sensing analysis [42, 57].

Recent GFM architectures apply self-supervised pretraining to satellite and aerial imagery [3, 20, 25, 40, 50]. Most of these approaches adopt masked image modeling or self-distillation objectives inspired by MAE and DINO, learning generic spatio-spectral-temporal representations without explicit semantic supervision. However, there is growing evidence that purely self-supervised objectives may exhibit diminishing returns to scale in EO settings [5]. To introduce richer training signals, some works incorporate pseudo-labels — such as land cover maps — generated by specialized segmentation models (e.g., U-Nets) [20, 34, 50]. Yet these pseudo-labels typically exhibit high spatial variability in quality (e.g., reduced accuracy in underrepresented regions such as Africa [22]) and are

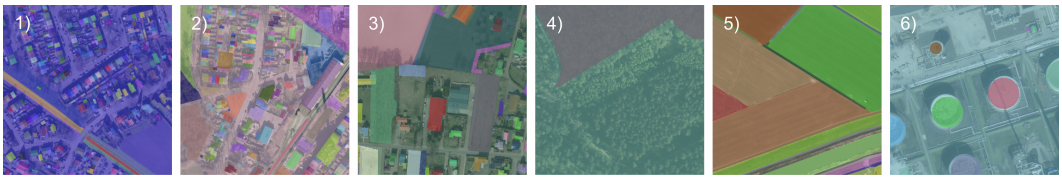

Figure 1: Examples of OpenStreetMap (OSM) annotations across our dataset. Each OSM object is visualized in a different color. 1)–3) show residential areas; 4) a forest landscape; 5) farmland; and 6) an industrial area. Tags present in the examples include: `building=apartments`, `building=house`, `building=industrial`, `amenity=school`, `amenity=parking`, `landuse=farmland`, and `landuse=forest`.

constrained to a small, fixed set of categories (e.g., nine land types in [4]), limiting their semantic diversity and generality.

In this work, we ask: *Can we jointly design data and model pretraining strategies that incorporate semantically rich and spatially precise supervision — beyond masked reconstruction or latent matching — to improve transfer on dense Earth observation tasks?*

To address this gap, we introduce `SatOSM`, a large-scale dataset and pretraining framework that leverages the rich, open-ended annotations of OpenStreetMap (OSM) to provide dense, semantically grounded supervision for high-resolution EO imagery. Our approach combines this new data source with a mask-grounded multimodal architecture that aligns OSM footprints and tags with image features through cross-attention and contrastive learning.

Our contributions are threefold:

- **Large-scale dataset:** We introduce `SatOSM`, the first large-scale, open-vocabulary, object-level dataset for Earth observation, pairing high-resolution imagery with fine-grained OSM footprints and tags to provide dense, semantically rich supervision.
- **Mask-grounded multimodal architecture:** We propose a novel pretraining framework that cross-attends image embeddings with OSM footprints and aligns them with text embeddings of OSM tags via a CLIP-style loss, learning transferable, object-level EO representations.
- **Empirical validation:** We show that a model pretrained on `SatOSM` outperforms specialized models and match or surpass existing GFMs on land cover segmentation, field delineation, and change detection.

2 `SatOSM`: Structured Supervision for Earth Observation

`SatOSM` is designed to enable scalable, open-vocabulary, and geographically diverse supervision for dense prediction tasks in EO. It pairs high-resolution EO imagery with OSM geometries and tags to create pixel-aligned, object-level annotations at a continental scale. `SatOSM` is publicly available on Hugging Face in MosaicML MDS format for direct streaming access.

Image sources We focus our first `SatOSM` release on aerial imagery rather than medium-resolution satellite data because many OSM features — such as buildings, roads, and small agricultural parcels — can only be delineated at sub-meter resolution. The spatial distribution of `SatOSM` is therefore jointly determined by aerial imagery availability and OSM availability. We use openly licensed aerial imagery from national mapping programs accessible through Google Earth Engine, including datasets from the Netherlands, Switzerland, Finland, Estonia, Latvia, Slovakia, Spain, and France. The imagery ranges from 6 cm to 50 cm per pixel and is divided into 550×550 pixel chips.

Data composition The current release contains 34 million image chips paired with 122 million OSM object instances across the eight countries. Each image chip is paired with all intersecting OSM geometries (polygons or lines) and their semantic tags. These geometries are stored as vector annotations and also rasterized into masks suitable for instance and semantic segmentation.

Semantic diversity Each OSM object is represented by its geometric footprint and associated semantic tags (e.g., `building=residential`, `landuse=farmland`). The semantic diversity of `SatOSM` stems from the open-ended tagging system of OSM, which we preserve without collapsing tags into a fixed ontology. Our dataset includes annotations with 2,219 unique key-value tags relevant for dense prediction, drawn from a wide spectrum of domains such as buildings, land use, roads, water bodies, and natural features. The tag distribution is heavily long-tailed: the top 1% of tags account

for over 88% of all annotations, while the median tag appears fewer than 100 times. Common tags such as `building=yes` (14M), `landuse=grass` (12.6M), and `highway=track` (5.5M) dominate the dataset, but `SatOSM` also includes thousands of fine-grained tags such as `natural=heath`, `amenity=parking_space`, and `generator:method=wind_turbine`, each contributing niche but semantically meaningful supervision signals. This imbalance reflects real-world semantic diversity and makes `SatOSM` a valuable testbed for learning from highly imbalanced, open-set label spaces.

