POLITECNICO DI TORINO
Repository ISTITUZIONALE

A Multi-Modal Deep Learning Framework for High-Resolution Alpine Land Use/Land Cover

Original
A Multi-Modal Deep Learning Framework for High-Resolution Alpine Land Use/Land Cover / Dabove, P., Rao, D.S.M., Olivotto, L., Pividori, L., Filippa, G., Morra Di Cella, U.. - In: REMOTE SENSING. - ISSN 2072-4292. - STAMPA. - 18:18(2026). [10.3390/rs18183203]

Availability:
This version is available at: 11583/3017582 since: 2026-11-05T11:58:58Z

Publisher:
MDPI

Published
DOI:10.3390/rs18183203

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

12 November 2026

remote sensing

MDPI

Article

A Multi-Modal Deep Learning Framework for High-Resolution Alpine Land Use/Land Cover

Paolo Dabove [1,*,†], Deepak Sairam Madhusudhana Rao [2,†], Luca Olivotto [2], Ludovico Pividori [2], Gianluca Filippa [3] and Umberto Morra di Cella [3]

[1] Department of Environment, Land and Infrastructure Engineering (DIATI), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy
[2] Geomatics & Data Science Division, DigiSky s.r.l., Strada della Berlia 500, 10146 Turin, Italy; deepak@digisky.it (D.S.M.R.); olivotto@digisky.it (L.O.); pividori@digisky.it (L.P.)
[3] Agenzia Regionale Protezione Ambiente Valle D'Aosta—ARPA Valle D'Aosta, Loc. La Maladière 48, Saint-Christophe (AO), 11020 Aosta, Italy; g.filippa@arpa.vda.it (G.F.); u.morradicella@arpa.vda.it (U.M.d.C.)
* Correspondence: paolo.dabove@polito.it; Tel.: +39-011-090-7662
† These authors contributed equally to this work.

Highlights

What are the main findings?

- The proposed weakly supervised multimodal framework, trained using only 2517 sparsely distributed annotation polygons and without requiring exhaustive wall-to-wall pixel annotations, achieved a validation mIoU of 0.777 across fourteen alpine land-cover classes.
- The ablation experiments identified the multispectral stream as the principal contributor to classification performance in the investigated dataset. Removing this stream reduced the validation mIoU by more than 35 percentage points, whereas removing the normalised Digital Surface Model (nDSMs) reduced it by fewer than 5 percentage points. These results indicate that spectral information provided substantially greater discriminative value than topographic information under the adopted weak-supervision setting.

What are the implications of the main finding?

- Learning from sparsely distributed annotation polygons substantially reduces the need for exhaustive pixel-wise labelling. This provides environmental agencies with a practical and scalable approach to producing high-resolution alpine land-cover maps, while retaining the need for expert involvement in defining the classification scheme, validating annotations, and reviewing the resulting maps.
- For weakly supervised alpine mapping, spectral breadth appears to matter more than backbone scale: the principal residual failure, Wetlands, is tentatively attributed to a moisture signal in the shortwave infrared that a six-band payload cannot acquire.

Abstract

Accurate Land Use and Land Cover (LULC) mapping in high-resolution alpine environments is challenging due to complex terrain, heterogeneous vegetation, seasonal snow and ice cover, and the limited spectral information provided by conventional aerial imagery. Although foundation models such as the Segment Anything Model (SAM) effectively capture structural features, their class-agnostic design, limits fine-grained semantic discrimination and typically requires large annotated datasets. This study proposes a multi-modal deep learning framework for alpine LULC mapping using sparse annotations, which would fall under the category of weakly supervised learning. The framework employs a dual-encoder

Check for updates

Academic Editors: Fei Xiao, Yonghui Yao and Wenzuo Zhou

Received: 27 July 2026
Revised: 5 September 2026
Accepted: 8 September 2026
Published: 17 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

architecture that integrates RGB imagery, six-band multispectral imagery, and custom adapters for spectral indices, and Digital Surface Models (DSMs). A SAM-based encoder extracts geometric and contextual features from RGB imagery, while a dedicated encoder learns complementary spectral representations from multispectral data. To address the boundary uncertainty introduced by sparse supervision, we propose post inference hybrid refinement strategy that combines a Canopy Height Model (CHM) derived from the DSM to improve tree crown delineation with edge-based refinement for low vegetation classes, such as shrubs, and mathematical methods to refine road and building edge delineation with DSMs. Experimental results across fourteen alpine classes highlight that the framework achieves a validation mIoU of 0.777, macro-averaged over the fourteen classes. For the present sensor choice and classification scheme, no comparable multi-modal baseline exists. Thus, results are reported in absolute terms. The present multi-modal data fusion sets the baseline for future scalable alpine LULC mapping applications.

Keywords: deep learning; semantic segmentation; multi-modal fusion; aerial multispectral imagery; aerial photogrammetry; remote sensing; Segment Anything Model (SAM); cryosphere; alpine environment; Digital Surface Model (DSM); Canopy Height Model (CHM)

1. Introduction

Alpine environments serve as highly sensitive indicators of global climate change. Monitoring Land Use and Land Cover (LULC) dynamics in these high-altitude regions is paramount for understanding shifting hydrological cycles, assessing biodiversity, and mitigating natural hazard risks [1]. As rapid environmental changes transform these landscapes at an unprecedented pace, the demand for accurate, scalable, and temporally consistent mapping techniques has become a critical priority within the Earth observation community and for the generation of information useful to land managers and policymakers.

Historically, LULC classification has relied heavily on medium-to-high resolution satellite imagery paired with traditional machine learning algorithms [2]. While these approaches achieve widespread success in flatter, homogeneous environments, they often prove more difficult in the alpine regions. Monitoring land use and land cover in the Alpine region is essential because mountain landscapes combine high biodiversity, strong ecological connectivity, intensive human use, and pronounced sensitivity to climatic and socioeconomic change [3,4]. Steep altitudinal gradients generate closely interspersed habitats, while agricultural abandonment or intensification, forest expansion, tourism development, settlement growth, and infrastructure construction can rapidly alter their spatial configuration. Consistent LULC information is therefore needed to detect habitat loss and fragmentation, evaluate ecological corridors, support the management of protected areas, monitor changes in mountain agriculture and forests, and inform spatial-planning and climate-adaptation strategies. However, producing reliable high-resolution LULC maps in mountain environments remains technically challenging. Complex terrain causes strong variations in illumination, cast shadows, and topographically induced differences in spectral response. Seasonal snow cover and vegetation phenology further reduce consistency among acquisition dates. At very high spatial resolution, increased spatial detail also produces substantial within-class heterogeneity, while ecologically distinct land-cover classes may show similar spectral characteristics [5]. Alpine landscapes are highly fragmented and contain objects over a wide range of spatial scales, making both contextual interpretation and accurate boundary delineation difficult. Multimodal analysis introduces additional challenges because optical, multispectral, and topographic datasets may differ in resolution,

spatial support, noise characteristics, and geometric alignment. Finally, reference data are often sparse, spatially clustered, imbalanced among classes, and affected by uncertainty at habitat boundaries. These characteristics motivate an approach that combines strong spatial priors with complementary spectral and terrain information rather than relying on a single image modality.

This limitation stems from extreme topographic variations, complex shadowing, frequent cloud cover, and highly dynamic snow and ice coverage, which collectively obscure spectral boundaries and undermine automated mapping methods. To mitigate these constraints, current research strongly advocates for a transition toward unmanned aerial vehicles (UAVs) and very high-resolution platforms, which allow for rapid, low-cost imaging below cloud ceilings. By capturing sub-meter, multi-sensor data including dense multispectral (MS) bands and precise Normalized Digital Surface Models (nDSMs), aerial acquisitions provide the physical and spectral depth required for complex terrain analysis. Recent studies have proven that fusing nDSM data with traditional RGB imagery significantly improves object-based land cover classification by enhancing class separability.

Parallel to advancements in aerial data acquisition, deep learning architectures have revolutionized LULC automation. Recently, vision foundation models such as the Segment Anything Model (SAM) [6] have exhibited high geometric fidelity when applied to very high-resolution aerial data. However, while SAM excels at delineating object boundaries based on visual contrast, its direct transfer to highly specialized remote sensing tasks is often hindered by profound spectral domain shifts, which suppress the high-frequency details essential for boundary delineation. While some recent methodologies have successfully adapted SAM for multispectral agricultural crop monitoring, alpine domains present unique challenges. In these high-altitude environments, a single macro-class like continuous snow, bare rock, or debris-covered ice heavily dominates the landscape, making it exceptionally difficult for standard AI architectures to maintain boundary precision for minority vegetation classes. Multimodal foundation models represent a major recent development in Earth observation, shifting the field from independently trained, task-specific networks towards reusable pretrained backbones that can be adapted to different sensors and downstream applications. SkySense is representative of this transition: it is a billion-scale model pretrained on 21.5 million multimodal temporal sequences using a factorized spatiotemporal encoder that jointly processes high-resolution optical imagery, multispectral time series, and synthetic aperture radar data [7]. Subsequent models have further expanded this paradigm. SkySense V2 and SkySense++ improve unified multimodal representation learning and semantic understanding [8,9]; RingMoE introduces modality-specialized, collaborative, and shared experts and was pretrained on 400 million images acquired by nine satellites [10]; DOFA uses wavelength-conditioned dynamic weights to accommodate heterogeneous sensors [11]; and SpectralGPT focuses specifically on transferable spatial–spectral representations [12]. Collectively, these studies demonstrate substantial progress towards models that can accommodate multiple sensors, resolutions, and Earth-observation tasks within a common pretrained framework. In parallel, multimodal fusion has evolved beyond simple feature concatenation towards asymmetric and attention-based mechanisms that preserve modality-specific information. ASANet, for example, explicitly addresses the semantic inconsistency between RGB and SAR representations [13], while LoGoCAF performs local-to-global cross-modal attention for HSI-X semantic segmentation [14]. HR-MM SegFormer is particularly relevant to high-resolution LULC mapping because it combines optical imagery with multispectral or DSM data through feature rectification and cross-attention [15]. In contrast to these approaches, our framework combines a SAM-derived spatial encoder with dedicated multispectral and topographic adapters. The proposed asymmetric dual-encoder design does not require het-

erogeneous inputs to be represented from the outset by a single shared backbone. Instead, it preserves the strong spatial and object-level priors learned by SAM while separately encoding spectral and terrain information that is not natively supported by the SAM image encoder. The subsequent feature fusion integrates these complementary representations for high-resolution Alpine LULC classification, where fine spatial boundaries, spectral similarity among classes, terrain variability, and sparse polygon-based supervision pose challenges that are not explicitly addressed by existing general-purpose multimodal foundation models.

Addressing these methodological gaps, this study introduces a highly specialized, multi-modal deep learning framework designed explicitly for the demanding complexities of high-altitude aerial mapping. While existing LULC foundation models typically rely on medium-resolution satellite data (e.g., 10 m/pixel Sentinel-2 imagery), our architecture is engineered to process synchronized aerial data acquired at centimetre-level ground sampling distance from custom Phase One RGB and MicaSense Multispectral payloads. By explicitly redefining how foundation models integrate with multi-sensor data, we propose a dual-encoder architecture that fuses SAM-derived spatial context with custom multispectral and topographic adapters. Ultimately, this work moves beyond conventional mapping limits, establishing a robust, highly novel baseline for instance-level LULC classification and precise environmental monitoring in the most challenging alpine terrains.

