M3DRS: Multi-Modal Multispectral Dataset for Remote Sensing

Shanci Li [1]
shanci.li@heig-vd.ch

Antoine Carreaud [1 2]
antoine.carreaud@epfl.ch

Adrien Gressin[1]
adrien.gressin@heig-vd.ch

Abstract

Transformer-based models depend on large datasets for effective image representation learning. A common strategy is to pretrain these models on large-scale data before task-specific adaptation. While acquiring natural imagery is costly and time-consuming, remote sensing offers abundant unlabelled data from open-access sources. To exploit this, we introduce M3DRS—a large-scale unlabelled aerial dataset comprising about 400,000 high-resolution orthophotos from Europe, enriched with near-infrared (NIR) spectral and normalized Digital Surface Model (nDSM) modality. Using self-supervised learning, we examine the impact of multimodal and multispectral information on semantic segmentation, evaluating several pretraining strategies with state-of-the-art deep networks to identify effective training practices for remote sensing models. The dataset and code used in this study are openly available.

1 Introduction

Large transformer-based vision models demand vast datasets to learn generalizable image representations, typically achieved through self-supervised pretraining followed by adaptation to downstream tasks. While large-scale natural image datasets exist, collecting and annotating imagery at very high resolutions remains resource-intensive. In contrast, remote sensing provides abundant unlabelled data from public sources, yet most available datasets focus on medium-resolution satellite imagery. Very high-resolution (10–20 cm) aerial datasets, especially those including additional modalities such as Near-Infrared (NIR) and Digital Surface Models (DSM)—are still scarce. Existing foundation models like DINOv3 [1] can be adapted for aerial imagery, but remain limited to RGB data, leaving a gap for multimodal, high-resolution representation learning. Recent efforts such as Million-AID [2] and fMoW [3] mark progress, yet they do not fully address this high-resolution, multimodal need.

To leverage this potential, we present M3DRS (Multi-Modal Multispectral Dataset for Remote Sensing), a new large-scale dataset comprising approximately 400,000 high-resolution orthophotos from European regions. Each image includes additional near-infrared (NIR) and normalized Digital Surface Model (nDSM) layers, enabling multimodal learning across spectral and spatial domains.

We employ self-supervised learning techniques [4] [5] [6] to assess the contribution of these additional modalities to downstream semantic segmentation performance. Multiple pretraining strategies, including unimodal RGB, multispectral (RGB + NIR), and multimodal (RGB + NIR + nDSM) configurations, are evaluated using state-of-the-art transformer-based architectures. Our experiments highlight how cross-modal fusion during pretraining enhances representation robustness and downstream segmentation accuracy, particularly in challenging land-cover classes such as vegetation and urban surfaces.

[1]University of Applied Sciences Western Switzerland (HES-SO / HEIG-VD), 1400 Yverdon-les-Bains, Switzerland

[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), 1010 Lausanne, Switzerland

2 Dataset

2.1 Sources and Statistics

The M3DRS dataset consolidates large-scale, unlabelled multi-modal remote sensing imagery from open-access sources in Switzerland, France, and Italy. It includes high-resolution RGB-NIR orthophotos and nDSM derived from LiDAR or elevation models, collectively covering 3,077 km^2. Each image measures 512×512 pixels with spatial resolutions between 10 and 25 cm.

Table 1: Composition and statistics of M3DRS dataset.

Data Source	Country	Area (km^2)	Resolution	No. of images	Size (GB)
Swisstopo [7] [8]	Switzerland	2,172	10/25 cm	282,243	346
Ferrara City [9] [10] [11]	Italy	95	10 cm	39,907	49
FLAIR #1 [12]	France	810	20 cm	77,762	96
Sum		3,077		399,912	491

The dataset spans diverse geographic and seasonal conditions, enhancing variability in vegetation, land cover, and lighting. Swiss imagery dominates the dataset, with sampling balanced according to national land-cover statistics [13]. French data originate from the FLAIR #1 dataset, containing RGB-NIR orthophotos and DSMs from IGN's ORTHO HR® and RGE ALTI DTM products [14, 15]. Italian data from Ferrara provide complementary urban and peri-urban samples at 10 cm resolution.

All aerial surveys were conducted between March and November, primarily from May to September, introducing natural seasonal variations.