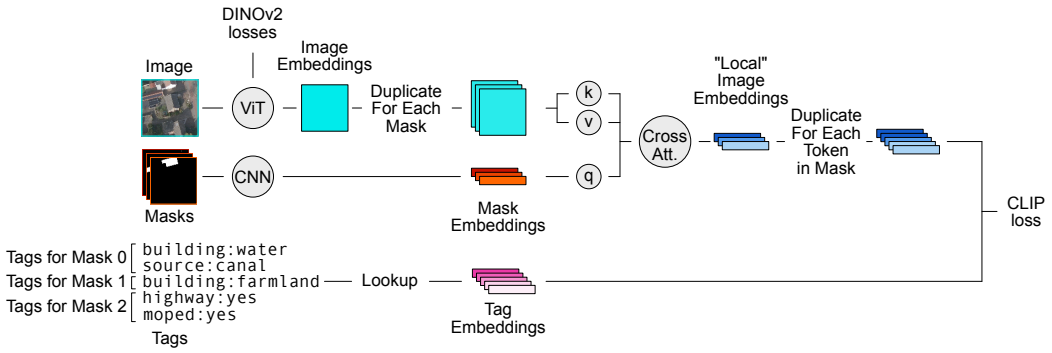

Figure 2: Schematic overview of our pretraining objective. The figure shows an example for one image with three masks, each mask with respectively two, one and two OSM tags.

3 Pre-training on `SatOSM`

We aim to design a pretraining framework that can learn representations useful for both image-level and dense prediction tasks in EO. Existing GFMs based on self-supervised objectives such as masked image modeling or self-distillation excel at learning global image features but lack explicit spatial grounding. Meanwhile, vision-language models provide semantic alignment but typically supervise at the image level. Our goal is to combine the strengths of both: to build a model that learns semantically meaningful, spatially grounded features while remaining effective for global representation learning.

We introduce `SatOSM-Net`, a multi-modal vision-transformer (ViT) architecture that integrates self-supervised visual learning with mask-grounded semantic alignment from `SatOSM`. In addition to standard image- and patch-level embeddings, `SatOSM-Net` extracts region-level embeddings corresponding to individual OSM objects (Figure 2).

Region-level embedding Starting from a ViT backbone [26], we add a learnable attention-pooling head that can extract an embedding vector for a specific region of the image, i.e., the footprint of an OSM object. The attention pooling head receives the query from a convolutional layer that turns an object mask into a query vector of the same embedding dimension as the ViT. The ViT patch tokens serve as keys and values, and a cross-attention operation combines them into a single region that represents the visual characteristics of the object.

Tag embedding Each OSM object is associated with one or more semantic tags. We embed these tags using a learnable lookup table and aggregate them into a tag embedding of the same dimension as the visual region embedding. This provides a weak semantic supervision signal from `SatOSM`.

Training objectives Our strategy to train this architecture is two-fold. (1) To align vision and text representations, we gather all regional visual and tag embeddings in the same batch and apply a CLIP contrastive loss [36] with a learnable temperature. (2) To learn rich visual features directly from the image data, we incorporate the objectives from DINOv2 [35]. This involves using a momentum encoder as a teacher network and applying both an image-level and a patch-level loss between the student and teacher outputs. This combined approach allows `SatOSM-Net` to learn general-purpose representations by leveraging both semantic alignment and self-supervised feature learning. We train our model on 8 GH200 GPUs on `SatOSM` for 125,000 steps with a local batch size of 256.

4 Experiments

Datasets To demonstrate the effectiveness of `SatOSM-Net` for dense prediction pretraining, we conduct experiments on three datasets that use high-resolution remote sensing images for segmentation:

Table 1: Segmentation results on AI4BOUNDARIES, MINIFRANCE and XVIEW2. (a) `SatOSM-Net` achieves the best overall performance on most high-resolution tasks. (b) Gradual unfreezing outperforms head-only and full fine-tuning.

(a)

Model	AI4BOUNDARIES		MINIFRANCE		XVIEW2	
	IoU	**mAP@0.5**	**mIoU**	**FWIoU**	**mIoU**	**FWIoU**
Non-pretrained						
U-Net	69.14	65.49	45.58	55.15	60.06	79.85
ViT	54.82	48.11	32.33	45.94	56.02	77.18
Pretrained GFMs (Gradual unfreezing)						
Scale-MAE	64.51	59.17	43.96	53.61	53.71	75.59
SkyCLIP-50	64.59	52.76	53.00	58.05	**65.19**	**82.43**
DOFA-CLIP	75.34	71.93	53.22	58.39	63.57	81.39
`SatOSM-Net`	**77.48**	**74.47**	**56.76**	**60.35**	64.77	81.48

(b)

`SatOSM-Net` **Fine-tuning**	AI4BOUNDARIES		MINIFRANCE		XVIEW2	
	IoU	**mAP@0.5**	**mIoU**	**FWIoU**	**mIoU**	**FWIoU**
Head-only	74.49	70.13	52.42	57.90	62.93	81.26
Full fine-tune	69.07	64.33	44.02	53.13	53.71	76.13
Gradual unfreezing	**77.48**	**74.47**	**56.76**	**60.35**	**64.77**	**81.48**

AI4Boundaries [10], **MiniFrance** [6] and **xView2** [18]. AI4Boundaries is a field delineation dataset spanning 7 countries pairing satellite and high-resolution aerial images (1m) with vectorized field boundaries. For our experiments, we train and evaluate on high resolution aerial images, which contains over 7500 samples at 512 pixel size. We also evaluate on MiniFrance, a high-resolution land use classification dataset (0.5 m). The data are provided as RGB tiles and comprise 14 semantically rich land use classes. We partition the original data for training and testing by extracting non-overlapping 512 pixel chips from a subset where annotations are available (575 tiles of size 10,000 pixel from Nice and Nantes/Saint-Nazaire). xView2 is a dataset focused on building damage caused by disasters such as flooding, hurricanes, earthquake and fires, covering 19 events in the training/validation set, and 10 events in the test set. The dataset contains high resolution pre- and post-event Maxar images, as well as five categories of labels (no building, no damage, minor damage, major damage and destroyed). Each image is provided as a RGB 1024 pixel sized chip with a $\sim$0.5m ground resolution. The train/validation set contains 9168 chips, which, following Marsocci et al., is split 90/10 and oversampled on building during training. Further evaluation details are provided in E.