2. Materials and Methods

2.1. Study Area

The study area comprises alpine situated within the Aosta Valley (Valle d'Aosta), an alpine region nestled in north-western Italy at the geographic heart of the European Alps. The Aosta Valley is unique among Italian regions in that it shares its borders with France to the west and south-west, Switzerland to the north, and the Piedmont region to the south and east, rendering it a strategic high-altitude corridor between three nations [1]. The valley's total administrative extent covers approximately 3263 km^2, although the effective inhabitable area is considerably reduced by the extreme alpine topography that characterizes the region, with steep-walled valleys, towering peaks, and deeply incised glacial troughs defining the landscape at nearly every elevation band [16]. The region encompasses the Italian flanks of some of Europe's most iconic mountains, including Mont Blanc, Monte Rosa, Gran Paradiso, and the Matterhorn, making the Aosta Valley Italy's highest-elevation administrative region by mean altitude and by summit elevation. Mont Blanc, the highest peak in the Alps and western Europe, rises to 4810 m, and its summit comprises a perennial ice-and-snow dome whose thickness varies inter-annually in response to prevailing climate conditions, making it an exceptionally sensitive indicator of regional cryospheric health [17]. The regional geology is further complicated by the presence of active neotectonic structures that contribute to ongoing uplift and seismic risk, adding another layer of complexity to the interpretation of surface processes in the context of cryospheric change [18]. The relief energy of the valley is extraordinary: the floor of the main Dora Baltea valley lies at approximately 600–800 m near the city of Aosta, while encircling ridgelines and peak walls rise to over 4000 m within horizontal distances of often less than 10 km, generating extreme topographic gradients that profoundly influence every aspect of the ecological systems present [19].

The climate of the Aosta Valley is characterised by strong altitudinal zonation and significant orographic complexity. Below approximately 1600 m, valley floor stations experience a humid subtropical climate with moderate precipitation and mild temperatures, while the intermediate alpine zone between 1600 and 2500 m transitions to a cold continental climate classified as Dfc (snow climate, fully humid, with cool summers) under

the Köppen–Geiger system [20], with a snow season that extends to eight or nine months at the upper limit of this band [21]. Above 2500 m, the climate transitions to a tundra type (ET), in which mean monthly temperatures remain below 10 °C in every month of the year, perennially suppressing melt and maintaining snow and ice cover across extensive areas [22]. At Plateau Rosa (3400 m), the coldest reliably monitored station in Italy, January mean temperatures reach −11.6 °C while July averages just 1.4 °C, documenting the extreme severity of the high-altitude environment and the challenging conditions that constrain both natural ecosystem dynamics and human data acquisition activities [16]. Precipitation in the region is strongly orographic in character: valley floor stations record approximately 1000–1200 mm per year, while high-altitude accumulation zones on the main Alpine crest receive in excess of 3000 mm per year, predominantly as snow. The pronounced seasonal and altitudinal temperature gradients ensure that snow cover in the Aosta Valley is not merely a winter phenomenon but rather a landscape-scale process that progresses from high elevations downward through autumn and retreats from low elevations upward through late spring and early summer, creating a complex temporal mosaic of snow-covered and snow-free terrain at any given point in the melt season with a strong impact on the vegetation cover [23] present within the valley [24].

The complex topography, pronounced shadow effects from steep valley-side walls, frequent cloud cover at intermediate elevations, and the highly dynamic seasonal snow coverage that characterises the Aosta Valley present significant and compounding challenges for automated LULC classification based on optical remote sensing data [16]. Topographic shadowing is particularly acute in the valley's lower-elevation reaches during the winter months and in the early and late portions of the growing season, when low solar angles cast steep shadows across the valley floor and lower valley walls for extended portions of the day, confounding pixel-level spectral signatures that are fundamental to conventional image classification algorithms [17]. Cloud cover in the Alps is notoriously variable and can persist for days or weeks during the summer months, making the scheduling of optical remote sensing acquisitions both critical and challenging, as even brief cloud-free windows during the snow-melt transition may represent the only viable opportunity for capturing snow-free or minimally snow-covered imagery within a given growing season [1]. The highly dynamic nature of seasonal snow cover in the region adds a further dimension of complexity, as snow lines retreat from valley floor to ridgeline over elevations spanning more than 2500 m within a single melt season, effectively altering the spectral characteristics of tens of square kilometres of landscape within a matter of weeks [21]. Permafrost is present in the upper reaches of all four target zones and exerts a fundamental control on surface stability, vegetation colonisation limits, and the thermal regime of exposed bedrock walls, all of which have direct consequences for the observed LULC class distributions [22]. The combination of extreme topographic complexity, rapid environmental change driven by climate warming, heterogeneous land management histories, and the fine spatial scale of many LULC class boundaries in the Alps together establishes the Aosta Valley as one of the most demanding environments for high-resolution LULC classification using deep learning methods [19].

2.2. Remote Sensing Data Acquisition

To overcome the spatial limitations of standard satellite imagery, high-resolution aerial imagery was acquired using a manned aircraft platform (Tecnam P92, Figure 1) equipped with a synchronized dual-sensor payload.



Figure 1. The Tecnam P92 aircraft utilized for the high-resolution aerial photogrammetric survey over the Aosta Valley.

2.2.1. Aircraft Platform

The aircraft's manoeuvrability and performance parameters make it well-suited for navigating the complex topography of the Aosta Valley. The key specifications of the platform are detailed in Table 1.

Table 1. Technical specifications of the Tecnam P92 aircraft utilized for the aerial survey.

Parameter	Specification
Length	6.4 m
Wing Span	8.7 m
Maximum Take-Off Weight (MTOW)	600 kg
Passenger Capacity	2 max
Engine	1 × Rotax 912 S2
Horsepower	100 HP per engine
Max Speed	222 km/h
Cruising Speed	219 km/h
Stall Speed	76 km/h
Autonomy	1100 km

2.2.2. Sensor Payload

To capture comprehensive spectral data, the aircraft was equipped with two sensors: a Phase One iXM-RS150F metrically calibrated optical camera with 50 mm focal lens (manufactured by Phase One, Copenhagen, Denmark) for ultra-high-resolution nadir optical imagery, and a MicaSense Altum-PT (manufactured by MicaSense, an AgEagle Aerial System company, in Seattle, WA, USA) for multispectral data (capturing Blue, Green, Red, Red Edge, and Near-Infrared bands, together with a co-registered thermal infrared band), as shown in Figure 2.

Both sensors were mounted simultaneously on board using DigiSky's proprietary "SmartBay" system. This integration hardware allows for the simultaneous housing of multiple sensors, ensuring that both the 150-megapixel optical data and the multispectral data were acquired in strict synchronization over the target areas.

These four hydrological basin zones (Figure 3) are not merely arbitrary survey extents—they represent established monitoring locations within the Aosta Valley where the Regional

Environmental Protection Agency (ARPA VdA) has conducted annual vegetation and cryospheric assessments in collaboration with research partners.



Figure 2. The DigiSky SmartBay proprietary system housing both the Phase One iXM-RS150F and the MicaSense Altum-PT sensors for synchronized data acquisition.

These zones were selected to provide spatially distributed, elevationally stratified coverage of the principal alpine landscape types present within the valley, encompassing the full gradient from valley-floor pastures to subalpine forest and periglacial terrain above the current tree line [1]. The annual monitoring programme at these sites has historically relied on a combination of high-resolution satellite imagery—primarily Sentinel-2—for broad-area contextual assessment, and Unmanned Aerial System (UAS/drone) deployments for ultra-detailed, plot-level botanical and structural characterisation [16]. This dual-platform approach has enabled ARPA VdA to leverage the synoptic coverage of satellite data for landscape-scale change detection while retaining the centimetric spatial detail of drone acquisitions for species-level observation. However, both legacy approaches carry fundamental and well-documented trade-offs: satellite sensors, despite ever-improving spatial resolutions, remain constrained by pixel sizes that average 10 m for openly accessible optical products, rendering them inadequate for resolving fine-scale LULC boundaries and ecosystems signals within structurally complex vegetation communities [21]. Conversely, UAS platforms, while capable of delivering centimetre Ground Sampling Distance (GSD) imagery over limited extents, are severely restricted in their operational coverage by flight endurance, regulatory airspace constraints, and the logistical complexity of deploying them across the large, remote, and topographically hazardous terrain that characterises the alpine target zones [24]. Consequently, neither platform, operating independently, is capable of delivering the simultaneous combination of broad spatial coverage, fine spatial resolution, and spectrally rich multi-modal input that is required to map the various ecosystem components and monitor their evolution over time [17].

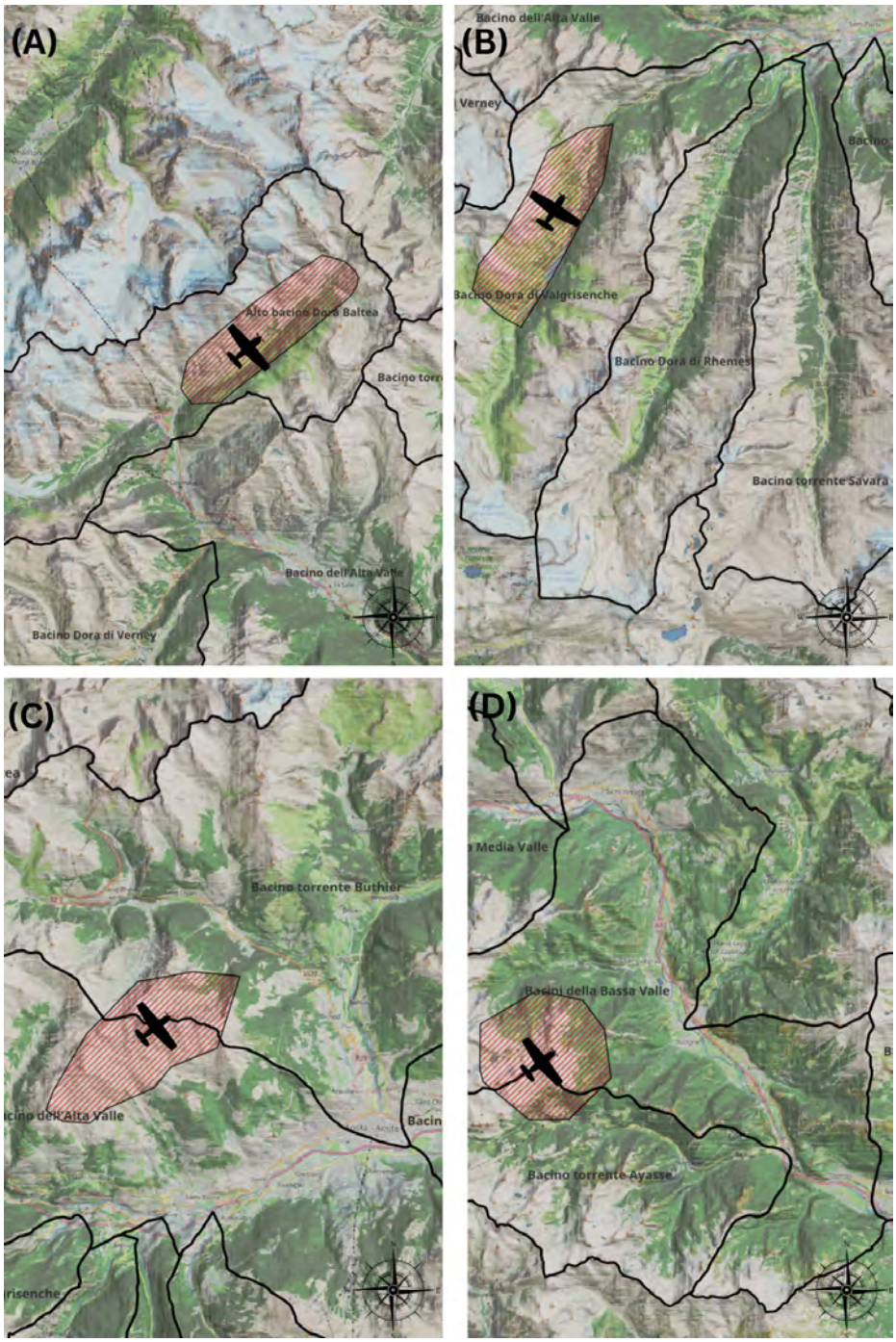


Figure 3. Location and extent of the four surveyed basin zones (**A**) Val Ferret (Courmayeur), (**B**) Valgrisenche, (**C**) Chaligne/Fallere (Aosta, Sarre), and (**D**) (Champorcher, Champdepraz) surveyed during the TEST BEFORE INVEST aerial campaign.

