2023 Dragon 5 Symposium

# Multi-Modal Deep Learning for Multi-Temporal Urban Mapping with a Partly Missing Modality

*Sebastian Hafner and Yifang Ban*

Division of Geoinformatics, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden

Paper & Dataset

## Introduction

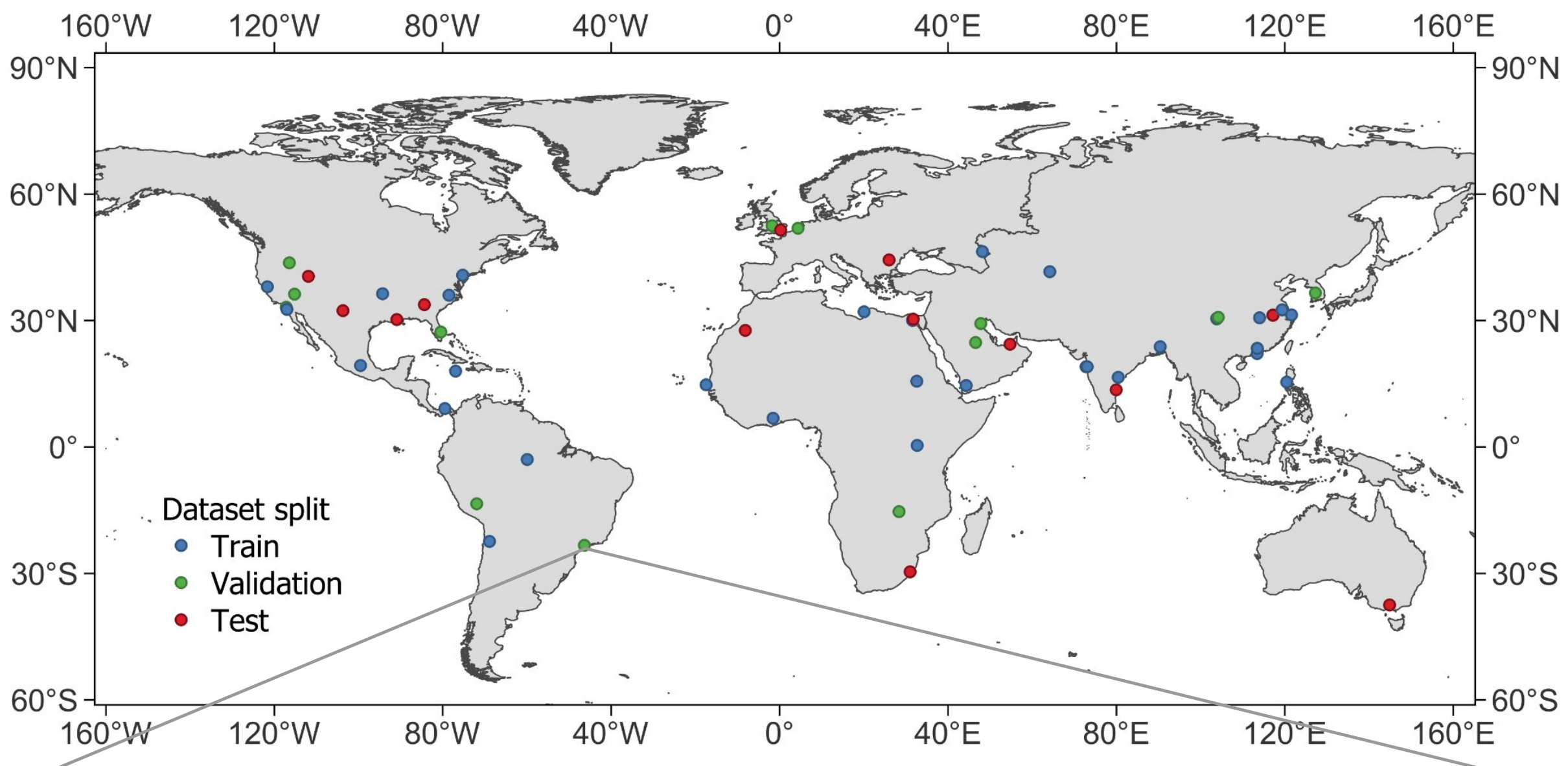
Urban mapping frameworks have made significant advancements through the integration of deep learning techniques and multi-modal Earth observation data from Synthetic Aperture Radar (SAR) and optical sensors (e.g., [1]). However, the presence of clouds often leads to partial loss of optical data, which poses challenges for multi-temporal urban mapping (see Fig. 2).