Baselines We include randomly initialized U-Net [39] and ViT [26] as baselines and compare against three GFMs designed to handle high-resolution inputs: **Scale-MAE** [38] (a masked autoencoder pretrained on large-scale satellite imagery), **SkyCLIP-50** [53] (a CLIP-style vision-language model using captions generated from OSM labels), and **DOFA-CLIP** [54] (a domain-adapted CLIP model trained on multimodal EO image-text datasets). `SatOSM-Net` and all other ViT-based models use a ViT-Large backbone followed by a Fully Convolutional Network (FCN) segmentation head. Concretely, we reshape the final ViT patch tokens into a spatial feature map and feed it into a lightweight FCN decoder, where the FCN head consumes the last-layer embeddings. For xView2, the FCN decoder is replaced by a Siamese UPernet decoder [32] which ingests the encoded transformer blocks. Downstream training and evaluation are performed on four NVIDIA L40S GPUs. Each method is trained for 20, 50 and 80 epochs on MiniFrance, Ai4Boundaries and xView2 respectively.

Fine-tuning `SatOSM-Net` on downstream tasks The effectiveness of pretrained vision models depends heavily on how they are fine-tuned. To evaluate how well `SatOSM-Net` 's pretrained features transfer to prediction tasks, we compared 3 fine-tuning procedures: **head-only (HO)**, **full fine-tuning (FT)**, and **gradual unfreezing (GU)**. In HO, the ViT backbone is frozen and only the decoder segmentation head is trained. In FT, the backbone and head are optimized jointly under the same hyperparameters. In GU, we first train only the head for 2 epochs, then unfreeze one ViT block every 2 epochs until the backbone is fully trainable.

Results On both AI4Boundaries and MiniFrance, `SatOSM-Net` achieves the best overall segmentation accuracy, outperforming DOFA-CLIP (the strongest prior model) by +2.1 IoU and +2.5 mAP@0.5 on AI4Boundaries, and slightly improving mIoU and FWIoU on MiniFrance. `SatOSM-Net` also achieves the second best performance on xView2, competitive with other GFMs. This indicates that grounding image representations with OSM masks and tags provides additional spatial and semantic signal beyond standard self-supervised or image-text pretraining.

GU yields the highest scores on all datasets, outperforming HO and FT by a large margin. This suggests that progressive adaptation allows the pretrained backbone to retain generalizable features from `SatOSM` while still specializing to the target domain. Together, these results demonstrate that `SatOSM-Net` adapted with GU effectively transfers to high-resolution segmentation tasks, surpassing existing high-resolution GFMs.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[2] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15619–15629, 2023.

[3] G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu. Anysat: An earth observation model for any resolutions, scales, and modalities. arXiv preprint arXiv:2412.14123, 2024.

[4] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. Scientific Data, 9(1):251, 2022.

[5] C. F. Brown, M. R. Kazmierski, V. J. Pasquarella, W. J. Rucklidge, M. Samsikova, C. Zhang, E. Shelhamer, E. Lahera, O. Wiles, S. Ilyushchenko, N. Gorelick, L. L. Zhang, S. Alj, E. Schechter, S. Askay, O. Guinan, R. Moore, A. Boukouvalas, and P. Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025. URL https://arxiv.org/abs/2507.22291.

[6] J. Castillo-Navarro, B. L. Saux, A. Boulch, N. Audebert, and S. Lefèvre. Semi-Supervised Semantic Segmentation in Earth Observation: The MiniFrance Suite, Dataset Analysis and Multi-task Network Study, Oct. 2020. URL http://arxiv.org/abs/2010.07830.

[7] C. Chambers, A. Raniwala, F. Perry, S. Adams, R. R. Henry, R. Bradshaw, and N. Weizenbaum. Flumejava: easy, efficient data-parallel pipelines. ACM Sigplan Notices, 45(6):363–375, 2010.

[8] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14455–14465, 2024.

[9] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems, 35:197–211, 2022.

[10] R. d'Andrimont, M. Claverie, P. Kempeneers, D. Muraro, M. Yordanov, D. Peressutti, M. Batič, and F. Waldner. AI4Boundaries: an open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography. Earth System Science Data, 15(1):317–329, Jan. 2023. ISSN 1866-3508. doi: 10.5194/essd-15-317-2023. URL https://essd.copernicus.org/articles/15/317/2023/.

[11] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008.

[12] GDAL/OGR contributors. GDAL/OGR Geospatial Data Abstraction software Library. Open Source Geospatial Foundation, 2025. URL https://gdal.org.

[13] S. Gillies et al. Rasterio: geospatial raster I/O for Python programmers. Mapbox, 2013–. URL https://github.com/rasterio/rasterio.

[14] Google Cloud. Dataflow: Stream and Batch Data Processing, 2024. URL https://cloud.google.com/dataflow. Accessed on 2025-05-15.

[15] Google Inc. Protocol Buffers. Google, 2008. URL https://protobuf.dev/. Version 2 initially released in 2008. Latest version and documentation available online.

[16] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment, 202: 18–27, 2017.

[17] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[18] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery.

In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 10–17, 2019.

[19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000–16009, 2022.

[20] J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. arXiv preprint arXiv:2504.11171, 2025.

[21] A. Kamath, J. Hessel, and K.-W. Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. Conference on Empirical Methods in Natural Language Processing, 2023. doi: 10.48550/arXiv.2310.19785.

[22] H. Kerner, C. Nakalembe, A. Yang, I. Zvonkov, R. McWeeny, G. Tseng, and I. Becker-Reshef. How accurate are existing land cover maps for agriculture in sub-saharan africa? Scientific Data, 11(1):486, 2024.

[23] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9404–9413, 2019.

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4015–4026, 2023.

[25] K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 4347–4355, 2025.

[26] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, 2021.

[27] A. Lacoste, N. Lehmann, P. Rodriguez, E. Sherwin, H. Kerner, B. Lütjens, J. Irvin, D. Dao, H. Alemohammad, A. Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. Advances in Neural Information Processing Systems, 36:51080–51093, 2023.

[28] S. Li, P. W. Koh, and S. S. Du. Exploring how generative mllms perceive more than clip with the same vision encoder. arXiv preprint arXiv: 2411.05195, 2024.