The aerial survey documented in this study was conceived and executed within the framework of the CHEDIH project (Changing European Highland Environment: Dynamics, Interactions, and Hazards), a large-scale European Commission-funded research initiative directed at improving environmental monitoring and hazard assessment across European alpine and highland environments [25]. The aerial campaign was specifically conducted under the TEST BEFORE INVEST sub-project, which was designed to evaluate the operational readiness of novel remote sensing platforms and methodologies including piloted aerial acquisition as demonstrated here prior to scaled deployment within ARPA VdA's routine monitoring programme [24]. The CHEDIH project brings together a consortium

of research institutions, environmental agencies, and technical partners to address critical knowledge gaps in high-altitude ecosystem dynamics and climate change impacts. Within this framework, the primary project beneficiary is ARPA VdA, whose mandate includes the long-term environmental monitoring of the Aosta Valley's ecological resources, while DigiSky S.r.l. serves as the technical erogatore—the implementation partner responsible for mission execution, sensor integration, and aerial data acquisition. Within this framework, the upgrade from UAS to piloted aerial acquisition represents a first-of-its-kind operational deployment for ARPA VdA in the context of alpine vegetation monitoring: before TEST BEFORE INVEST, no aerial campaign had been conducted by or on behalf of the agency specifically for the environmental surveillance of high-altitude forest ecosystems in the Aosta Valley [24]. The motivation for this transition is directly tied to the coverage-to-resolution constraint described above. A single aerial survey flight line, flown at altitude from the Tecnam P92 platform, can encompass tens of linear kilometres of terrain in a single sortie, delivering contiguous, georeferenced optical and multispectral coverage at 15 cm GSD across an entire basin zone—a feat that would require dozens of individual UAS flights to replicate, and that satellite platforms cannot achieve at any temporal frequency compatible with growing-season monitoring [16]. The resolution achieved by the aerial acquisition is furthermore sufficient to resolve individual shrub crowns, erosion trails, and fine linear features within the vegetation mosaic Studying alpine ecosystems Earth observation techniques presents specific and compounding difficulties that extend well beyond the generic challenges of alpine remote sensing: the extreme heterogeneity of surface reflectance characteristics arising from admixtures of bare rock, moss, lichen, sparse shrubs, grassland, debris and snow; the pervasive topographic shadowing that varies with aspect and elevation These difficulties have historically limited the availability of high-quality, consistent, multi-temporal benchmark datasets for alpine vegetation monitoring, and they underscore the value of the aerial multi-modal dataset generated within TEST BEFORE INVEST as a rare and methodologically novel contribution to the field [19].

2.2.3. Mission Planning and Execution

Depending on the onboard equipment, aerial surveys can easily achieve decimeter to centimeter spatial resolutions in the optical domain. By definition, the GSD is the ratio between the product of the flight altitude and the sensor pixel size, divided by the focal length of the coupled lens. Consequently, the effective GSD varies across sites, function of terrain elevation changes, safety-mandated obstacle clearance, and airspace access restrictions. The flight plan was developed and managed using the Skymetry Mission Planning tool, which enables DigiSky operators to accurately coordinate flight activities and design flight lines over specific target areas. To verify adequate separation conditions and ensure strict adherence to aviation safety regulations within the mountainous terrain, the planner incorporated a Digital Elevation Model (DEM) derived from the Shuttle Radar Topography Mission (SRTM) at a resolution of 3 arc-seconds (approximately 30 m).

Acquisition timing was carefully scheduled to minimize sun-angle shadows and target snow-free conditions where possible and to maximize the differentiation of species (or groups of species) based on phenological characteristics. The mission was executed according to the following guaranteed acquisition parameters:

- Phase One iXM-RS150F (Nadir Optical): Utilizing a 50 mm focal length and an effective resolution of 150 MPix, the mission achieved a guaranteed minimum GSD of 15 cm/pix. Flight lines were planned with a longitudinal overlap exceeding 85% and a lateral overlap (sidelap) exceeding 45%. This ultra-high resolution is sufficient to resolve fine-scale LULC features of interest such as individual shrubs, erosion trails, and trees.

- MicaSense Altum-PT (Multispectral): With an effective resolution of 3.2 MPix, the multispectral acquisition achieved a guaranteed GSD of under 86 cm/pix on target (strictly maintained below 1 m/pix). The multispectral flight parameters ensured a longitudinal overlap greater than 85% and a lateral overlap greater than 40% [1,16].

Aerial imagery was acquired using a manned aircraft platform equipped with a 6-band multispectral sensor capturing Red (R), Green (G), Blue (B), Red Edge (RE) and Near-Infrared (NIR), and Thermal Infrared (T) bands spectral bands and another high-resolution RGB Phase One. Mission planning ensured a minimum of 80% forward and lateral overlap between adjacent flight lines, achieving a GSD of 15 cm for the RGB imagery and below 86 cm for the multispectral imagery. At acquisition these resolutions are sufficient to resolve fine-scale LULC features such as individual shrubs, erosion trails, and road edges, although the common analysis grid described in Section 2.2.4 is necessarily coarser. Acquisition timing was scheduled to minimize sun-angle shadows and snow-free conditions where possible.

2.2.4. RGB Orthomosaic Generation

Raw aerial image blocks were processed using Structure-from-Motion (SfM) photogrammetry via a commercial photogrammetric software suite to generate ortho-rectified image products [2,24]. Dense image matching was performed to produce a digital surface model (DSM) capturing above-ground surface heights while ground control points were used to derive a DEM representing the bare-ground terrain. All products were georeferenced to ETRS89/UTM zone 32N (EPSG:32632) and re-sampled onto a common analysis grid of 0.64 m for consistency. This resolution was chosen to lie close to the native ground sampling distance of the multispectral sensor rather than that of the RGB sensor: upsampling the multispectral bands towards 15 cm would have fabricated spectral detail the instrument did not record, whereas downsampling the RGB imagery discards spatial detail without introducing information that was never measured. Every tile therefore covers an identical ground extent of 655 m $\times$ 655 m across all four modalities and the annotation mask, with pixel-level correspondence guaranteed by construction.

2.3. Digital Elevation and Surface Models

A DEM and a DSM were derived from the same aerial image block via Agisoft Metashape Version 2.3.0. The DEM represents the bare-ground terrain elevation excluding vegetation and structures, while the DSM captures top-of-canopy and built-structure heights. These products served as ancillary input layers to provide topographic context—a critical consideration in alpine environments where elevation, slope, and aspect are strongly correlated with LULC class distributions [17,19].

2.4. Feature Engineering

2.4.1. Spectral Indices

In addition to the six raw spectral bands, a suite of 13 spectral index proxies was computed at native multispectral image resolution to enhance class discriminability, particularly for spectrally similar landcover classes (Table 2).

2.4.2. Rationale for Spectral Index Proxies

The red-edge band is particularly valuable in Alpine environments because it captures spectral variations related to vegetation chlorophyll content and canopy structure that may not be adequately represented by conventional visible and near-infrared bands [21,22]. Accordingly, $NDVI_{\mathrm{RE}}$ and CI_{RE} exploit red-edge information to improve the discrimination of sparse shrubland, krummholz, and vegetation occurring near snow-covered surfaces

from bare rock and seasonal snow. The set of spectral indices was selected to address the specific discrimination problems encountered in the study area rather than to provide a general-purpose vegetation-monitoring product. Although NDVI is one of the most widely used vegetation indices, it can saturate under moderate-to-high canopy density or biomass, reducing its sensitivity to differences among the dense woody vegetation classes occurring in the surveyed areas. Furthermore, indices based on green or red-edge reflectance may be more sensitive than conventional red-based NDVI to variations in chlorophyll content within dense vegetation, which can assist in distinguishing meadow, heath, shrub, and woody vegetation types. These limitations motivated the inclusion of $NDVI_{RE}$ and GNDVI, which replace the red band with the red-edge and green bands, respectively, and retain sensitivity under vegetation conditions in which conventional NDVI may provide limited discrimination. Because the original Red and NIR bands were already included in the multispectral input, NDVI would have provided a largely redundant transformation of information already available to the model. Its exclusion therefore limited predictor redundancy and input dimensionality while prioritising indices designed for the classification challenges encountered in the Alpine study area. The WS_AWEI index was adopted to better discriminate open water bodies (Lakes, Wetlands) from bright snow surfaces—a known challenge in cryospheric LULC mapping [17,23].

Table 2. Spectral indices used as input features for supervised classification.

Index	Abbreviation	Primary Application	Formula/Reference
Built-up Land Features Extraction Index—Proxy	BLFEI_Proxy	Built-up land	See [26]
Bare Soil Index—Proxy	BSI_Proxy	Bare soil/rock	See [27]
Chlorophyll Index—Red Edge	CL_RE	Vegetation chlorophyll content	(NIR/RE) − 1 [28]
Enhanced Built-Up & Bareness Index—Proxy	EBBI_Proxy	Built-up and bare surfaces	See [29]
Green NDVI	GNDVI	Vegetation vigour	(NIR − Green)/(NIR + Green) [30]
Index-based Built-up Index—Proxy	IBI_Proxy	Built-up areas	See [31]
Normalized Difference Built-up Index—Proxy	NDBI_RE	Built-up (Red Edge variant)	(RE − NIR)/(RE + NIR) [32]
NDVI Red Edge	NDVI_RE	Vegetation (Red Edge)	(RE − Red)/(RE + Red) [33]
Perpendicular Impervious Surface Index	PISI	Impervious surfaces	0.8192 Blue − 0.5735 NIR + 0.0750 [34]
Transformed Chlorophyll Absorption	TCARI	Vegetation chlorophyll	See [35]
Urban Index—Proxy	UI_Proxy	Urban/built-up	(RE − Green)/(RE + Green) [36]
Water Ratio Index—Proxy	WRI_Proxy	Open water bodies	(NIR + RE) [37]
Automated Water Extraction Index—Proxy	WS_AWEI	Water/snow differentiation	See [23]

Indices labelled Proxy are adaptations of the cited formulation in which the shortwave infrared bands required by the original index are replaced by the Red Edge, Red or Blue bands, as the MicaSense Altum-PT payload does not acquire SWIR. These adaptations are therefore not equivalent to the published indices and are used here purely as learned input features. RE denotes the Red Edge band.

2.4.3. Integration of Elevation and Topographic Data

The DEM and DSM were resampled onto the common 0.64 m analysis grid and terrain attributes including slope and aspect were derived from the DEM using standard gradient operators and included as model inputs. Topographic features are especially relevant for specific classes (e.g., Snow/seasonal snow accumulation zones) and rock–vegetation boundaries in steep alpine terrain [18,19].

2.5. Land Use Land Cover Classification Schema

The classification schema adopted here follows the CORINE Land Cover (CLC) nomenclature. For this specific type of classification, no pre-exsisting alpine land cover taxonomy was provided by ARPA Valle d'Aosta. CLC was selected because it is the established reference for land cover mapping in Europe, especially. Initiated in 1985 and now delivered through the Copernicus Land Monitoring Service, it provides a hierarchical nomenclature of five Level 1 categories, 15 Level 2 categories, and 44 Level 3 categories [38,39]. Adopting it means the resulting maps aggregate directly into national and European land cover reporting, thus has been pursued within this current work. What CLC lacks is that it alone cannot describe the surface at the resolution acquired here. Its nomenclature is designed for

regional and continental mapping from satellite imagery and applies to minimum mapping unit of 25 hectares, with a minimum width of 100m for linear features, so features such as road corridors, individual tree crowns, isolated shrub patches, narrow water bodies etc. that are central to alpine monitoring at 15 cm ground sampling distance of this survey with respect to a 100m width corresponds roughly to 660 pixels, as such therefore the current work retains the CLC hierarchy and extends it with a project specific custom level 4, in which each class nests within an existing CLC parent and is resolvable at centimeter scale (Table 3). Class definitions themselves were established in three stages. An unsupervised k-means clustering of the multispectral imagery was first used to surface the spectral groupings actually present in the surveyed basins; domain experts at ARPA VdA then reviewed these groupings to determine which corresponded to operationally meaningful land cover types and mapped them onto the CLC hierarchy; final polygon labels were assigned by manual photo-interpretation, supported by DEM/DSM topographic context and the high-resolution reference RGB orthomosaics.

Fourteen LULC classes spanning natural, semi-natural, and anthropic land cover types were defined (Table 3). Class labels were assigned through manual photo-interpretation supported by field knowledge provided by experts at ARPA VdA, and visual analysis on DEM/DSM topographic context, and high-resolution reference RGB orthomosaics.

