51. Multimodal Deep Learning for Alpine Land Cover: Benchmarking Foundation Encoders for Very-High-Resolution Multi-Sensor LULC Classification

Paolo Dabove, Deepak Sairam Madhusudhana Rao, Stefano Gamba, Luca Olivotto, Alessandro Zolli, Ludovico Pividori, Umberto Morra di Cella, Gianluca Filippa

1 Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Italy; 2 Geomatics & Data Science Division, DigiSky s.r.l., Turin, Italy; 3 ARPA Valle d'Aosta, Italy

Mapping land cover in mid- to high-altitude terrain of the European Alps is difficult because the classes environmental agencies monitor, among them dwarf shrub, wetland, exposed rock and seasonal snow, form mosaics only a few metres across and fall below what tenmetre satellite products resolve. Sub-metre airborne multi-sensor survey resolves them, at the cost of a processing chain whose components are seldom weighed against one another. Seven configurations of a weakly supervised network were compared. The network fuses airborne RGB, six-band multispectral imagery, spectral indices and a photogrammetric surface model into a fourteen-class product nested in the CORINE hierarchy. Only the image encoder differs between configurations; the sensor streams, the fusion stage and the training recipe are held fixed. The payload itself is assessed on the same imagery using standard radiometric and spatial quality measures. The best configuration maps the fourteen classes at a mean Intersectionover-Union of 0.727 on held-out tiles, with snow, water bodies, roads and buildings above 0.85 and dwarf shrub below 0.30. Under a fixed and equal training budget, substituting the image encoder, over a range from 28 to 636 million parameters and three architectural families, moves that mean by less than the variation the pipeline shows between consecutive training epochs. Of the six spectral bands, the three visible ones carry nearly the same information, leaving about three independent channels. The quality of the resulting land-cover product is therefore governed by the sensor payload and the annotation rather than by the size of the image encoder.