[29] I. Loshchilov and F. Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts, May 2017. URL `http://arxiv.org/abs/1608.03983`. arXiv:1608.03983 [cs].

[30] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization, Jan. 2019. URL `http://arxiv.org/abs/1711.05101`. arXiv:1711.05101 [cs].

[31] K.-K. Maninis, K. Chen, S. Ghosh, A. Karpur, K. Chen, Y. Xia, B. Cao, D. Salz, G. Han, J. Dlabal, et al. Tips: Text-image pretraining with spatial awareness. arXiv preprint arXiv:2410.16512, 2024.

[32] V. Marsocci, Y. Jia, G. L. Bellier, D. Kerekes, L. Zeng, S. Hafner, S. Gerard, E. Brune, R. Yadav, A. Shibli, H. Fang, Y. Ban, M. Vergauwen, N. Audebert, and A. Nascetti. PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models, Apr. 2025.

[33] D. Mizrahi, R. Bachmann, O. Kar, T. Yeo, M. Gao, A. Dehghan, and A. Zamir. 4m: Massively multimodal masked modeling. Advances in Neural Information Processing Systems, 36:58363–58408, 2023.

[34] V. Nedungadi, A. Kariryaa, S. Oehmcke, S. Belongie, C. Igel, and N. Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In European Conference on Computer Vision, pages 164–182. Springer, 2024.

[35] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.

[37] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024.

[38] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4088–4099, 2023.

[39] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[40] J. Schmude, S. Roy, W. Trojak, J. Jakubik, D. S. Civitarese, S. Singh, J. Kuehnert, K. Ankur, A. Gupta, C. E. Phillips, et al. Prithvi wxc: Foundation model for weather and climate. arXiv preprint arXiv:2409.13598, 2024.

[41] H. Si, Y. Wan, M. Do, D. Vasisht, H. Zhao, and H. F. Hamann. Towards scalable foundation model for multi-modal and hyperspectral geospatial data. arXiv preprint arXiv: 2503.12843, 2025.

[42] S. Singh, M. Fore, and D. Stamoulis. Geollm-engine: A realistic environment for building geospatial copilots. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 585–594, 2024.

[43] D. Szwarcman, S. Roy, P. Fraccaro, Þ. E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. d. S. Almeida, R. Sedona, Y. Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. arXiv preprint arXiv:2412.02732, 2024.

[44] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.

[45] T. M. M. Team. streaming. <https://github.com/mosaicml/streaming/>, 2022.

[46] TensorFlow Authors. Tfrecord and tf.train.Example, 2024. URL https://www.tensorflow.org/tutorials/load_data/tfrecord. Accessed: 2025-05-14.

[47] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9568–9578, 2024.

[48] J. Topf. Osmium tool: Command line tool for working with openstreetmap data, 2025. URL https://github.com/osmcode/osmium-tool. Version 1.18.0, released March 17, 2025.

[49] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.

[50] G. Tseng, A. Fuller, M. Reil, H. Herzog, P. Beukema, F. Bastani, J. R. Green, E. Shelhamer, H. Kerner, and D. Rolnick. Galileo: Learning global and local features in pretrained remote sensing models. arXiv preprint arXiv:2502.09356, 2025.

[51] L. Waldmann, A. Shah, Y. Wang, N. Lehmann, A. J. Stewart, Z. Xiong, X. X. Zhu, S. Bauer, and J. Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. arXiv preprint arXiv:2503.10845, 2025.

[52] Y. Wang, Z. Xiong, C. Liu, A. J. Stewart, T. Dujardin, N. I. Bountos, A. Zavras, F. Gerken, I. Papoutsis, L. Leal-Taixé, et al. Towards a unified copernicus foundation model for earth vision. arXiv preprint arXiv:2503.11849, 2025.

[53] Z. Wang, R. Prabha, T. Huang, J. Wu, and R. Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 5805–5813, 2024.

[54] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. L. Saux, G. Camps-Valls, and X. X. Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. arXiv preprint arXiv:2403.15356, 2024.

[55] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In 2nd USENIX workshop on hot topics in cloud computing (HotCloud 10), 2010.

[56] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11975–11986, 2023.

[57] C. Zhang and S. Wang. Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7839–7849, 2024.

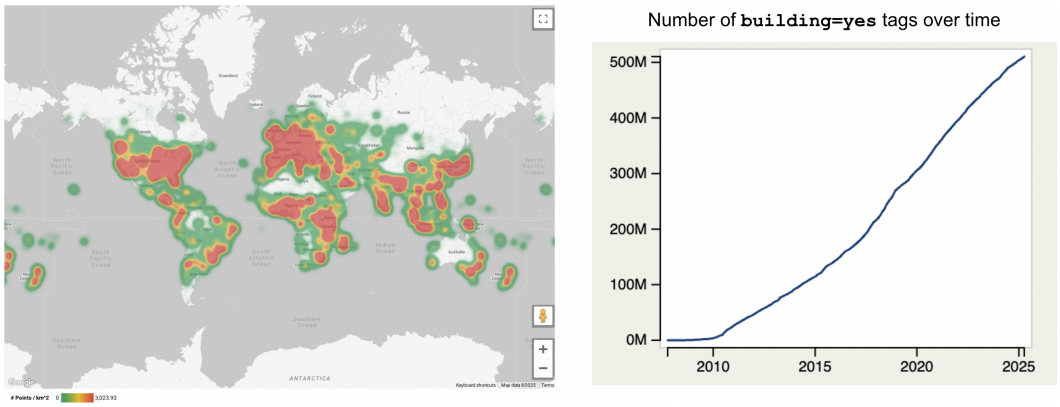

Figure 3: (Left) OpenStreetMap annotations are global; density is measured in number of annotations per km^2. (Right) The number of OSM objects has increased steeply since 2010 (shown for tag `building=yes`).