Table 3. Land cover classification schema, ordered by class identifier.

Class ID	Class Name	CLC Level 1
0	Wetlands	Wetlands
1	Heath	Forest and semi-natural areas
2, 3	**MERGED_MEADOW**	Agricultural areas
2	∟ *Non_Irrigated_Meadow*	
3	∟ *Irrigated_Meadow*	
4	Shrub_Evergreen	Forest and semi-natural areas
6	River	Water bodies
7	Lakes	Water bodies
8	Buildings	Artificial surfaces
9	Road	Artificial surfaces
10	Rock	Forest and semi-natural areas
11	Rocky_Foundation	Forest and semi-natural areas
12	**SNOW (Merged Snow Class)**	Forest and semi-natural areas
12	∟ *Snow*	
12	∟ *Dirty_Snow*	
15, 16 [a]	Trees	Forest and semi-natural areas

[a] Broad-leaved (15) and coniferous (16) forest, merged as Trees.

3. Dataset Preparation and Ingestion Pipeline

The efficacy of multi-modal segmentation architectures heavily relies on the strict spatial alignment of input sensors and robust handling of class imbalance. To facilitate this, we developed a two-stage data pipeline comprising an offline dataset generation phase for spatial alignment, followed by a dynamic runtime loading phase incorporating priority sampling for rare alpine typologies.

3.1. Dataset Description and Weak Supervision

The dataset was collected across the designated alpine study areas (for full geographical definitions and zonal delineations, please refer to Section 2.1). Rather than utilizing dense, pixel-perfect masks, annotations were provided as polygonal boundaries delineated directly on the RGB orthomosaic products. This weakly supervised paradigm significantly reduces labeling effort compared to dense pixel annotation while preserving critical spatial boundary precision.

A total of 2517 polygons were manually delineated across the land-cover classes. During preliminary data exploration, it was observed that specific meadow sub-typologies

namely, *Dry_Meadow* () and *Irrigated_Meadow* exhibited extreme spectral similarity and lacked distinct spatial boundaries in both the RGB and MS domains. To prevent radiometric confusion and improve model convergence, these sub-categories were discarded prior to final modal configuration retaining *Irrigated_Meadow* and *Non_Irrigated_Meadow* as distinct classes grouped under a common meadow macro-class.

These polygons were subsequently converted to dense binary masks via rasterization during the tiling process. Crucially, all unlabeled regions falling outside these polygons were assigned an *ignore_index* (value of 255) in the loss calculation. This ensures the model only backpropagates gradients from confidently annotated pixels and is not penalized for predicting classes in ambiguous, unlabeled regions.

3.2. Offline Dataset Generation and Modality Alignment

Raw geospatial orthomosaics and corresponding vector annotations were systematically processed into a spatially aligned, tile-based dataset. The Multispectral (MS) orthomosaic was designated as the master spatial reference grid. Because MS bands typically possess a coarser native resolution than RGB imagery, interpolating RGB up to the MS grid prevents the artificial generation of spectral information that would occur if MS data were spatially aggregated.

A sliding window approach was applied across the MS master grid to extract non-overlapping tiles of 1024×1024 pixels. Vector polygon shapefiles were reprojected to the MS Coordinate Reference System (CRS) and clipped strictly to the tile boundaries. All auxiliary modalities were subsequently aligned to this master bounding box. RGB imagery was resampled using bicubic interpolation to match the MS tile geometry, preserving the native 8-bit unsigned integer format. Spectral indices and nDSM matrices were read directly via coordinate windows.

The tiles were randomly partitioned into training and testing sets using an approximate 80/20 split applied on a per-class basis. As detailed in Table 4, the final curated dataset comprises 646 annotated tiles (501 training, 145 testing). The shadowed snow class (*Dirty_Snow*) is grouped with the primary snow class (*Snow*) under a common macro-class, while remaining separately evaluated.

Table 4. Class distribution and annotation statistics across the alpine dataset, reported prior to the macro-class merging described in Section 3. Indented rows represent original raw annotations. Classes subsequently consolidated or excluded do not appear in Table 5.

Class Name	Polygons	Train Tiles	Test Tiles
Wetlands	90	30	13
Heath	198	23	6
MERGED_MEADOW	**199**	**91**	**26**
⌞ *Non_Irrigated_Meadow*	*114*	*60*	*16*
⌞ *Irrigated_Meadow*	*85*	*31*	*10*
Shrub_Evergreen	221	19	5
River	107	25	5
Lakes	75	41	19
Buildings	174	51	16
Road	202	59	20
Rock	193	44	9
Rocky_Foundation	90	17	4
SNOW (Merged Snow Class) *	**118**	**23**	**5**
⌞ *Snow*	*85*	*13*	*4*
⌞ *Dirty_Snow*	*33*	*10*	*1*
Trees	850	78	17
Total	**2517**	**501**	**145**

* Sub-classes merged due to high spectrmodel convergence.

Table 5. Per-class Intersection over Union at Epoch 30.

Class Name	Train mIoU	Val mIoU
Wetlands	1.00	0.00
Heath	0.85	0.80
MERGED_MEADOW		
⌞ *Non_Irrigated_Meadow*	*0.88*	*0.82*
⌞ *Irrigated_Meadow*	*0.92*	*0.88*
Shrub_Evergreen	0.95	0.40
River	0.90	0.85
Lakes	0.96	0.96
Buildings	0.97	0.93
Road	0.95	0.91
Rock	0.93	0.88
Rocky_Foundation	0.89	0.80
SNOW (Merged Snow Class)		
⌞ *Snow*	*0.94*	*0.90*
⌞ *Dirty_Snow*	*0.92*	*0.85*
Trees	0.94	0.90
Macro average (14 classes)	**0.93**	**0.777**

3.3. Dataset Visual Exemplars

To illustrate the visual heterogeneity and spatial complexity of the alpine environment, Figure 4 presents representative RGB patches for the target land-cover classes. As depicted, the ground-truth annotations are provided as delineated polygonal boundaries (weak supervision) rather than dense pixel-wise masks. This highlights the intra-class variance and the challenge of distinguishing spectrally similar typologies (e.g., separating various meadow types from heath or evergreen shrubs) in high-resolution aerial imagery.

3.4. Runtime Data Loading and Priority Sampling

During the training phase, a custom data ingestion pipeline applies stochastic augmentations, modality-specific normalization, and targeted sampling.

Modality-Specific Normalization: Because the model fuses data from heterogeneous sensors, varied normalization strategies were strictly applied to preserve the mathematical integrity of each modality:

- RGB Imagery: Standardized using ImageNet channel means and standard deviations to align with the pretrained weights of the SAM backbone.
- Multi-Spectral and Spectral Indices: Scaled to a continuous $[0, 1]$ range using 1st and 99th percentile min–max clipping per band. This preserves relative spectral reflectance differences while remaining robust to radiometric outliers.
- nDSM Topology: Scaled to $[0, 1]$ using a top-hat transformation to represent relative topographical height rather than absolute elevation above sea level.

To ensure the network does not learn ambiguous edge artifacts, any pixel flagged as *NoData* in any single modality is uniformly masked across all tensors and assigned the *ignore_index* in the ground truth annotation.

Priority Sampling for Class Imbalance:
The polygon distribution reveals severe class imbalance typical of natural environments. *Trees* alone accounts for approximately 34% of all annotations, while rare classes such as snow and wetlands are heavily underrepresented. A further difficulty follows from the weakly supervised design: only the interior of the delineated polygons carries a label, and the remainder of each tile is assigned the *ignore_index* and excluded from the loss, so the supervised signal available in any given tile is proportional to the annotation it contains. To direct training towards the classes

least represented in the annotation set, a polygon-centroid priority sampling procedure was implemented, showcased in priority sampling.

During an initial scan the dataset is traversed once and the centroid of every annotated polygon is recorded together with its class label and source tile, creating a register of candidate sampling locations. The same scan accumulates per-class instance counts, which characterise the imbalance reported in Table 4; classes holding fewer than 150 instances are flagged as minority typologies, and the register is restricted to their polygons. During stochastic gradient descent, half of all training samples are drawn from this register rather than from the sequential tile order, so that tiles containing minority-class annotations are revisited substantially more often than their share of the dataset would otherwise allow, while the remaining half preserves the natural distribution.

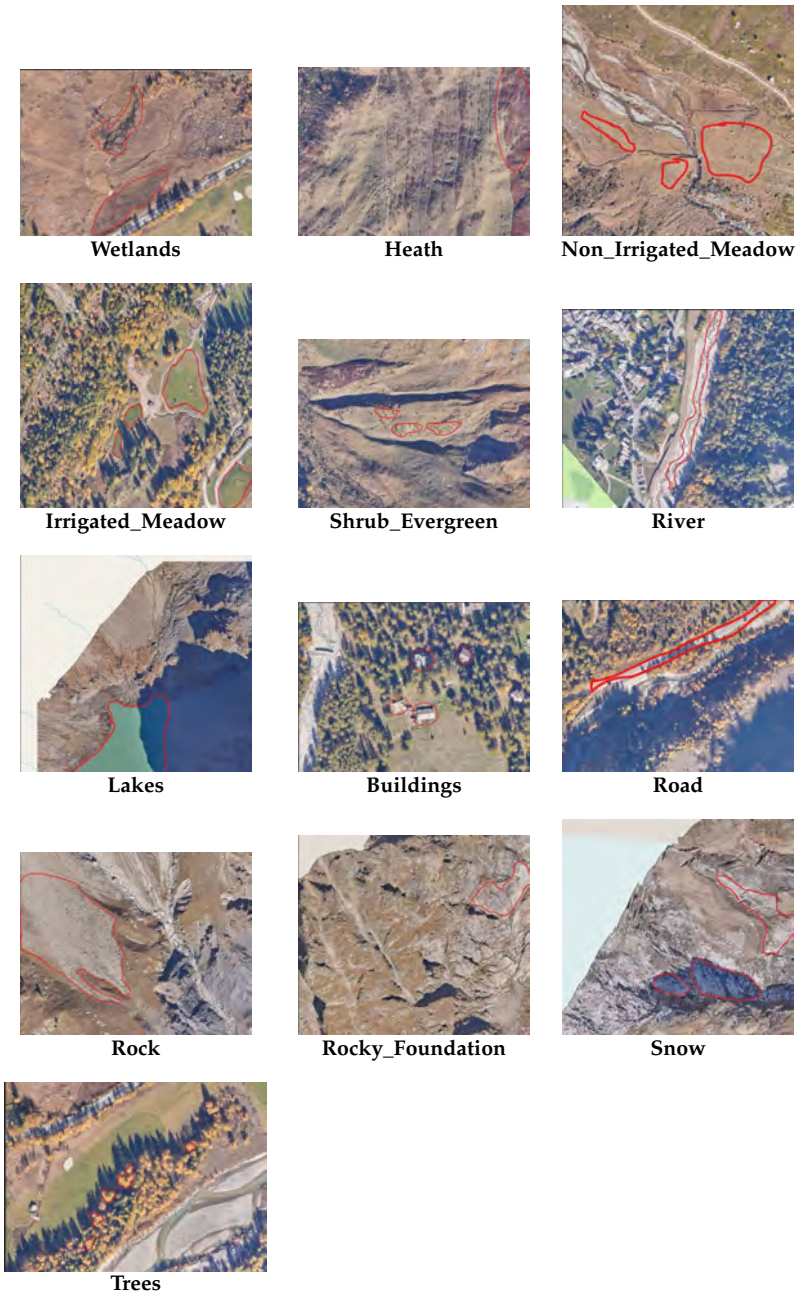


Figure 4. Representative 1024 × 1024 pixel RGB image patches showcasing 13 of 14 classes in the Aosta Valley study area; Polygons overlaid on the imagery denote the general boundary representations for interpretation of the respective LULC classes.

