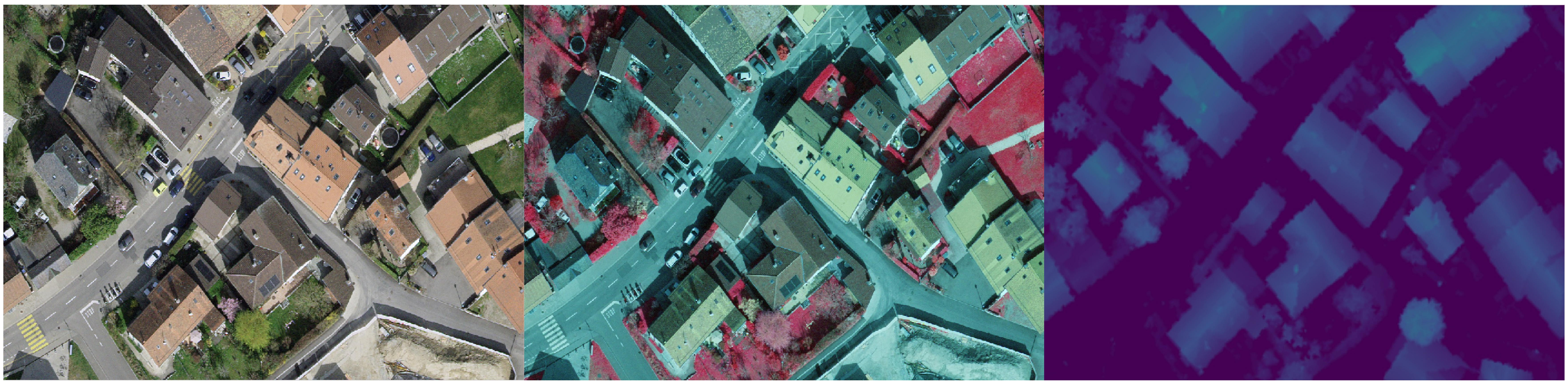
M3DRS: Multi-Modal Multispectral Dataset for Remote Sensing

Shanci Li [1], Antoine Carreaud [1,2], Adrien Gressin [1]
[1]University of Applied Sciences Western Switzerland (HES-SO / HEIG-VD)
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL)



RGB **NIR-R-G** **Normalized DSM**

Introduction

This study introduces M3DRS, a large-scale multimodal dataset designed to advance self-supervised representation learning in remote sensing. Integrating high-resolution RGB-NIR-nDSM orthophotos across Europe, it enables the development of transformer-based models that jointly learn from spectral, spatial, and geometric cues.

Motivation

❑ Foundation models in computer vision rely on massive datasets and self-supervised pretraining.
❑ Remote sensing offers abundant open-access imagery, yet most datasets are medium-resolution and spectral-only.
❑ There is a lack of **high-resolution multimodal** datasets that include spectral (NIR) and geometric (nDSM) information for aerial imagery.