In this study, we present a novel approach that leverages SAR data to predict missing optical features, addressing the limitations of the optical modality. Our proposed method has been evaluated on a multi-temporal urban mapping dataset, utilizing Sentinel-1 SAR and Sentinel-2 MSI data.

## Study Areas and Satellite Data

We created a multi-temporal urban mapping dataset by enriching the SpaceNet 7 dataset [2] with data from the Sentinel-1 SAR and Sentinel-2 MSI missions. The SpaceNet 7 dataset includes monthly Planet composites acquired between 2017 and 2020, along with manually annotated building footprints. Data from the Sentinel missions were preprocessed in Google Earth Engine, resulting in time series of data triplets for 60 sites (Fig. 1). The triplets (Fig. 2) were resampled to match the 4-meter resolution of the SpaceNet 7 dataset for consistency and comparability.

Figure 1: Location of the 60 distinct study sites spread across the globe. The sites were split into 41 training, 15 validation and 14 test sites.



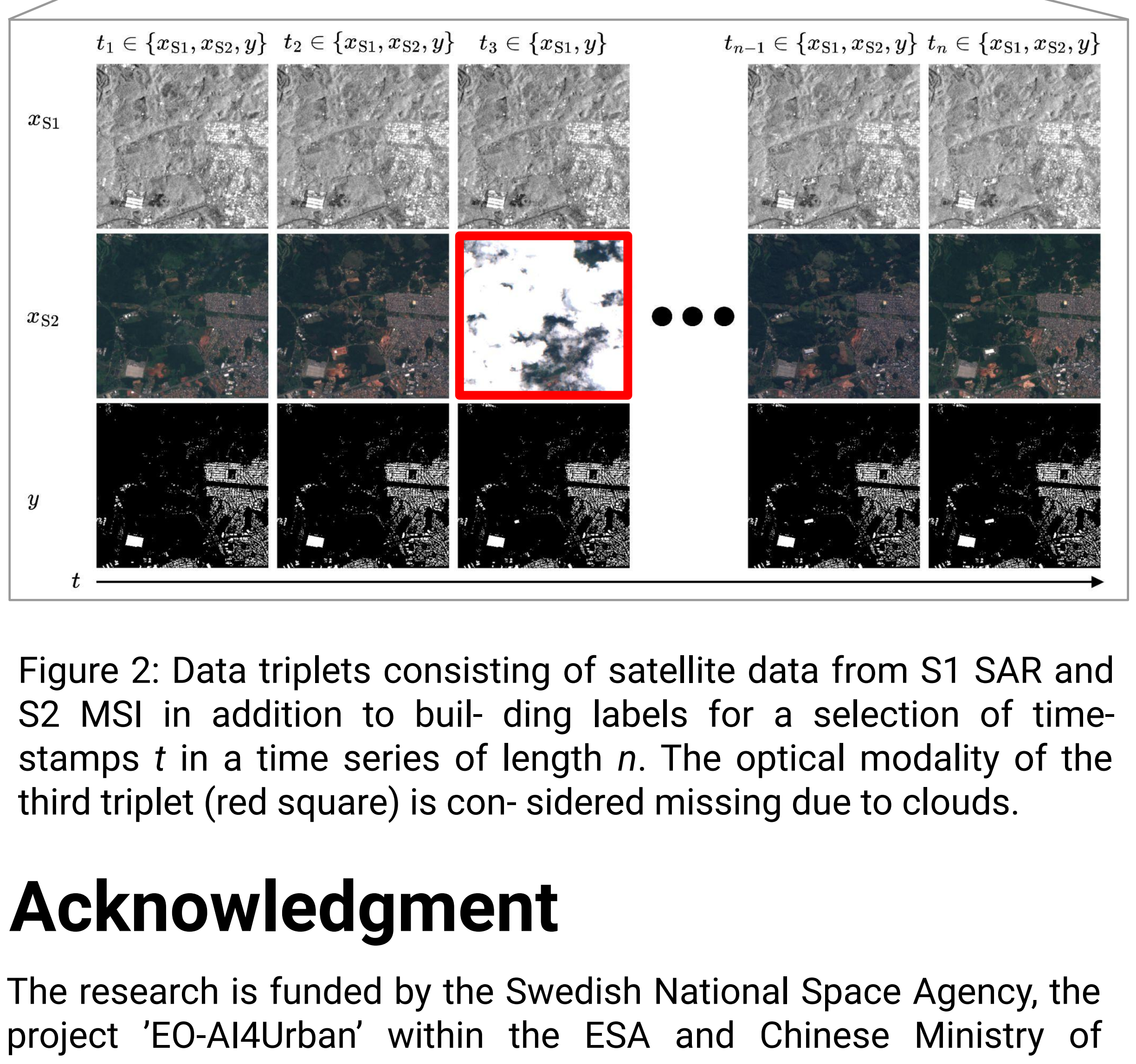

$t_1 \in \{x_{S1}, x_{S2}, y\}$
$t_2 \in \{x_{S1}, x_{S2}, y\}$
$t_3 \in \{x_{S1}, y\}$
$t_{n-1} \in \{x_{S1}, x_{S2}, y\}$
$t_n \in \{x_{S1}, x_{S2}, y\}$
$x_{S1}$
$x_{S2}$
$y$
$t$

Figure 2: Data triplets consisting of satellite data from S1 SAR and S2 MSI in addition to buil- ding labels for a selection of time-stamps $t$ in a time series of length $n$. The optical modality of the third triplet (red square) is con- sidered missing due to clouds.

## Acknowledgment

The research is funded by the Swedish National Space Agency, the project 'EO-AI4Urban' within the ESA and Chinese Ministry of Science and Technology's Dragon 5 Program, Digital Futures within the project 'EOAI4GlobalChange', and the EU Horizon 2020 project 'HARMONIA'.

## Proposed Approach

The proposed model consists of two networks to extract features from the SAR and optical input separately, in addition to a third network that reconstructs the optical features from the SAR input in order to cope with a missing optical modality.

To train the model and perform inference, two cases are considered: (1) the multi-modal case where both the S1 image and S2 image are available (see (a & c), and (2) the missing modality case where only the S1 image is available (b & d).