Data Augmentation:
To maximize generalization, spatial augmentations are synchronously applied across all modalities and annotation masks to maintain geometric alignment. Data augmentation was performed stochastically at loading time rather than by generating and storing a fixed set of transformed images. Consequently, the nominal training set remained unchanged at 510 tiles. Each tile was revisited once per epoch and independently transformed on each visit, resulting in 15,300 augmentation draws over the 30 training epochs. This value represents the number of presentations to the augmentation pipeline and not a guaranteed number of unique transformed samples. Horizontal and vertical flips and 90° rotations were each applied with a probability of 0.5. A shift–scale–rotate transformation, with maximum translation, scaling, and rotation of ±5%, ±10%, and ±15°, respectively, was applied with a probability of 0.5. All geometric transformations were applied jointly to the co-registered RGB, multispectral, spectral-index, and nDSM inputs and to the corresponding annotation mask. Photometric jitter was applied only to the RGB channels, with a probability of 0.7, to avoid modifying the physical values of the other input modalities. Because the transformations were generated dynamically, augmentation did not create a fixed number of additional samples per class and did not alter the nominal class distribution. Instead, it increased the variability of the spatial configurations presented to the model.

Additionally, photometric distortions (brightness, contrast, and HSV shifts) are independently applied exclusively to the RGB tensors to simulate varied lighting conditions without corrupting the physical spectral reflectance values of the MS bands.

3.5. Implementation Details and Hardware Specifications

All computational experiments, encompassing the offline dataset generation, modality alignment, and deep learning model training, were executed on a Dell G15 5530 mobile workstation. The system hardware comprised a 13th Generation Intel Core i5-13450HX processor (10 cores, 16 threads) paired with 16 GB of system RAM, operating on Microsoft Windows 11 Home.

Neural network optimization, gradient calculations, and inference were strictly accelerated utilizing an NVIDIA GeForce RTX 3050 Laptop GPU equipped with 6 GB of dedicated VRAM. Training a 1.1-billion parameter network (SAM ViT-H) alongside complex multi-modal fusion tensors typically requires 8–10 GB of VRAM for a single training batch in standard 32-bit floating-point (FP32) precision. To successfully engineer and train this massive architecture within the strict 6 GB VRAM constraint without triggering Out-Of-Memory (OOM) exceptions, several aggressive memory-optimization strategies were implemented:

- **Backbone Freezing and Selective Fine-Tuning:** The vast majority of the SAM backbone was completely frozen. By not storing gradient graphs for these parameters, approximately 4.4 GB of VRAM was conserved. For domain adaptation to the alpine environment, only the final 2 of the 32 transformer blocks were unfrozen for fine-tuning.
- **Automatic Mixed Precision (AMP):** Forward and backward passes were executed using half-precision (FP16) floating-point format via PyTorch (Version 2.9) AMP. This drastically reduced the tensor memory footprint by roughly 50% while maintaining numerical stability during loss scaling.
- **Gradient Accumulation:** Due to spatial memory limits, the physical mini-batch size was restricted to 1 sample per forward pass. To ensure stable gradient descent, gradient accumulation was set to 2, effectively simulating a batch size of 2 while saving approximately 2 GB of VRAM compared to a true batch of 2.
- **Differential Learning Rates:** To preserve the pretrained spatial priors while adapting the newly initialized fusion components, a differential learning rate strategy was

employed. The partially unfrozen transformer blocks received a significantly lower learning rate compared to the cross-attention modules to maintain optimization stability.

- **Modality Dropout Regularization:** To foster training diversity and force the network to rely on generalized cross-sensor representations without the memory overhead of loading additional augmented data, dynamic modality dropout was applied at rates of 15% for both the RGB and nDSM streams.

3.6. Deep Learning Segmentation Architecture

The main steps of this system architecture are represented in Figure 5, and described in detail below.

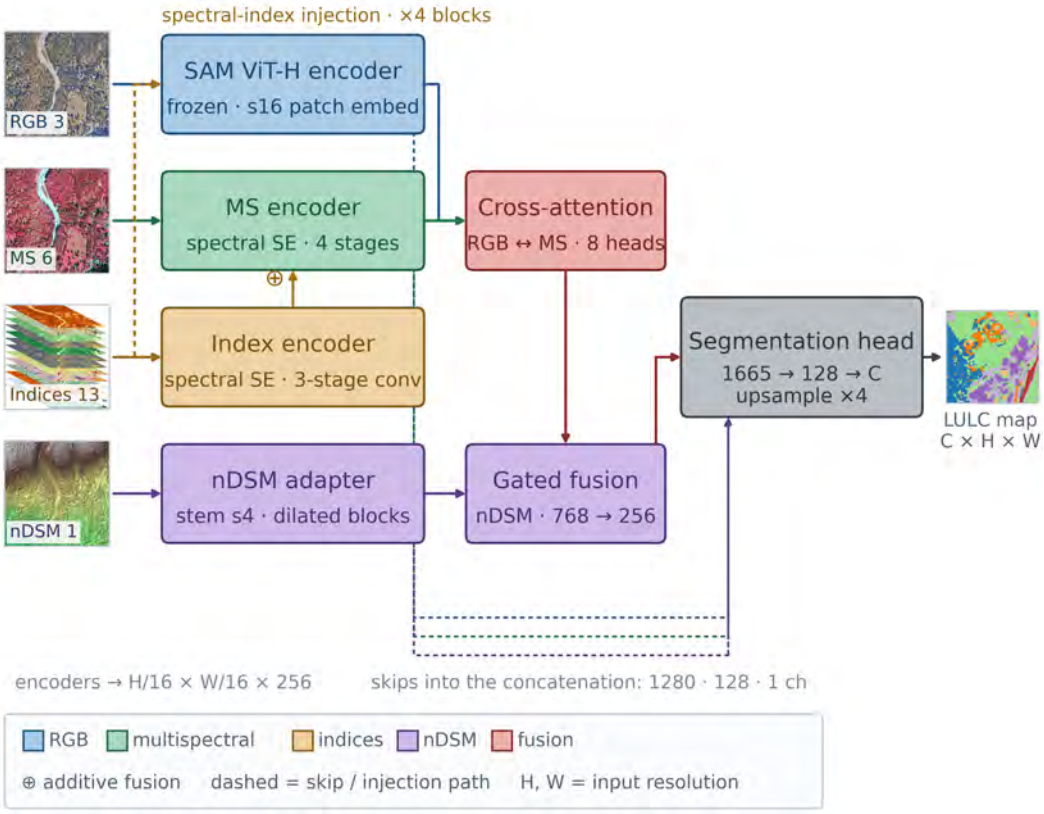


Figure 5. Detailed schematic of the proposed dual-stream multi-modal segmentation architecture, illustrating the parallel extraction streams, cross-modal attention mechanisms, and multi-scale feature aggregation.

3.6.1. Primary Feature Encoders

RGB Foundation Backbone (SAM ViT-H): To leverage highly robust, domain-agnostic spatial representations, the RGB stream utilizes the Vision Transformer Huge (ViT-H) backbone [40] from the Segment Anything Model (SAM) [6]. The frozen encoder accepts RGB patches of 1024 × 1024 pixels, processing them through a 16 × 16 patch embedding layer to produce a 64 × 64 token grid. The sequence is subsequently processed by 32 transformer blocks equipped with relative position biases [41]. To adapt the generic SAM features for land-cover classification, the backbone is selectively modulated via features injected from the Spectral Indices adapter at four equidistant transformer blocks. A convolutional neck ultimately reduces the massive 1280-dimensional channel depth to a manageable 256-dimensional feature map at 1/16 resolution. Furthermore, raw texture features from the initial block (Block-0) are bypassed directly to the segmentation head via a skip connection to preserve uncompressed spatial fidelity.

Multi-Spectral (MS) Encoder: While the RGB stream relies on a massive pretrained transformer, the 6-band MS data (Blue, Green, Red, Red Edge, NIR and Thermal Infrared) is processed by a custom, lightweight ResNet-style Convolutional Neural Network (CNN) [42] trained from scratch. This stream is designed to strictly evaluate localized spectral distributions. The input initially passes through a Squeeze-and-Excitation (SE) block [43] across the channel dimension to dynamically recalibrate the importance of specific spectral bands. The spatial encoding occurs across four stages: Stage 1 utilizes a 7×7 depthwise separable convolution [44] to reduce computational overhead, followed by three stages of residual blocks enhanced with SE attention. This encoder yields two vital outputs: an intermediate high-resolution feature map at 1/4 scale (128 channels) used for deep supervision and skip connections, and a deep semantic feature map at 1/16 scale (256 channels). Prior to cross-modal fusion, this deep MS feature map is additively fused with features derived from the Spectral Indices.

3.6.2. Auxiliary Modality Adapters

Spectral Indices Adapter: To prevent the model from having to re-learn well-established empirical geophysical thresholds, we introduce a dedicated Spectral Indices stream. This stream explicitly processes a multi-channel tensor containing pre-computed indices specifically selected to isolate challenging land-cover variations in the study area. These indices are grouped into three primary functional categories:

- **Urban and Impervious Surfaces:** To distinguish complex anthropogenic features, the input includes the Built-up Land Features Extraction Index Proxy (BLFEI_Proxy), Enhanced Built-Up & Bareness Index Proxy (EBBI_Proxy) [29], Index-based Built-up Index Proxy (IBI_Proxy) [31], Urban Index Proxy (UI_Proxy) [36], Perpendicular Impervious Surface Index (PISI) [34], and a Red-Edge variant of the Normalized Difference Built-up Index (NDBI_RE) [32].
- **Bare Soil and Rock:** The Bare Soil Index Proxy (BSI_Proxy) [27] is utilized to differentiate exposed mineral surfaces from surrounding non-photosynthetic structures.
- **Vegetation Vitality and Chlorophyll:** To capture fine-grained variations in plant health and canopy chemistry, the stream ingests the Green NDVI (GNDVI) [30], NDVI Red-Edge (NDVI_RE) [33], Chlorophyll Index Red-Edge (CL_RE) [28], and the Transformed Chlorophyll Absorption in Reflectance Index (TCARI) [35].

This input tensor first passes through a spectral Squeeze-and-Excitation (SE) layer [43] to dynamically prioritize active indices given specific regional or seasonal conditions. Features are then spatially downsampled through a 3-stage convolutional network consisting of a 7×7 convolution (stride 4), followed by two successive 3×3 convolutions (stride 2 each), mapping the dimensions directly to a 1/16 resolution grid (64×64 pixels with 256 channels).

The latent feature map serves a dual-routing purpose. First, for Multi-Spectral fusion, it passes through a 1×1 convolution layer and is additively combined with the deep Multi-Spectral representation to enrich local spectral representations. Second, it is injected into the SAM stream, where it passes through a separate 1×1 projection layer that expands the channel depth from 256 to 1280. This tensor is duplicated across four dedicated, block-specific batch normalization layers [45]. These projected features are directly injected into four equally-spaced transformer layers within the frozen SAM backbone, dynamically altering the internal token states of the ViT based on local macro-spectral context.

Topographic and Height Adapter (nDSM): To resolve visual ambiguities between classes with identical spectral reflections but drastically different vertical profiles (such as low-lying wetlands versus dense forest canopies), a photogrammetrically derived nDSM is ingested

as a single-channel height topology map. The nDSM values are min–max normalized to a continuous range of $[0, 1]$ before extraction.

The structural network for the height stream relies heavily on dilated convolutions [46] to capture structural geometries without aggressively losing fine spatial maps. The nDSM stem applies a 7×7 convolution with a stride of 4, followed by a BatchNorm-ReLU sequence and an SE-block channel recalibration layer, reducing the spatial grid to 1/4 resolution ($256 \times 256 \times 64$).

The features are then processed sequentially through two Dilated Residual Blocks (DilatedResBlock). Each block implements a 3×3 convolution using a dilation rate of 2 and a stride of 2. This parameterization expands the effective receptive field of the network to 5×5 pixels at each localized step, matching the physical metric scale of trees, buildings, and urban canopies in the aerial imagery. The final stage maps the height representation to the uniform 1/16 spatial bottleneck ($64 \times 64 \times 512$), which is projected via a 1×1 convolution down to 256 channels to match the target fusion space.