A Related Works

Pretrained Vision Models on Earth Observation Data Earth observation communities have also made significant efforts to reproduce the success of large-scale pre-training in language and vision, with substantial progress in developing sensor-agnostic vision foundation models that can flexibly ingest data from arbitrary satellite sensors. For example, Panopticon [51] extends DINOv2 [35] to handle any combination of optical and radar bands. Copernicus-FM [52] introduces a dynamic hypernetwork that generates model weights conditioned on the sensor type. DOFA [54] uses hypernetworks to adapt a single transformer to five different sensors. AnySat [3] adapts I-JEPA [2] to pretrain a transformer backbone on heterogeneous data. Galileo [50] introduces a family of models designed for both global and local feature learning in remote sensing. LESS-ViT [41] further improves the scalability of EO pre-training by factorizing spectral–spatial attention. Furthermore, TerraMind [20] adapts 4M [33] to train an any-to-any generative model on a diverse set of EO modalities. Prithvi-EO-2.0 [43] and SatMAE [9] employ a masked autoencoder approach to learn spatiotemporal representations from satellite imagery. SkyScript [53] proposes a rule-based method to generate synthetic captions for high-resolution aerial images based on OSM.

Although prior works have explored various self-supervised training methods adapted from computer vision, they largely omitted the possibility of using semantically rich and dense aware supervision signals in a self-supervised pre-training pipeline.

Diversifying Supervision Signals in Vision Pretraining Contrastive text-image pretraining [36, 49] has become one of the de facto standards, along with vision-centric pre-training, for training a strong vision backbone. However, vision models trained with purely text-image supervision often contain "glitches", such as erroneous agreements in the latent space [47] and a lack of spatial awareness [21, 28], limiting their usefulness in dense prediction tasks. To mitigate the lack of spatial awareness in common vision-language models, various methods ranging from synthetic text captions focused on dense understanding [31], incorporating diverse pretext tasks that encourage spatial coherence [31, 49], and using a carefully curated dataset and task mixture [8]. This work's philosophy aligns with the prior works in this area, aiming to bring more diverse supervision signals into EO pretraining.

B Preliminaries: OpenStreetMap and Earth Observation Datasets

OpenStreetMap (OSM). OSM is a globally crowdsourced geospatial database containing over 10 billion objects annotated by a community of more than 10 million contributors (Figure 3). Each object is associated with a geometry (point, line, or polygon) or a collection of geometries and a set of key-value tags describing its semantics (e.g., `building=house`, `landuse=orchard`). These annotations are created through manual digitization, GPS-based surveys, and satellite imagery annotation, and are continuously refined through community validation efforts. The resulting dataset offers a rich source

of supervision that is spatially precise and semantically open-ended. The construction of SatOSM takes advantage of the following features of OSM.

1. **Diverse and compositional semantic labels.** Unlike conventional remote sensing datasets, which are typically constrained to a fixed taxonomy of object classes, OSM spans hundreds of thousands of unique tag combinations. While a small set of tags dominate in frequency (e.g., `building=yes`), the long tail includes infrastructure, land use categories, natural features, and agricultural attributes (Table 2). This richness enables us to supervise models with detailed, instance-level distinctions (e.g., `amenity=school` vs. `amenity=hospital`) that are not available in fixed-class benchmarks. Furthermore, the open tagging schema enables hierarchical and compositional semantics; for example, a structure may be tagged as both `building=retail` and `shop=bakery`, enriching the contextual signal beyond what is possible with standard class labels.
2. **High-precision spatial alignment.** OSM annotations are defined with meter-level accuracy and can be aligned directly with high-resolution EO imagery. This allows us to generate pixel-level segmentation masks by rasterizing vector geometries onto image tiles. Because OSM features are not restricted to any particular schema, we can produce instance, semantic, or panoptic segmentation labels depending on downstream needs.
3. **Global geographic coverage.** OSM's geographic footprint spans most of the inhabited world, with dense coverage across Europe, North America, and urban regions in other continents (Figure 3). This global reach is driven by a distributed community of contributors, including disaster response initiatives (e.g., post-earthquake mapping by the Humanitarian OpenStreetMap Team), targeted campaigns in under-mapped regions (e.g., by Missing Maps), and local mappers who contribute contextual knowledge. OSM is therefore well-suited for the development of EO datasets that enable models to generalize globally.

At the same time, OSM has several limitations from the perspective of training computer vision models for EO. First, OSM geometries are not always perfectly aligned with satellite or aerial imagery, leading to potential spatial misregistration. Second, its labels are incomplete; coverage varies by region, and many objects or classes of interest may be missing entirely. Third, OSM exhibits geographic and socioeconomic biases, with disproportionately dense labeling in high-income and Western countries compared to lower-income or rural regions. Finally, as a crowd-sourced dataset, OSM is susceptible to labeling noise and inconsistencies, which can impact supervision quality. Despite these imperfections, OSM remains the most comprehensive open global database available and contains large potential for EO model development.

Open Aerial and Satellite Imagery. Publicly available Earth observation imagery spans a wide range of spatial resolutions and modalities, from coarse global products to sub-meter-level aerial surveys. High-resolution imagery is particularly well-suited for tasks that require detailed spatial understanding, such as object detection. When paired with OSM, high-resolution imagery enables precise alignment between visual features and labeled geometries like building footprints, roads,

Table 2: Examples of OpenStreetMap tags, from most common (`building=yes`) to least common. Tags may include descriptions of the object and abstract information, such as administrative geocoding.

Tag	Count	Percent
`building=yes`	510,470,126	4.66%
`highway=residential`	67,708,948	0.62%
`natural=tree`	28,487,649	0.26%
...		
`leaf_cycle=deciduous`	3,629,444	0.03%
`tunnel=culvert`	3,526,881	0.03%
`amenity=parking_space`	3,455,545	0.03%
...		
`generator:method=wind_turbine`	410,729	0.00%
`NHD:FCode=39004`	407,206	0.00%
`railway=abandoned`	406,010	0.00%

or field boundaries. In contrast, coarser imagery (e.g., 10–30 m per pixel) may fail to resolve small or narrow features. For the construction of SatOSM, we focus on pairing OSM labels with publicly-available high-resolution remote sensing data.