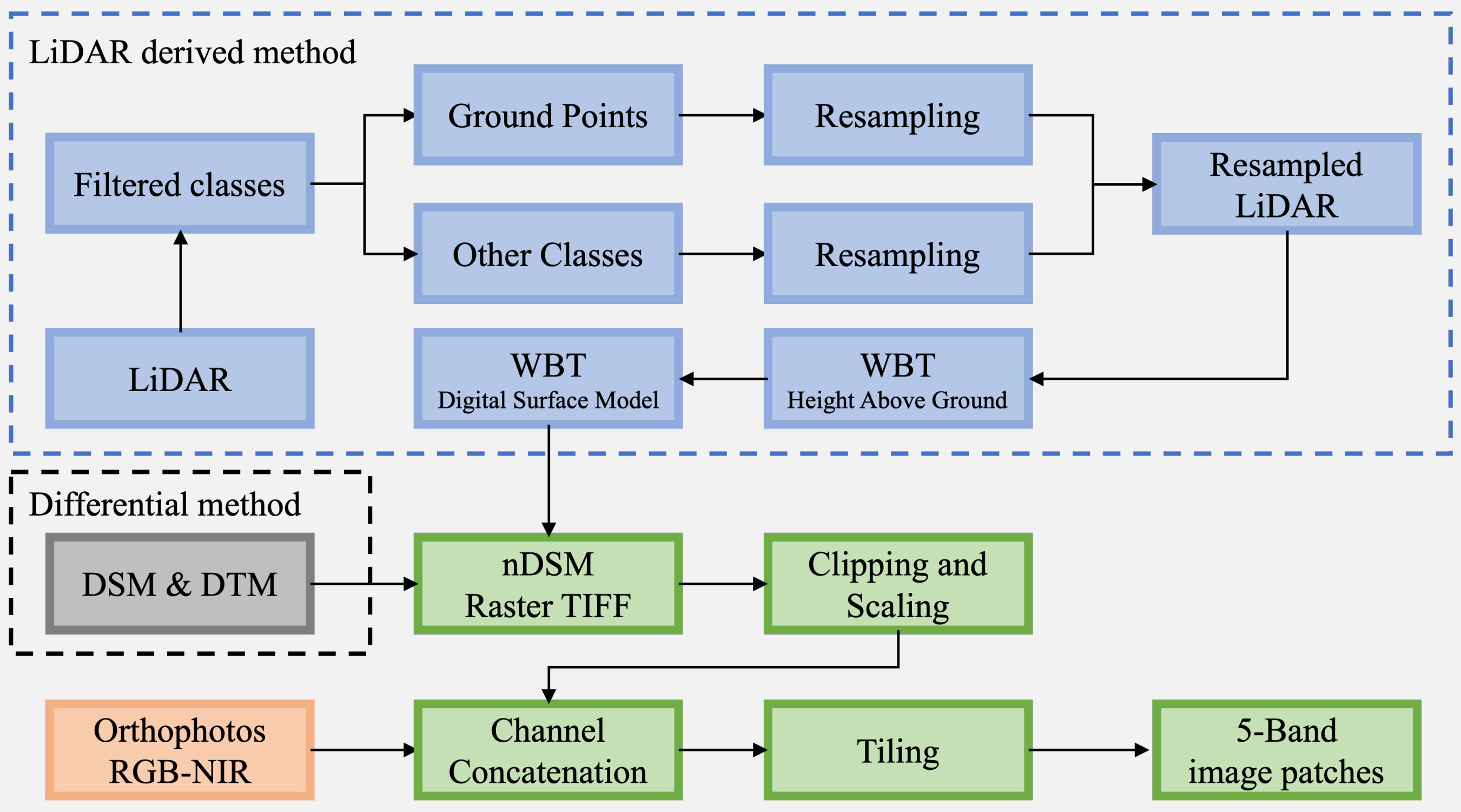
Contributions

❑ ~400,000 unlabeled high-resolution (10–25 cm) orthophotos across Europe.
❑ Five bands: RGB + NIR + normalized Digital Surface Model (nDSM).
❑ Covers 3,077 km² across Switzerland, France, and Italy.
❑ Enables **self-supervised** multimodal pretraining for transformer models.