3.6.3. Cross-Modal Fusion Mechanisms

A fundamental challenge in multi-modal remote sensing architectures is preventing highly parameterized streams (e.g., the 1.1 billion parameter SAM ViT-H backbone) from semantically overwhelming the smaller, specialized streams during feature aggregation. To address this dimensional and representational imbalance, the architecture eschews naive concatenation at the primary 1/16 bottleneck in favor of a specialized cross-modal attention and gating sequence.

Bidirectional Cross-Attention Fusion (RGB and MS): The primary semantic fusion between the spatial (RGB) and spectral (MS + Indices) domains is achieved via a Bidirectional Cross-Attention mechanism [47]. Rather than independently processing features, this module forces the modalities to explicitly query one another, promoting a symmetric exchange of context.

The module comprises two parallel attention pathways. In the first pathway (RGB $\rightarrow$ MS), the spatial RGB features act as queries (Q), while the fused Multi-Spectral features act as the keys (K) and values (V). This enables the model to selectively embed localized spectral signatures into the spatial geometries detected by the ViT-H backbone. Conversely, the second pathway (MS $\rightarrow$ RGB) uses the MS features as queries and RGB features as keys/values, anchoring the spectral data with high-resolution spatial boundaries.

Each pathway consists of two cascaded Cross-Attention blocks implementing scaled dot-product attention [47]. To ensure representational diversity, multi-head attention is utilized with 8 independent heads, each computing a 32-dimensional latent subspace. The outputs of both pathways are concatenated (yielding a 512-dimensional tensor) and passed through a projection block consisting of linear transformations, Layer Normalization [48], and GELU activations [49] to compress the fused representation back to 256 channels.

Gated Height Fusion (nDSM): While height topology is critical for distinguishing certain land-cover classes (e.g., differentiating elevated forest canopies from spectrally similar ground-level wetlands), it acts as noise for others where surface elevation varies arbitrarily. To allow the network to dynamically decide when to incorporate topological data, the nDSM features are integrated using a learnable spatial Gated Fusion module.

The module computes a dense spatial gate map by passing the combined RGB-MS features through a convolutional layer with a sigmoid activation function. To prevent the gating mechanism from silencing the DSM stream entirely during early training epochs, the gate bias is initialized at 2.0, ensuring an initial sigmoid output near 0.8 (predominantly

open). The deep nDSM features are then filtered via a Hadamard product (element-wise multiplication) with this gate map.

Following the gating operation, a weighted concatenation is performed. The raw RGB features, the cross-attention fused RGB-MS features, and the gated nDSM features are channel-stacked and scaled by learnable scalar weights, forming a 768-dimensional tensor. A sequence of 1×1 convolutions acts as a bottleneck, reducing the tensor to the final 256-channel fused representation. Finally, an alpha-scaled residual connection from the original RGB features is added to the output to ensure the continuous flow of uncorrupted foundational spatial gradients during backpropagation [42].

3.6.4. Compound Loss Formulation and Regularization

Given the inherent complexities of high-resolution remote sensing including extreme class imbalances, label noise, and multi-modal feature dominance the network is optimized using a carefully weighted compound loss function alongside strategic modality dropout.

Compound Loss Function:
To stabilize training across both dominant background classes and rare land-cover typologies, the total loss function ($\mathcal{L}_{total}$) is formulated as a linear combination of four distinct objectives. During training, a component-wise loss dictionary is logged at each step to closely monitor individual optimization dynamics and ensure gradient stability across the multi-modal streams:

$$\mathcal{L}_{total} = \lambda_{focal}\mathcal{L}_{focal} + \lambda_{dice}\mathcal{L}_{dice} + \lambda_{ce}\mathcal{L}_{ce} + \lambda_{aux}\mathcal{L}_{aux}$$

where:

- $\mathcal{L}_{focal}$ is the Focal Loss ($\lambda_{focal} = 2.0$), heavily weighted to penalize hard-to-classify pixels and mitigate the effects of class imbalance.
- $\mathcal{L}_{dice}$ is the Dice Loss ($\lambda_{dice} = 1.0$), ensuring global topological agreement and maximizing the Intersection over Union (IoU) metric.
- $\mathcal{L}_{ce}$ is the Categorical Cross-Entropy loss ($\lambda_{ce} = 0.5$) utilized for baseline pixel-wise probability maximization. A label smoothing factor of 0.05 is applied to this component to slightly regularize predictions and prevent overconfident network activations.
- $\mathcal{L}_{aux}$ is an auxiliary Cross-Entropy loss ($\lambda_{aux} = 0.4$) computed on the intermediate 1/4 resolution output of the Multi-Spectral encoder (which is upsampled to full resolution prior to loss calculation). This deep supervision injects gradients directly into the early layers of the network, preventing vanishing gradients and ensuring the MS stream learns meaningful representations independently of the massive SAM backbone.

The relative weighting of these four terms was assigned from the role each plays within the objective, rather than by exhaustive search over the weight space, which the training cost of the full configuration made impractical. The focal term carries the largest weight as it is the only component that reweights individual pixels by classification difficulty, and is therefore the principal mechanism through which under-represented alpine typologies contribute to the gradient despite occupying a small fraction of the annotated area. Dice term instead is retained at unit weight as the reference of the formulation, since it optimises regional overlap directly and is the closest differentiable component for the Intersection over Union used to evaluate the model. Categorical cross-entropy is delibrately reduced to half the value: applied at full strength it is dominated by the spatially extensive classes, particularly the meadow and rock typologies that occupy majority of the annotated surface, and it serves here to maintain calibrarted pixel-wise probabilities rather than to drive the optimisation. The auxiliary term is assigned the smallest weight, its purpose being to supply gradient to the early layers of the multispectral encoder rather than to shape the

final prediction; the value is large enough to prevent vanishing gradients in that batch while remaining small enough that the intermediate 1/4-resolution output does not compete with the full resolution objective. The resulting ordering, $\lambda_{focal} > \lambda_{dice} > \lambda_{ce} > \lambda_{aux}$, therefore reflects a deliberate ranking of the four objectives against the characteristics of this dataset. A systematic sensitivity analysis over these weights was not performed and is identified in Section 6.3 as a target for subsequent work.

Modality Dropout Regularization:
Because the 1.1-billion parameter SAM backbone provides exceptionally strong spatial priors, there is a risk that the network will over-rely on the RGB stream and ignore the MS or DSM inputs during training. To force the network to extract generalized representations from all available sensors, we implement a specialized Modality Dropout during the training phase. At each iteration, the Multi-Spectral features have a 20% probability of being zeroed out, while the RGB and DSM features each have a 10% probability of being dropped. This sensor-level dropout ensures that the cross-attention fusion module remains highly robust to missing data and learns deeply synergistic, rather than redundant, cross-modal representations.

3.7. Inference Architecture and Operational Deployment

The final stage of the deep learning pipeline involves operational deployment for large-scale LULC mapping across extended alpine regions, a process termed inference. To ensure mathematical validity, the operational deployment phase utilizes an identical architectural framework, hyperparameter setting, and multi-modal feature engineering stack as the training phase described in Section 4. Maintaining strict architectural parity is an absolute prerequisite; the classification process relies entirely on the optimized, frozen internal parameters ("weights") acquired during training. Deploying an unaligned feature space or an uninitialized layer configuration would fundamentally disrupt weight mappings, leading to arbitrary and invalid outputs.

During large-scale operational deployment, the model operates over the full unannotated geospatial dataset. To successfully construct the co-registered multi-modal input required by the dual-encoder attention mechanism comprising the three-channel RGB tensor, the six-band multispectral tensor, the thirteen derived spectral index proxies listed in Table 2, and the single-channel normalised surface model the structural integrity of the multi-sensor payload must be preserved. Consequently, the input pipeline enforces a mandatory, co-registered three-part input stack for any active inference footprint:

1. **RGB Orthomosaic:** Captured via the high-resolution Phase One camera, utilized for extracting visual geometry and generating visible-band spectral proxies.
2. **Multispectral (MS) Orthomosaic:** Captured via the MicaSense sensor (Blue, Green, Red, RedEdge, NIR, and Thermal IR bands), serving as the foundational spectral data stream for computing the complex vegetation, soil, and moisture indices.
3. **Digital Surface Model (DSM):** Utilized for engineering normalized topographic proxies (such as relative height, slope, and aspect matrices) and driving post-inference correction logic.

Because the model's learned weights are intrinsically mapped to a multi-dimensional feature space rather than raw imagery channels alone, omitting any single sensor layer makes it impossible to reconstruct the 16-channel input layer. Operational deployment is explicitly contingent upon perfectly replicating the exact multi-modal feature environment that governed the training cycle.

To maximize classification precision and minimize spectral confusion across unseen high-altitude regions, the deployment pipeline freezes and loads the final optimal weights obtained at Epoch 30. These parameters represent the optimal network configuration

verified during the evaluation phase, striking a robust balance between spatial boundary precision and spectral generalization across the diverse topographic variations in the alpine test basins.

4. Results

4.1. Evaluation Context and Benchmarking Limitations

In traditional deep learning computer vision studies, novel architectures are typically benchmarked against standardized public datasets. However, the specific configuration of this study combining a custom aerial payload (synchronized Phase One RGB and MicaSense Multispectral sensors alongside nDSM data), sub-metre spatial resolution, and a highly specialized alpine LULC taxonomy renders direct benchmarking against standard remote sensing datasets (e.g., ISPRS Vaihingen/Potsdam, SEN12MS) scientifically incompatible.

Standard datasets lack either the precise multi-modal alignment, the extreme topographical variance, or the specific high-altitude vegetational and geological classes (e.g., highly granular meadow and rock categorizations) targeted in this research. Consequently, this study establishes a novel foundational baseline for high-resolution, multi-modal alpine segmentation. Model efficacy is therefore rigorously evaluated through internal validation metrics, tracking optimization convergence and spatial overlap fidelity across a spatially stratified, unseen testing subset.

4.2. Analysis of Loss Convergence

The Loss Function serves as the primary objective metric during the backpropagation phase, quantifying the mathematical disparity between the model's predicted semantic probabilities and the ground-truth annotations. The central objective of the stochastic gradient descent process is the systematic minimization of this value.

As depicted in Figure 6, optimal convergence is definitively evidenced by the simultaneous, monotonic reduction in both the Training Loss and the Validation Loss over the 30-epoch training cycle. The descending trajectory of the Training Loss (stabilizing near 0.50) confirms that the dual-encoder architecture effectively optimized its multi-modal cross-attention parameters based on the observed spatial and spectral priors. Concurrently, the analogous and tightly coupled reduction in Validation Loss (stabilizing near 0.60–0.80) signifies robust generalization. This synchronous decline is the critical indicator that the SAM-based architecture learned generalized, underlying physical properties of the alpine classes, successfully avoiding the pitfalls of overfitting or rote memorization of the training distribution.



Figure 6. Training and Validation Loss convergence over 30 epochs. The synchronous decline indicates robust optimization without overfitting to the training distribution.

4.3. Mean Intersection over Union (mIoU) Progression

While Loss quantifies mathematical error, the Mean Intersection over Union (mIoU) serves as the definitive, human-interpretable metric for evaluating semantic segmentation spatial precision. It quantifies the exact geographical overlap between the predicted LULC mask and the expert-delineated ground truth. Figure 7 tracks this performance metric across the training duration. Both the Training mIoU and Validation mIoU demonstrate a steady, continuous ascent from an initial untrained state of roughly 0.10. Crucially, the Validation mIoU closely tracks the Training mIoU throughout the entire process, indicating that the model achieved high classification competence not only on its training subset but maintained spatial boundary precision on previously unseen geographical regions.

The series plotted in Figure 7 is the mIoU logged during training, which our implementation computes across all fifteen output channels of the network, including an unlabelled background channel that carries no annotated support under the weakly supervised regime adopted here. That series is therefore computed on a different basis from the per-class figures reported below and sits correspondingly lower. Evaluated over the fourteen classes reported in Table 5, the macro-averaged validation mIoU at Epoch 30 is 0.777, and all mIoU values quoted elsewhere in this article refer to the fourteen-class quantity unless stated otherwise.



Figure 7. Mean Intersection over Union (mIoU) progression across the training cycle, computed across all fifteen output channels including the unlabelled background channel. Macro-averaged values over the fourteen evalauted land cover classes are reported in Table 5.