A growing number of high-resolution satellite and aerial imagery sources are now openly available for research use. Several European national mapping agencies provide aerial imagery at sub-meter resolution under permissive licenses, including the Netherlands and Switzerland, among others. These datasets are typically produced through government-led aerial surveys and offer consistent, orthorectified imagery suitable for training. In the United States, the National Agriculture Imagery Program (NAIP) provides openly licensed aerial imagery at sub-1m resolution, covering the continental U.S. on a biannual basis. On the satellite side, however, most very high-resolution imagery remains commercially licensed, limiting its accessibility for open research and large-scale dataset construction. A notable exception is the Maxar Open Data Program, which releases high-resolution satellite imagery for select disaster response and humanitarian use cases. This limited availability presents a key challenge for extending OSM-imagery pairings to under-mapped or lower-income regions where open aerial data is scarce. In this work, we focus SatOSM on European regions where high-resolution aerial imagery is both publicly available and well-aligned with OSM. Expanding the dataset to additional geographies remains an important direction for future work.

C SatOSM: Structured Supervision for Earth Observation

C.1 Design Principles

The design of SatOSM is informed by the goal of enabling scalable, open-vocabulary, and geographically diverse supervision for dense prediction tasks in EO. To this end, we outline three core design principles:

1. **Open-vocabulary supervision.** SatOSM departs from the conventional fixed-class paradigm in remote sensing datasets by leveraging the open-ended tagging schema of OpenStreetMap (OSM). Tags such as `building=house`, `landuse=orchard`, or `shop=bakery` capture both class and attribute-level information. We preserve the original OSM key-value tagging semantics to support explorations in open-vocabulary pre-training.
2. **Pixel-aligned dense labels.** Supervision in SatOSM is provided in the form of per-pixel semantic or instance masks. We align vector geometries from OSM with high-resolution (sub-meter) aerial imagery and rasterize the polygons into dense label masks. By focusing on high-fidelity aerial data, we ensure that the fine-scale features annotated in OSM (e.g., building outlines, road segments, farmland boundaries) are resolvable in the imagery.
3. **Geographic diversity.** To promote generalization across geographies within the constraints of open high-resolution EO data, SatOSM samples imagery from a diverse set of regions across Europe, prioritized by OSM annotation density and class diversity. We target both urban and rural areas, selecting scenes that include a range of land cover types, infrastructure, and agricultural features. Our dataset currently includes scenes from seven countries—the Netherlands, Finland, Latvia, Estonia, Spain, Switzerland, and Slovakia—selected for their rich and high-quality OSM coverage and availability of high-quality aerial images (Figure 5). The resulting dataset includes over 2,000 unique tags and spans more than 34 million image patches (Appendix C.3).

These principles are operationalized via a scalable data engine that aligns geospatial vector data with high-resolution imagery, extracts open-vocabulary label masks, and standardizes metadata for downstream model training and benchmarking (Appendix C.2).

C.2 Data Curation

OSM Preprocessing We begin by extracting OSM features from its *planet file* [1], a collection of three types of OSM records: nodes (i.e., points), ways (mainly lines and polygons), and relations (a collection of ways). The planet file differs from conventional geospatial data formats (e.g., GeoParquet and GeoPackage) in that it is not organized into a tabular or columnar storage format with each row

[1] `https://wiki.openstreetmap.org/wiki/Planet.osm`

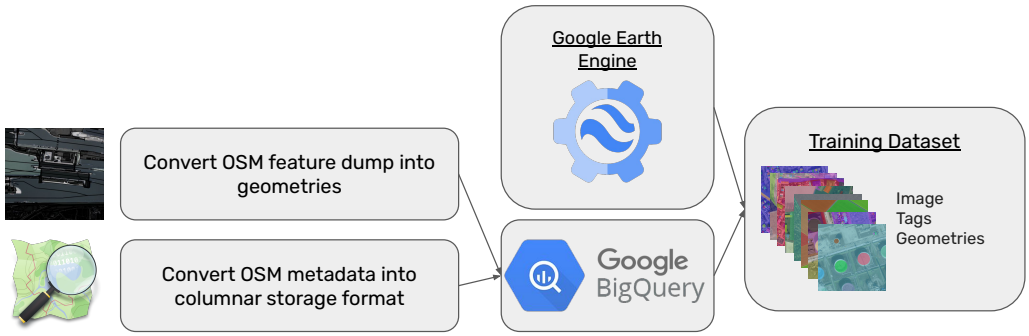

Figure 4: Overview of the data curation pipeline for training-ready OSM datasets. The data pipeline is implemented in a MapReduce fashion [11] with a modular design to suit different data processing requirements. The implementation is compatible with various MapReduce engines such as Apache Spark [55] and Google Cloud Dataflow [7, 14]. The final pipeline can produce `SatOSM` within four hours, discounting Google Earth Engine downloading time.

containing a record. Instead, the OSM planet file stores a series of serialized Protocol Buffers [15] describing the records.

Therefore, to convert the OSM planet dump into a tabular format suitable for analysis and machine learning, we use `osmium-tool` [48] and GDAL [12] to interact with the planet file. These tools reconstruct and extract geometries from the OSM planet dump in a bottom-up fashion: they start by building an index of all the nodes (i.e., points) in the OSM planet file, and then reconstructs the ways (i.e., lines, polygons) referencing nodes and relations (i.e., a collection of ways) that reference ways. Using GDAL and `osmium-tool`, we convert all the metadata into a set of sharded Parquet files, indexed by OSM object ID. In parallel, we convert nodes, ways, and relations into geometries compatible with Open Geospatial Consortium (OGC) standards while discarding those OSM objects that cannot be converted into a valid geometry.

Finally, we ingest the geometries and metadata of node, way, and relation OSM objects into BigQuery to join them as one table ready for analysis (Figure 4).