2.2 Data Generation Workflow

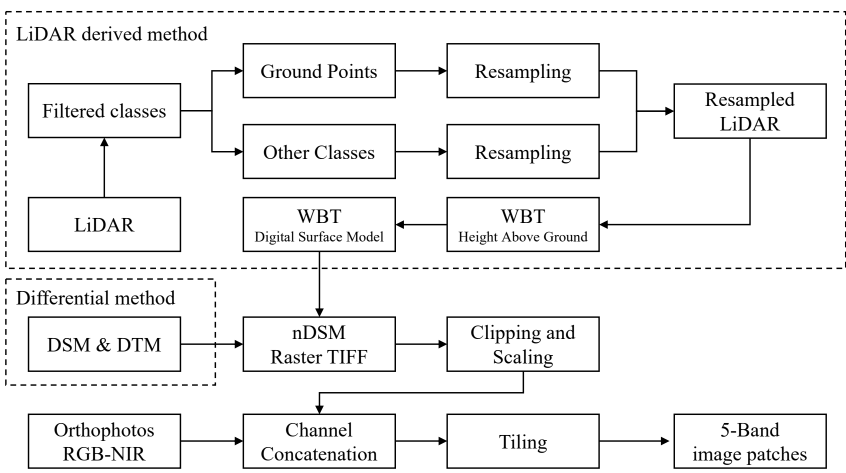

Figure 1: Workflow to generate 5-band images with classified LiDAR and orthophotos.

For each region, nDSM layers were derived by combining DSM and DTM or directly from classified LiDAR point clouds using PDAL [16] and WhiteboxTools [17], as shown in Figure 1. Only points with LAS classification codes between 2 and 17 were retained to ensure reliability. LiDAR data (averaging 15-20 $points/m^2$) were resampled to match orthophoto resolution.

Ground points (class 2) and off-ground points were rasterized separately, then combined using height-above-ground estimation to form nDSM rasters. Following FLAIR [12], 32-bit float values were scaled by a factor of 5 and encoded as 8-bit integers, preserving height information up to 51 m with 0.2 m resolution.

The final 5-band tiles (RGB, NIR, nDSM) were generated as 512×512 GeoTIFFs. The workflow ensures consistent alignment across modalities and scalability for future open-data integration.

3 Benchmark

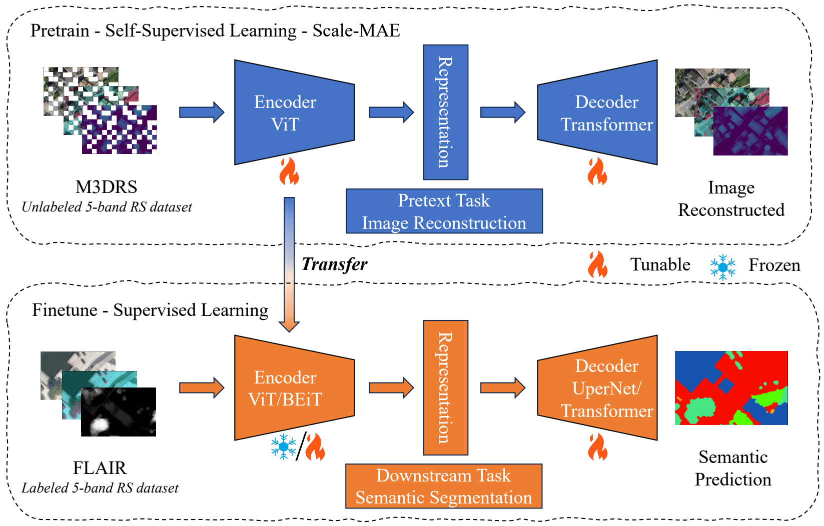

Figure 2: Experimental workflow.

To evaluate the benefits of multi-spectral and multi-modal inputs, we conducted experiments using transformer-based foundation models pretrained on M3DRS. Self-supervised learning (SSL) enables effective utilization of large-scale unlabelled datasets. Table 4 in **Appendix** summarize some representative models and their features. Among state-of-the-art methods, Scale-MAE [6], with Ground Sample Distance based positional encoding and multiscale features, was selected for pretraining due to its superior performance [18].

Our experimental pipeline (Figure 2) uses a ViT encoder pretrained via Scale-MAE to reconstruct masked image patches. The encoder is then paired with a UperNet decoder for semantic segmentation and fine-tuned on the labelled FLAIR dataset. For comparison, we also implemented ViT-Adapter [19] with BEiT backbone and Masked Attention Transformer decoder (Mask2Former [20]). Both 3-band and 5-band ViT-Adapter variants were evaluated.

3.1 Pre-training

Scale-MAE was adapted for 5-band inputs by initializing RGB channels with weights pretrained on FMoW-RGB [3]. The encoder was trained on M3DRS dataset for 600 epochs. Reconstruction results confirm effective feature learning for both optical and additional channels.