Dataset

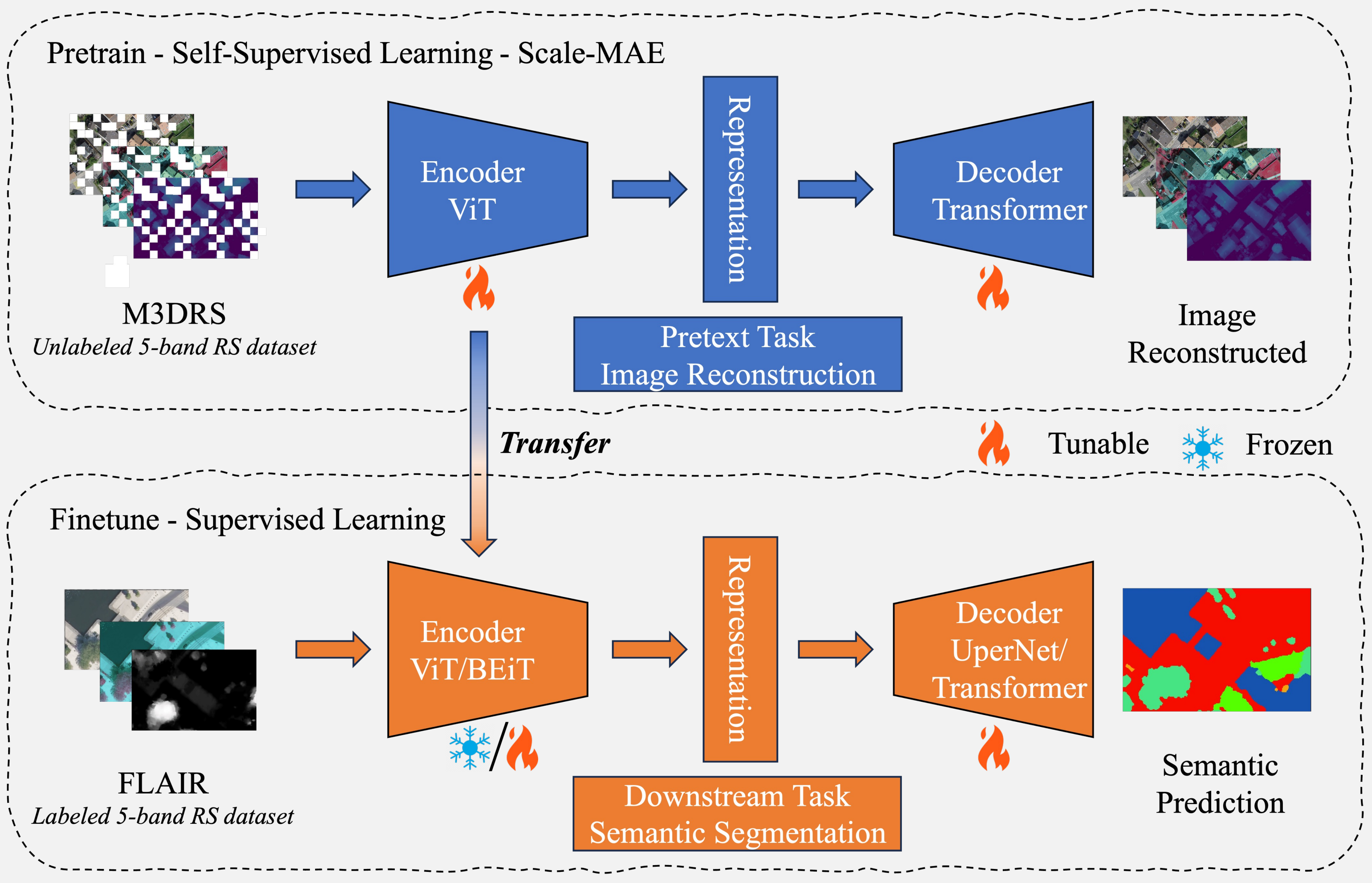
Dataset Composition:
diverse landscapes, lighting, and seasonal conditions (Mar–Nov).

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



Benchmark & Conclusion

Objective
❑ Evaluate effect of multimodal multispectral input on semantic segmentation.
Technical Specs
❑ 4 × NVIDIA A40 (48 GB)
Pretraining
❑ Scale-MAE with GSD-aware positional encoding.
Fine-tuning
❑ Transfer ViT encoder to FLAIR dataset using ViT-UperNet or ViT-Adapter (BEiT backbone + Mask2Former).


Pretrain - Self-Supervised Learning - Scale-MAE
M3DRS
Unlabeled 5-band RS dataset
Encoder ViT
Representation
Decoder Transformer
Image Reconstructed
Pretext Task Image Reconstruction
Transfer
Tunable
Frozen
Finetune - Supervised Learning
FLAIR
Labeled 5-band RS dataset
Encoder ViT/BEiT
Representation
Decoder UperNet/ Transformer
Semantic Prediction
Downstream Task Semantic Segmentation

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

Ablation Study: 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**

Conclusion

❑ Multi-modal pretraining enhances **robustness** and **performance**.
❑ NIR contributes most reliably to semantic distinction, while nDSM adds spatial depth information when properly **fused**.
❑ **Decoder** architecture remains critical for downstream performance.

Limitations & Future Work

❑ Current dataset lacks text modalities and temporal image sequence, limiting multi-source pretraining.
❑ Further improvement may come from better geometric fusion techniques and temporal or multi-seasonal data to capture dynamic surface changes..

Source Code: https://github.com/swiss-territorial-data-lab/proj-vit
Dataset: https://huggingface.co/datasets/heig-vd-geo/M3DRS
Acknowledgement: Funded by STDL - Swiss Geoinformation Strategy
Contact: adrien.gressin@heig-vd.ch

EurIPS HEIG-VD SCHOOL OF ENGINEERING AND MANAGEMENT EPFL Swiss Territorial Data Lab