OSM Filtering Although the free tagging system of OSM can provide us with rich and diverse labels, it also leads to a large number of tags that are not relevant for dense prediction tasks using EO imagery. Some common examples include `source=Bing`, `source=GPS`, and `contact:phone=*`, which represent metadata about the data collection process or contact information of the OSM object. In addition, free tagging can also lead to disputes over the meaning of tags and the conventions of tagging objects. To maintain the richness of the dataset while filtering out noisy or irrelevant tags, we use the tags with well-defined conventions documented in the OSM Wiki [2]. After going through the OSM Wiki, we identified a set of 2219 tags that are reasonable for dense prediction tasks. We filter out the objects that do not contain any of those tags. We note that the majority of OSM objects contain at least one of the tags in our list, as the filtering reduced the total number of global OSM way objects from 1,047,831,606 to 1,032,515,758.

Image Curation We collect high-resolution aerial imagery from open-access sources across Europe, primarily focusing on national mapping agencies that provide orthorectified imagery at sub-meter resolution (Table 3). For each target country, we query available imagery through Google Earth Engine (GEE) [16], prioritizing the most recent collections. Concretely, we form covering grids following the geographical footprint of the available images and obtain the overlapping members of the GEE `ImageCollection`. Then, our data engine requests images from GEE through its `getPixels` API. We carefully choose the spatial extent of each request to achieve the upper bound on GEE's rate limits while respecting the number of concurrent connections, minimizing the number of retries for requests failed due to rate limits (Figure 4). After fetching images, we split large images into smaller chips with the desired shape if needed. We store the processed images in their original

[2] `https://wiki.openstreetmap.org/wiki/Map_features`

spatial resolution and projection and retain their metadata, such as spatial extents, in BigQuery (Figure 4).

Table 3: Summary of image sources in SatOSM. Resolution is in cm/pixel. We report the number of 550×550 pixels tiles extracted from those image sources.

Country/Region	Earth Engine Image Collection	Year	Resolution (cm)	Tile Count
The Netherlands	Netherlands/Beeldmateriaal/LUCHTFOTO_RGB	2022	6 - 8	19,283,311
Switzerland	Switzerland/SWISSIMAGE/orthos/10cm	2017 - 2021	10	10,201,412
Finland	Finland/SMK/V/50cm	2020 - 2023	50	2,079,521
Estonia	Estonia/Maamet/orthos/rgb_low_flying	2020 - 2021	40	1,288,918
Latvia	Latvia/Maamet/orthos/rgb	2016 - 2018	20	1,159,523
Slovakia	Slovakia/orthos/25cm	2019 - 2020	50	385,280
Spain	Spain/PNOA/PNOA10	2018 - 2019	10	331,910

Spatial Join and Rasterization Finally, we perform a spatial inner join between the footprints of processed images and OSM records. We then can obtain a table whose rows contain the metadata of each image chip, along with a list of OSM records collected into a variable-length array. We split this table into training and test sets based on the MD5 hash value of the S2 cell ID of the centroid of the image chip. After splitting, we clip the OSM geometries into the spatial extent of the image chips and rasterize them into binary masks along with the original OSM tags. We perform rasterization using the `rasterio` library [13], producing binary masks with the same spatial resolution as the images. For polygons, we fill all the pixels inside the geometry, while for line strings, we make a buffer of one pixel. In the current version of SatOSM, we discard image chips that do not contain any OSM records. There are 9,072,878 such chips in total. We leave their inclusion for future work, as unlabeled images can still be used for self-supervised pre-training.

Software Library We run the aforementioned data pipeline on Google Cloud Platform (GCP) using Cloud Dataflow [7, 14]. Our pipeline is designed to be modular, enabling easy adoption for different data processing requirements. We note that our implementation is compatible with various batch data processing engines, including Apache Spark [55]. The final pipeline can produce SatOSM within four hours, excluding the time taken to download images from Google Earth Engine (Figure 4).

C.3 Dataset Overview

Dataset Artifacts We deliver the final dataset as a collection of Zstandard-compressed TFRecord [46] and MosaicML MDS [45] shards[3], allowing easy integration into PyTorch and TensorFlow/JAX workflows[4]. After compression, the training set in MDS format comprises 236,324 data shards, totaling 19.46 TiB, while the test set comprises 14,133 data shards, totaling 1.07 TiB. We also provide the user with dataloaders that support streaming access to the dataset hosted on Google Cloud Storage and Hugging Face. As a result, users can perform their training tasks without downloading the full dataset to a local storage system.

Spatial Distributions The spatial distribution of SatOSM is jointly determined by aerial imagery availability and OSM availability. Our dataset currently includes scenes from eight countries—the Netherlands, Finland, Latvia, Estonia, Spain, Switzerland, and Slovakia. Among the eight countries, the Netherlands and Switzerland feature the highest overall annotation density, quantified by the number of primary OSM objects in each S2 cell (Figure 5).

Tag Distribution The semantic diversity of SatOSM stems from the open-ended tagging system of OSM, which we preserve without collapsing tags into a fixed ontology. Our dataset includes annotations with 2,219 unique key-value tags relevant for dense prediction, drawn from a wide spectrum of domains such as buildings, land use, roads, water bodies, and natural features. The distribution of tags is heavily long-tailed: the top 1% of tags account for over 88% of all annotations, while the median tag appears fewer than 100 times. Common tags such as `building=yes` (14M), `landuse=grass` (12.6M), and `highway=track` (5.5M) dominate the dataset, but SatOSM

[3] `https://storage.googleapis.com/deepfried-dd/`

[4] While we make the dataset fully open on GCS, we advise our esteemed reviewers to exercise caution when downloading data larger than 10 GiB from the bucket due to the high networking cost.