3.2 Fine-tuning

The FLAIR dataset with 5-band imagery and semantic masks was used to fine-tune both encoder and decoder. Table 2 summarizes the results.

Table 2: Segmentation performance (mIoU) on FLAIR dataset.

Model	CNN	ViT-UperNet	ViT-Adapter
Bands	5	5	3
Backbone	ResNet34	ViT-L	BEiT-L
Decoder	U-Net	UperNet	Mask2Former
Pretraining	Supervised	ScaleMAE	ImageNet + SL
mIoU	55.70	62.15	62.80

CNN model (ResNet + U-Net) performs 7% worse than ViT-based models. ViT-UperNet pretrained on M3DRS approaches but does not surpass ViT-Adapter, which benefits from supervised ImageNet pretraining, multi-scale understanding, and advanced decoder architecture. Limitations of BEiT tokenizers prevented 5-band ViT-Adapter pretraining on M3DRS; to analyze the impact of additional spectral channels, ablation studies were conducted on plain ViT-based models.

All experiments were executed on a single node with Intel Xeon Silver 4310 CPU, 256 GB RAM, and 4 NVIDIA A40 GPUs.

3.3 Ablation Study

Table 3: Ablation study of pretraining dataset and bands.

Bands	Method	Dataset	mIoU
RGB	-	-	53.73
RGB	MIM + SL	ImageNet	60.52
RGB	Scale-MAE	M3DRS	60.54
RGB	Scale-MAE	fMoW-RGB	60.61
RGB + NIR	Scale-MAE	M3DRS	61.52
RGB + nDSM	Scale-MAE	M3DRS	60.87
RGB + NIR + nDSM	-	-	53.86
RGB + NIR + nDSM	MIM + SL	ImageNet	61.58
RGB + NIR + nDSM	Scale-MAE	M3DRS	**62.15**

Table 3 presents the performance of foundation models pretrained on different datasets and spectral configurations. The 5-band model pretrained with the M3DRS dataset achieved the highest mIoU, confirming the benefit of incorporating multi-spectral and multi-modal information during pretraining.

Without pretraining, 3-band and 5-band models performed similarly, indicating that additional NIR and nDSM channels offer limited improvement under purely supervised learning on the FLAIR dataset. This aligns with the ViT-Adapter's stronger results on RGB inputs, suggesting that much of the discriminative information in the extra channels overlaps with RGB.

Pretraining substantially improved performance for all 3-band models, by at least 6.73% mIoU, regardless of dataset or method, emphasizing its importance. The comparable outcomes between fMoW-RGB and M3DRS indicate that large-scale pretraining enhances general feature extraction, while naive autoencoder-based architecture remains largely insensitive to geographic context.

For 5-band models, M3DRS pretraining yielded only a modest 0.57% gain over natural imagery, as ImageNet-style models lack exposure to NIR and nDSM modalities. Analysis of 4-band variants shows NIR improving mIoU by 0.98% and nDSM by 0.33%, showing less benefits comes from nDSM. Overall, NIR contributes more consistently, while leveraging structural data like nDSM requires careful alignment and large-scale domain-specific pretraining.

4 Conclusion and Outlooks

We introduced M3DRS as a new large-scale unlabelled aerial dataset featuring high-resolution multi-spectral and multi-modal data from European regions. Its integration of NIR and nDSM modalities facilitates multimodal representation learning and provides a foundation for further exploration of self-supervised strategies in remote sensing. Pretrained baseline models show that while RSFMs hold promise, their advantage over natural image foundation models remains moderate without advanced decoder designs or domain-specific objectives.

Future work should focus on improving multi-modal fusion architectures and pretraining objectives that explicitly capture spatial, spectral, and geometric relationships. The superior performance of ViT-Adapter compared to ViT-UperNet underscores the importance of decoder design in transferring pretrained representations to downstream tasks. Additionally, leveraging large-scale time-series datasets and unsupervised geo-context learning [21] could enhance geo-awareness and temporal reasoning. While constructing such datasets remains resource-intensive, the ongoing expansion of open-access Earth observation archives makes scalable, multimodal pretraining increasingly feasible. Ultimately, M3DRS contributes toward building robust and domain-adaptive foundation models for Earth observation and environmental monitoring.

A Appendix

Table 4: A summary of state-of-the-art RSFMs and natural FMs.