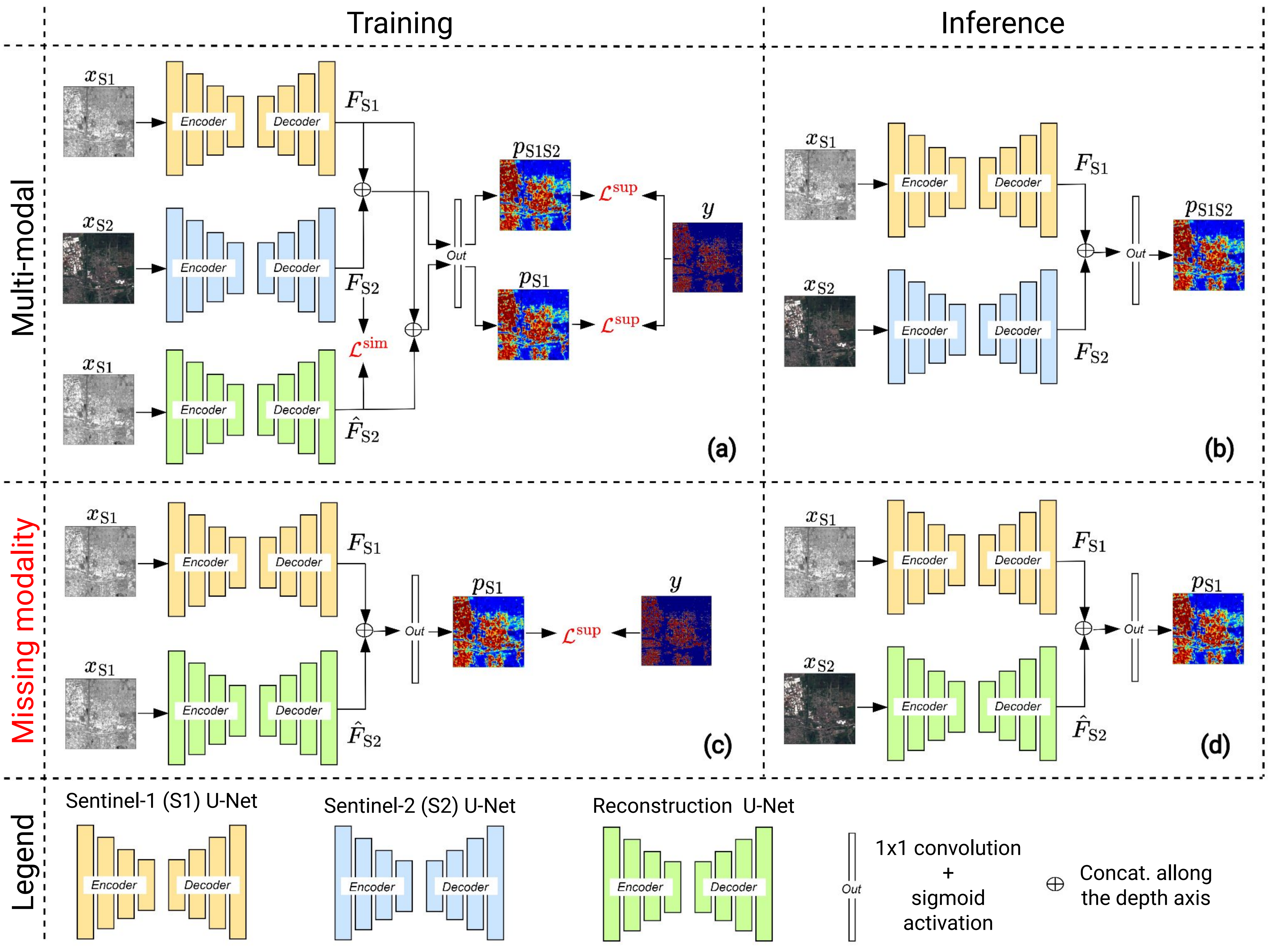


Figure 3: Overview of the proposed approach for multi-modal urban extraction with a partly missing optical modality.

## Results

- Comparison of three models: U-Net S1, Dual Stream (DS) U-Net [1], and the proposed model.
- The proposed model achieves the highest accuracy values: we outperform the DS U-Net and U-Net on multi-modal sample, while also performing better than DS U-Net on samples with a missing optical modality.
- Limitation: we fall short of U-Net on the missing modality samples.

Table 1: Quantitative test results in terms of Inter- section over Union (IoU) and F1 score. Values represent μ±σ of 5 runs. The highest mean values are boldfaced.

| Method | All samples | | Multi-modal samples | | Missing modality samples | |
|---|---|---|---|---|---|---|
| | F1 ↑ | IoU ↑ | F1 ↑ | IoU ↑ | F1 ↑ | IoU ↑ |
| U-Net S1 | 0.362 ± 0.005 | 0.221 ± 0.003 | 0.364 ± 0.005 | 0.222 ± 0.004 | **0.348 ± 0.004** | **0.210 ± 0.003** |
| DS U-Net | 0.411 ± 0.008 | 0.259 ± 0.006 | 0.426 ± 0.008 | 0.271 ± 0.006 | 0.233 ± 0.033 | 0.132 ± 0.021 |
| Proposed | **0.423 ± 0.006** | **0.269 ± 0.005** | **0.435 ± 0.006** | **0.278 ± 0.005** | 0.327 ± 0.008 | 0.195 ± 0.006 |

## Conclusion

This poster addressed the problem of a partly missing optical modality during inference time for multi-modal urban mapping from S1 and S2 data. To that end, a model that utilizes a reconstruction network to approximate the features of the optical modality when its missing was proposed.

The model achieved improved performance over a uni-modal approach trained on S1 SAR data and a multi-modal approach using zero values in case of a missing optical modality. Despite these improvements, our findings indicate that further research is needed to improve the performance of missing modality samples.

## References

[1] Hafner, S., Ban, Y. and Nascetti, A., 2022. Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data. Remote Sensing of Environment, 280, p.113192.

[2] Van Etten, A., Hogan, D., Manso, J.M., Shermeyer, J., Weir, N. and Lewis, R., 2021. The multi-temporal urban development spacenet dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6398-6407).