4.4. Loss-Accuracy Trade-Off (mIoU vs. Validation Loss)

To further validate the mathematical integrity of the training pipeline, a diagnostic analysis was conducted to map the relationship between objective function minimization (Loss) and the principal spatial metric (mIoU).

Figure 8 illustrates this critical correlation across the validation subset. The horizontal axis represents the Validation Loss, while the vertical axis represents the Validation mIoU. The temporal progression of the training epochs is color-encoded, transitioning from the initial epochs (dark purple) to the final, converged state (yellow). The resulting plot reveals a stark, linear trajectory from the bottom-right quadrant (high error, low accuracy) directly to the top-left quadrant (low error, high accuracy). This near-perfect negative correlation validates that the chosen loss formulation is highly effective for this specific task; as the mathematical prediction error was forced down, it translated directly and reliably into superior geographical segmentation accuracy.

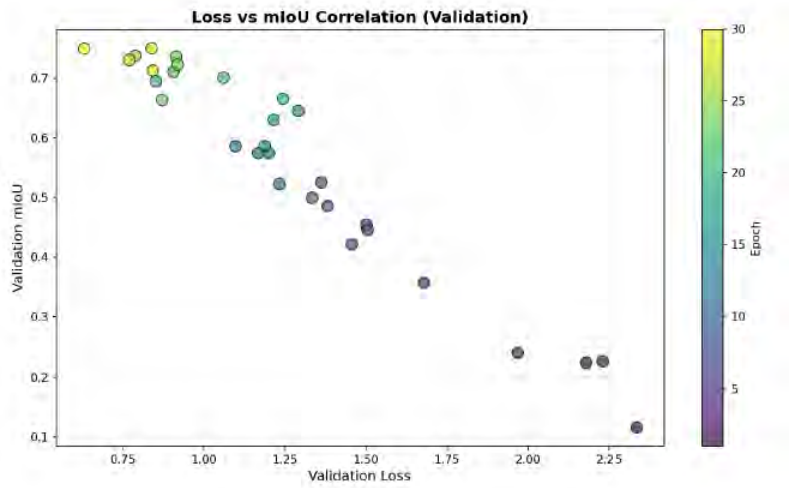


Figure 8. Diagnostic correlation between Validation Loss and Validation mIoU. The temporal color gradient (purple to yellow) tracks epoch progression, revealing a highly effective optimization trajectory.

4.5. Performance Analysis by Class

While global metrics provide an overarching view of model health, evaluating the per-class mIoU is critical for understanding the architecture's generalization capabilities across distinct topological features. A high, steady Training IoU confirms that the architecture is sufficiently deep to capture complex class-specific features, while the corresponding Validation IoU dictates its true predictive power in the wild.

Table 5 details the final mIoU scores extracted at Epoch 30. Overall, the network achieved exceptional performance, with 11 of the 14 monitored classes exhibiting robust generalization.

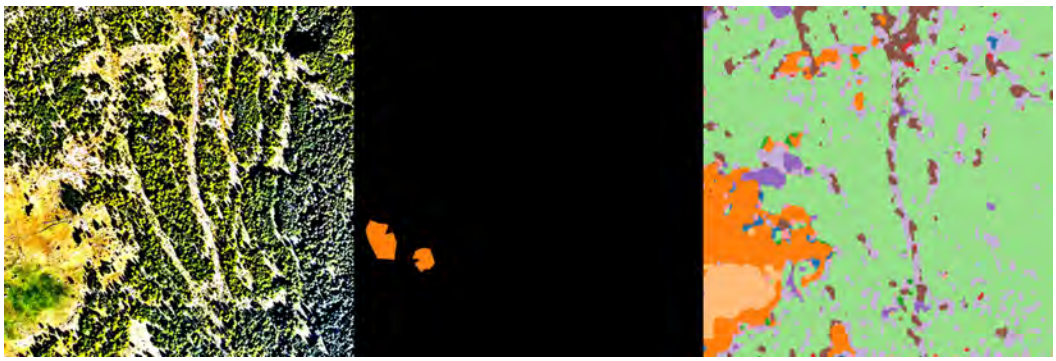
Robust Generalization (Top Performers): The vast majority of classes notably *Irrigated_Meadow*, *Snow*, *Lakes*, *Road*, *Buildings*, and *Rock* demonstrated excellent convergence. For these categories, the Validation IoU closely tracked the Training IoU, consistently stabilizing between 80% and 96%.

Anomalous Classes: Two classes depart markedly from this general pattern. *Shrub_Evergreen* achieves an IoU of 0.95 on the training set but only 0.40 on the validation set. This large generalisation gap is consistent with overfitting and suggests that the training samples do not adequately represent the intra-class variability encountered in the validation tiles. The class is also difficult to distinguish from other woody and shrubby vegetation because of their similar spectral and structural characteristics, particularly in transitional or shaded areas. An even larger discrepancy is observed for *Wetlands*, which achieves a training IoU of 1.00 but a validation IoU of 0.00. This result indicates that the model fitted the available training annotations but failed to transfer the learned representation to the validation samples. Wetlands are particularly challenging because they are spatially heterogeneous and may contain different proportions of vegetation, saturated soil, and shallow water. Their spectral response can therefore overlap with that of moist grassland, open water, and shadowed surfaces. In addition, the limited number and spatial diversity of the annotated wetland samples may make the validation IoU highly sensitive to individual prediction errors. These results indicate that additional expert-validated samples covering a broader range of wetland and evergreen-shrub conditions will be required to improve generalisation. The learning behaviour of these classes, together with the delayed learning onset observed for *Rocky_Foundation*, is further examined in Section 5.1.

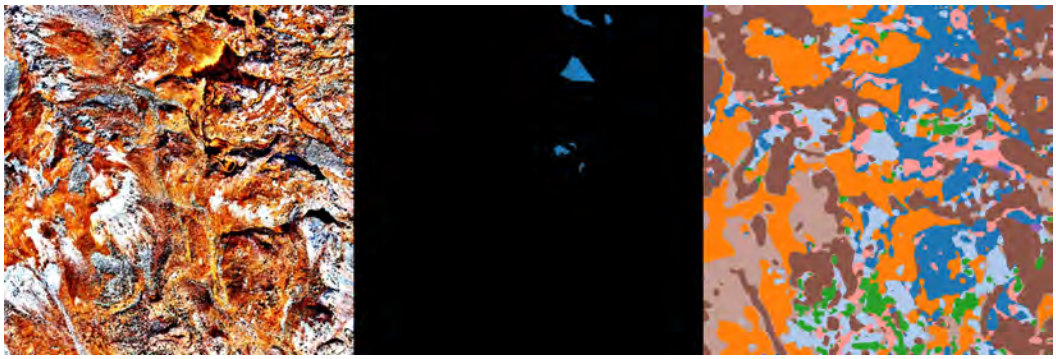
4.6. Qualitative Visual Assessment and Weakly Supervised Inference

While quantitative evaluation provides a statistical summary of network performance, qualitative visual assessment is critical to confirm the spatial fidelity and ecological boundary precision of the generated land cover maps. A defining challenge of this high-resolution alpine dataset is the logistical impossibility of hand-painting exhaustive, pixel-level training masks across expansive mountain basins. To address this limitation, the framework relies entirely on a weakly supervised learning paradigm. The training dataset consists only of sparse geometric polygon seeds, forcing the neural network to execute complex "guesswork" to infer class boundaries across the vast, unannotated background distribution based purely on learned multi-modal feature weights.

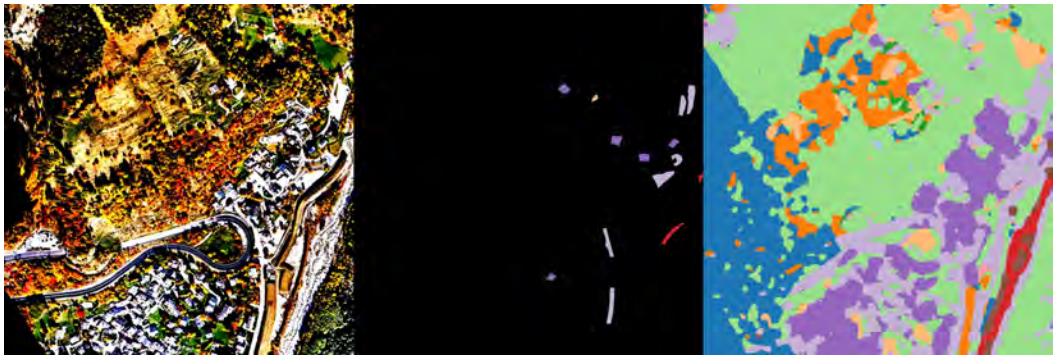
Figure 9 displays three representative validation tiles captured at Epoch 30. Each panel contrasts the raw aerial mosaic input against the sparse training polygons and the resulting dense classification map.



(**a**) Meadow typing (*Non_Irrigated_Meadows*).



(**b**) Wetland delineation (*Wetlands*).



(**c**) Infrastructure mapping (*Road, Buildings*).

Figure 9. Visual exemplars of dense semantic mapping driven by sparse, weakly supervised annotations at Epoch 30. Within each subfigure: (**Left**) displays the high-resolution RGB/Multispectral composite; (**Middle**) shows the manually delineated sparse training seeds; (**Right**) demonstrates the dense LULC prediction inferred across the background.

As qualitatively demonstrated across all three scenes, despite receiving minimal spatial priors (the isolated orange and purple training seeds in the center panels), the dual-encoder attention layer successfully modeled the underlying spectral–spatial patterns. By Epoch 30, the network successfully generalized its representations to classify the unmapped background, accurately capturing irregular boundaries and transitions in the vegetation without suffering from salt-and-pepper noise or breaking down in regions lacking dense label density.

4.7. Model Interpretability via Spatial Grad-CAM Diagnostics

To validate the internal decision-making mechanics of the dual-encoder framework prior to examining final map outputs, Gradient-weighted Class Activation Mapping (Grad-CAM) was executed during full-scene inference. Weakly supervised deep learning models can occasionally achieve high quantitative scores by leveraging dataset biases or spurious spatial context. Grad-CAM mitigates this "black box" risk by calculating the gradients of a targeted LULC class score with respect to the final convolutional feature maps in the multi-modal encoder. By aggregating and pooling these gradients, the algorithm projects a spatial attention heatmap directly over the imagery, explicitly detailing which physical features are driving the semantic categorization.

Figure 10 illustrates the spatial focus of the model's cross-attention encoders across four distinct alpine typologies during the inference stage on the full basin orthomosaics.

The diagnostic heatmaps confirm that the multi-modal fusion layer accurately isolates relevant physical properties for each class without relying on surrounding background artifacts. For example, in the *Buildings* map (Figure 10a), the high-intensity activation zones are strictly confined to the geometric boundaries of structural roofs, demonstrating the spatial precision of the frozen backbone. In the *FIUME* diagnostic (Figure 10c), the attention pathway seamlessly tracks the sinuous, low-reflectance flow lines of the river channel.

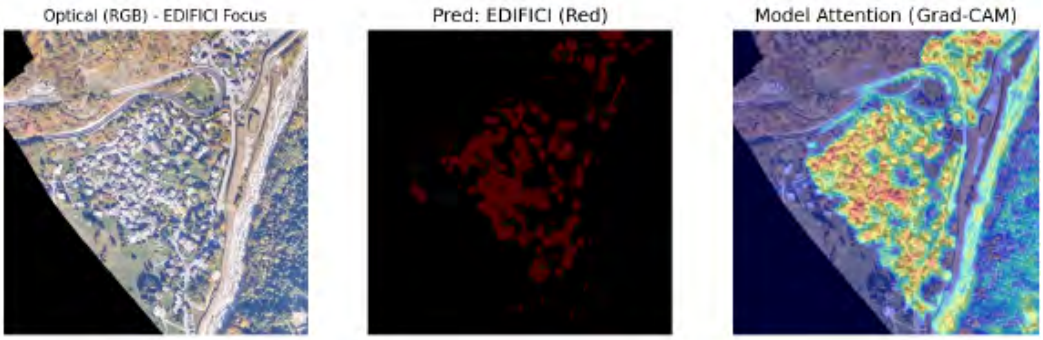
Crucially, for vegetative sub-classes like *Heath* and *Irriagted_Meadow* (Figure 10b,d), the network successfully separates distinct micro-topologies. It focuses on coarse texture transitions for heath layers and strong near-infrared (NIR) and red-edge signatures for cultivated meadows. This targeted focus confirms that the model's semantic interpretation is physically grounded, providing a highly reliable foundation for large-scale geographic deployment.