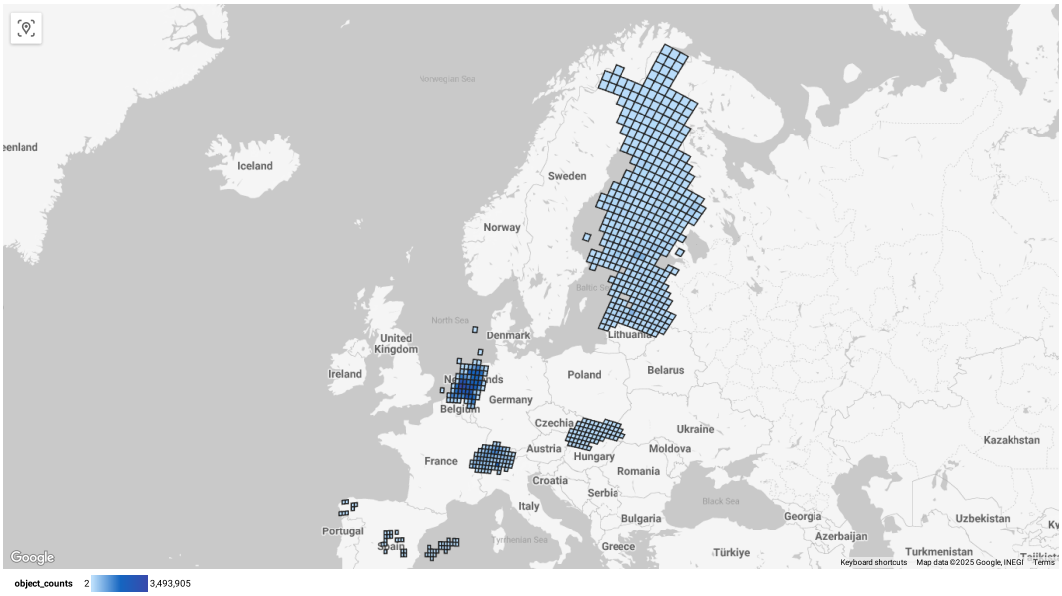

Figure 5: Number of primary OSM objects per level-eight S2 cell (1297.17 km^2 on average) in the study region. The Netherlands and Switzerland have high annotation densities. An interactive version is linked here.

also includes thousands of fine-grained tags such as `natural=heath`, `amenity=parking_space`, and `generator:method=wind_turbine`, each contributing niche but semantically meaningful supervision signals. This imbalance reflects real-world semantic diversity and makes `SatOSM` a valuable testbed for learning from highly imbalanced, open-set label spaces.

D Potential Usage of `SatOSM`

Multi-class Semantic Segmentation Unlike conventional datasets that enforce a one-label-per-pixel constraint, `SatOSM` reflects the real-world complexity of human-labeled geography. OSM annotations often contain layered or nested semantic concepts — for example, a region may be labeled both as `landuse=orchard` and `crop=apple`, or a building may simultaneously be `building=retail` and `shop=bakery`. Our dataloader supports generating one vs. all style labels that can be supervised with binary cross-entropy loss. In addition, we also support generating an empirical distribution over possible tags for each pixel, which can be supervised with a cross-entropy loss whose targets are class probabilities. This compositional labeling reflects how humans perceive and describe places, and provides a fertile testbed for developing models that can disentangle and predict multiple semantic attributes per pixel. `SatOSM` thus enables experimentation with models that are aware of label hierarchy, coexistence, and semantic entanglement.

Table 4: Distribution of common OSM primary tags in `SatOSM`. An interactive version of this table is here.

Geometry Type	Tag	Count	Percent
Polygon	`building=yes`	14,021,314	10.79%
	`landuse=grass`	12,620,427	9.71%
	`landuse=farmland`	9,335,709	7.18%
	`building=house`	8,845,403	6.81%
	`landuse=forest`	8,230,442	6.33%
	`natural=water`	4,473,949	3.44%
LineString	`highway=track`	5,527,953	4.25%
	`highway=service`	3,512,871	2.70%
	`highway=path`	3,091,460	2.38%

Instance & Panoptic Segmentation Each OSM object in SatOSM is uniquely identified with well-defined vector geometries and raster masks, allowing for direct supervision of instance-level segmentation. This is critical for building footprint extraction, road instance segmentation, or counting discrete land parcels. Moreover, because SatOSM annotations can be rendered as semantic (e.g., land cover and land use tags) and instance (e.g., field boundary and building footprint) masks, the dataset supports panoptic segmentation [23], combining pixel-wise classification with object-level delineation.

Compositional Object Retrieval SatOSM 's key advantage over traditional datasets lies in the compositionality of its label space. OSM tags follow an open key-value schema where each object can be described using multiple semantically orthogonal dimensions — e.g., `building=yes`, `shop=supermarket`, `roof:shape=flat`, `opening_hours=24/7`. This allows users to retrieve objects and image regions based on rich semantic queries beyond flat class labels. For example, one can retrieve image patches containing `landuse=farmland` with `crop=rice`, or `building=residential` with `roof:material=metal`. These compositional queries are naturally expressed as conjunctive filters over structured metadata, enabling novel applications in multi-modal and retrieval-augmented vision systems.

E Experiments

Downstream Tasks The AI4Boundaries dataset used in evaluation consists of 7,500 512px images. These images were partitioned into subsets of size 60%/20%/20% of the original dataset for training/-validation/testing respectively. For MiniFrance, samples with no labels were discarded before training. The resulting dataset included over 100,000 512px images using training/validation/testing subsets of size 60%/20%/20%. Lastly, xView2 evaluations included over 9,000 1024px images, which were split into a training and testing sets of portions 90% and 10% of the original dataset respectively.

Our results include evaluations on two versions of SatOSM-Net. Precisely, testing on AI4Boundaries was conducted using a model of similar architecture but trained on a Netherlands-only subset of SatOSM. This model was trained on 4 GH200 GPUs for 56,600 steps with a local batch size of 64. All other evaluations were conducted with the model described in the main text.

Fine Tuning For GU, we apply layer-wise learning-rate decay: the head uses a learning rate of 1e-3; the backbone starts at 1e-4 with decay across deeper layers; weight decay is 0.05 with no-weight-decay groups for norm and bias. For HO and FT, we use AdamW [30] with betas (0.9, 0.999) and a cosine-annealing scheduler [29] over the full training horizon with a minimum learning rate of 1e-5.

Because SatOSM-Net performs best under GU, we also evaluate the baseline GFMs using GU.