Type	Model	Backbone	Pretained Dataset	Model Parameter	GPU	GPU Hours	Features
Natural FMs	Swin	Swin-B Swin-L	ImageNet (150G)	88 M 197 M	V100	1.4 / 9 every epoch	Swin Transformer block;
	I-JEPA	ViT-H14	ImageNet (150G)	632 M	16 * A100 80G	1152	Training Efficient 5.3x faster than MAE
	MAE	ViT-L	ImageNet (150G)	307M	64 * V100	2688	Masked Autoencoders
	SAM	ViT-H/16	SA-1B (11M images / 1B masks)	632 M	256 * A100	17408	Large-scale training dataset; Prompt encoder; Contextual understanding; Long-range dependencies modeling
RSFMs	RVSA	ViT-B	MillionAID	86 M	8 * A100 80G	-	plain ViTs; Rotated varied-size window attention;
	SatMAE	ViT-L	fMoW-RGB (200G)	307 M	8 * V100 16G	960	Temporal Encoding multi-spectral RS image input
	ScaleMAE	ViT-L	fMoW-RGB (200G)	307 M	8 * V100 16G	960	GSD Positional Encoding; Super-resolution; Multiscale Features
	Cross-Scale MAE	ViT-B ViT-L	fMoW-RGB (200G)	86.6 M 304.4 M	A6000	-	Multi-Scale Augmentation; Cross-Scale Information Consistency; Contrastive learning; GSD Positional Encoding
	Multi-MAE	ViT-B	ImageNet (150G)	88 M	8 * A100 80G	1280	Multi-task; cross-modality: depth and semantic; pseudo labeling
	SkySense	ViT-L + Swin-H	SkySense (300TB)	2.06 B	80 * A100 80G	24600	Geo-Context Prototype Learning; Multi-Modal; Multi-spectral; Multi-Granularity Contrastive Learning

Acknowledgments and Disclosure of Funding

This project was made possible thanks to a fruitful collaboration between HEIG-VD and the Swiss Territorial Data Lab (STDL). Funding was provided by STDL through the (Swiss Geoinformation Strategy) [22].

References

[1] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[2] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4205–4230, 2021.

[3] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[4] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[6] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor 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.

[7] Swisstopo. swisssurface3d, Jan 2024. URL `https://www.swisstopo.admin.ch/en/height-model-swisssurface3d`. The classified point cloud of Switzerland.

[8] Federal Office of Topography. Swissimage rs, 2024. URL `https://www.swisstopo.admin.ch/fr/orthophotos-swissimage-rs`.

[9] City of Ferrara. Ortofoto 2022 area urbana, 2024. URL `https://dati.comune.fe.it/dataset/ortofofo2022`. License: CC BY 4.0.

[10] City of Ferrara. Modello digitale delle superfici (dsm) 2022, 2024. URL `https://dati.comune.fe.it/en/dataset/dsm-2022`. License: CC BY 4.0.

[11] City of Ferrara. Modello digitale terreno (dtm) 2022, 2024. URL `https://dati.comune.fe.it/en/dataset/dtm-2022`. License: CC BY 4.0.

[12] Anatol Garioud, Stéphane Peillet, Eva Bookjans, Sébastien Giordano, and Boris Wattrelos. Flair# 1: semantic segmentation and domain adaptation dataset. *arXiv preprint arXiv:2211.12979*, 2022. License: Apache 2.0.

[13] Section Geoinformation. Swiss land use statistics, 2024. URL `https://www.bfs.admin.ch/bfs/en/home/services/geostat/swiss-federal-statistics-geodata/land-use-cover-suitability/swiss-land-use-statistics.html`.

[14] IGN. Bd ortho®, 2024. URL `https://geoservices.ign.fr/bdortho`.

[15] IGN. Rge alti®, 2024. URL `https://geoservices.ign.fr/rgealti`.

[16] PDAL Contributors. Pdal point data abstraction library, August 2022.

[17] Dr. John B. Lindsay. Whiteboxtools version 2.3.0, March 25 2023.

[18] Licheng Jiao, Zhongjian Huang, Xiaoqiang Lu, Xu Liu, Yuting Yang, Jiaxuan Zhao, Jinyue Zhang, Biao Hou, Shuyuan Yang, Fang Liu, et al. Brain-inspired remote sensing foundation models and open problems: A comprehensive survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[19] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[20] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[21] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *arXiv preprint arXiv:2312.10115*, 2023.

[22] swisstopo. Swiss geoinformation strategy, 2024. URL `https://www.geo.admin.ch/fr/strategie-et-mise-en-oeuvre`.