Post-Refinement via Non-AI Mathematical Operations

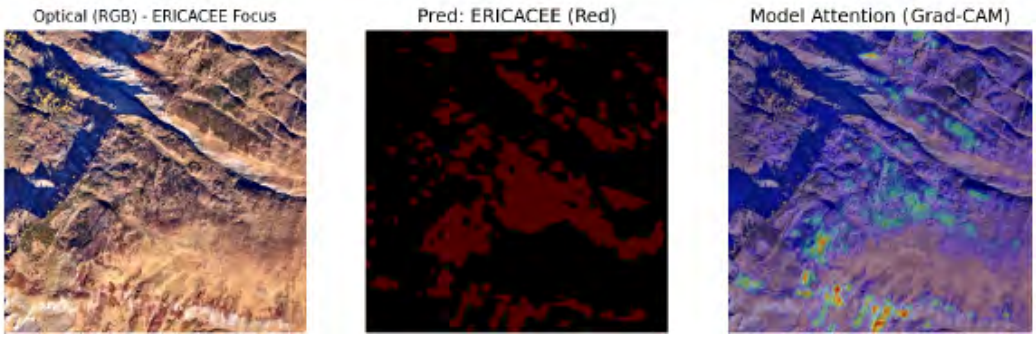
Because weakly supervised models are inherently designed to speculate across unannotated territory, raw neural network outputs naturally exhibit a degree of spatial bleeding or "blob-like" artifacts along highly complex transition margins. To protect the geometric integrity of the final map product, a completely non-AI, mathematically deterministic post-refinement stage was appended to the workflow using standard geospatial GIS operations.

First, rule-based height thresholding was applied directly to a Canopy Height Model (CHM) derived from the aerial nDSM layer. This purely mathematical filter dynamically overrides classification bleeding by sorting tall, vertical tree canopies (*Trees*) from completely flat meadow classes (*Meadow*) based on a strict geometric height cutoff, independently correcting the neural network's spectral outputs. Second, traditional mathematical edge-clipping and morphological filtering operations were performed on rigid architectural and linear features (such as roads and buildings) to restore clean, sharp structural vectors.

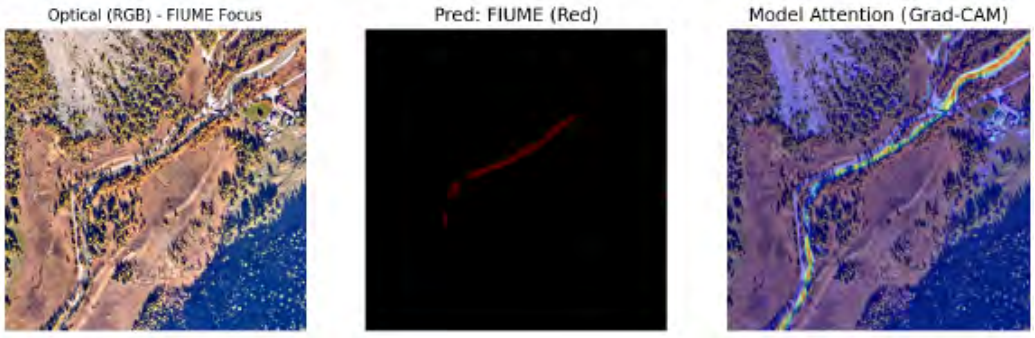
This strict separation of duties utilizing the deep learning network exclusively to solve the semantic categorization guesswork and using traditional, deterministic mathematical GIS rules for geometric edge refinement allows the pipeline to produce continuous, instance-level LULC orthomosaics without requiring costly, fully hand-painted training sets.



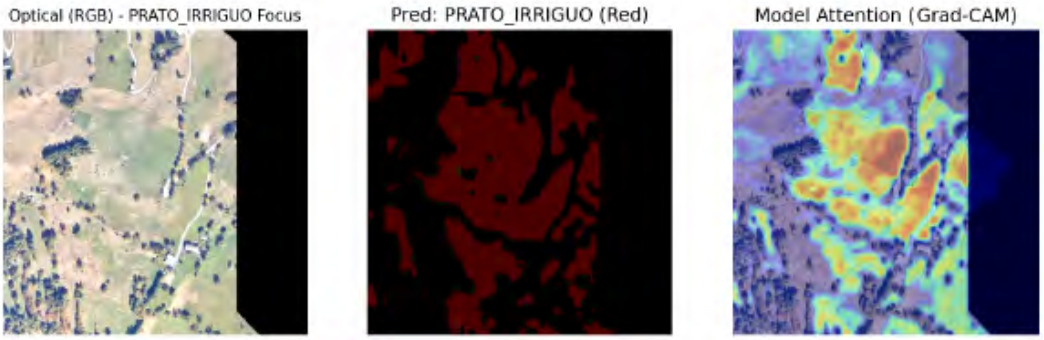
(**a**) Attention mapping for built infrastructure (*Buildings*).



(**b**) Attention mapping for alpine heath (*Heath*).



(**c**) Attention mapping for fluvial channels (*River*).



(**d**) Attention mapping for irrigated meadows (*Irrigated_Meadow*).

Figure 10. Grad-CAM activation maps displaying the internal attention focus of the neural network across diverse alpine land cover classes during full-mosaic inference. High-intensity zones (red/yellow) demonstrate that the model successfully isolates correct class-specific spectral and geometric structures.

5. Discussion

5.1. Interpretation of Learning Dynamics and Anomalies

The per-class metrics presented in Section 4 confirm that the multi-modal fusion mechanism successfully learned intrinsic physical parameters (such as elevation signatures from the nDSM and spectral reflectance from the MS bands) rather than memorizing spatial locations for the majority of the dataset. However, an analysis of the epoch-by-epoch logs reveals distinct temporal learning patterns and data anomalies that warrant deeper discussion.

Delayed Learning Patterns in Complex Topologies: Typologies such as *Rocky_Foundation* and *Shrub_Evergreen* exhibited a "delayed start," where validation accuracy remained near zero until approximately Epoch 10 before initiating a sharp ascent. This behavior suggests these classes possess subtle or morphologically ambiguous boundaries. The network naturally prioritized optimizing weights for mathematically distinct, high-contrast features (e.g., lakes, buildings) before successfully resolving the granular, high-variance feature space of complex terrestrial transitions.

Overfitting and Evaluation Artifacts: While the global network successfully generalized, two specific anomalies were identified. First, *Shrub_Evergreen* exhibited clear signs of overfitting, reaching 0.95 mIoU during training but stalling at 0.40 during validation. This significant divergence indicates that the intra-class variance present in the validation set was not adequately represented in the training distribution, causing the model to memorize specific shrub topologies rather than learning a generalized representation.

More generally, the per-class results in Table 5 indicate that residual error is concentrated between spectrally and structurally adjacent typologies rather than distributed uniformly. *Shrub_Evergreen* lies at the boundary of the surrounding vegetation classes, sharing much of its spectral response with heath and the meadow typologies while differing principally in canopy structure at a scale close to the resolution limit of the patch-based backbone. *Rocky_Foundation*, at 0.80, is similarly the weaker member of a pair with *Rock* at 0.88, the two differing in surface continuity rather than in mineral spectral signature. The anthropic classes exhibit the same pattern in milder form, *Buildings* and *Road* both being impervious surfaces separated largely by geometry and elevation. The classes that converged most strongly are correspondingly those with an unambiguous physical signature: open water, snow and dense canopy. This distribution of error is consistent with the interpretation advanced in Section 5.2, that the multispectral stream carries the discriminative burden, since the classes that remain difficult are precisely those whose separation is not primarily spectral.

The *Wetlands* class presents a more severe case. It registers a training IoU of 1.00 against a validation IoU of 0.00, a combination indicating that the network fitted the training annotations without acquiring any representation transferable to unseen tiles. Wetlands are widely reported as among the most difficult typologies to delineate from optical imagery: their spectral response is governed by a variable mixture of vegetation, saturated substrate and standing water, and the resulting intra-class variability routinely defeats classifiers that perform well on adjacent land cover [50,51]. The difficulty is compounded by the sensor configuration employed here. The Altum-PT provides blue, green, red, red-edge, near-infrared and thermal bands but no shortwave infrared channel, and the moisture-sensitive indices most commonly used to separate saturated ground from dry vegetation of which the Modified NDWI (Normalised Difference Water Index) is the canonical example are defined on shortwave infrared reflectance and could therefore not be computed for this dataset [52]. The residual confusion observed for this class does not involve open water, from which *Wetlands* remains well separated, but the low vegetation typologies

Shrub_Evergreen itself the second weakest class in the schema and *Heath*, which occupy the wetland margins and into which a substantial proportion of the annotated wetland extent grades. We attribute this tentatively to sub-canopy moisture that the available bands cannot observe directly, although the present data do not allow the mechanism to be established. Comparable studies report that spectral features alone are insufficient to resolve co-occurring wetland vegetation and require the addition of textural, geometric or LiDAR-derived information [53,54].

The consequence for the aggregate metric is substantial: excluding this class alone raises the macro-averaged validation mIoU from 0.777 to 0.837, indicating that the remaining thirteen typologies were learned robustly and that a single ambiguous class accounts for the greater part of the shortfall. We do not regard this result as a settled characterisation of the class, and identify the wetland typology as the principal target for further work, requiring a substantially larger and more geographically diverse annotation set together with ancillary evidence multi-temporal moisture dynamics, or shortwave infrared reflectance from a complementary sensor that a single-date six-band acquisition cannot supply; this is identified as a priority in Section 6.3.

5.2. Ablation and Architectural Optimization Dynamics

To peek inside the "black box" of the dual-stream architecture, a post-hoc diagnostic analysis was conducted to evaluate gradient flow, learning rate scheduling, and individual modality contributions. These internal metrics reveal critical insights into how the network prioritizes heterogenous geospatial data based on its underlying code structure.

Modality Importance and Spectral Dominance:
A permutation-based ablation study was executed at Epochs 10 and 20 to quantify the reliance of the fusion module on each specific input stream. As illustrated in Figure 11, the network exhibits an overwhelming reliance on the Multispectral (MS) and Spectral Index tensors. Removing the MS stream resulted in a catastrophic mIoU drop of over 35%. Conversely, masking the nDSM resulted in a marginal performance drop of less than 5%.

This behavior is a direct, intended result of both the environmental physics and the architectural design. In alpine environments, elevation is highly ambiguous for land cover (e.g., both forests and exposed rock exist across vast altitudinal gradients). Spectral signatures directly measure vegetation health and mineral composition, making them inherently more discriminative. Architecturally, the 'DSMAdapter' is intentionally lightweight, comprising only two DilatedResBlocks projecting to 256 channels. In contrast, the MS encoder utilizes four full stages with Squeeze-and-Excitation (SE) attention across all spectral bands. Furthermore, because the applied top-hat transform extracts relative topological features but removes absolute elevation context, the network's 'GatedDSMFusion' module dynamically learns to suppress the nDSM stream, confirming that the MS data is the genuinely reliable primary signal for land cover classification.

Gradient Flow and Frozen Backbone Dynamics:
Tracking the L_2 norm of the gradients per component (Figure 12) provides a mechanical explanation for the model's convergence. The MS Encoder consistently registered gradient norms an order of magnitude higher than the SAM RGB backbone. This is the expected behavior of the configuration paradigm: the 1.1-billion parameter SAM ViT-H backbone was strictly frozen to preserve its generalized spatial priors and does not compute dense gradients. Its minor gradient footprint is derived solely from adapter injections and cross-attention backpropagation. Conversely, the MS encoder is trained entirely from scratch; its randomly initialized weights must rapidly adapt to extract meaningful spectral features, resulting in the massive learning signals observed in the gradient logs.

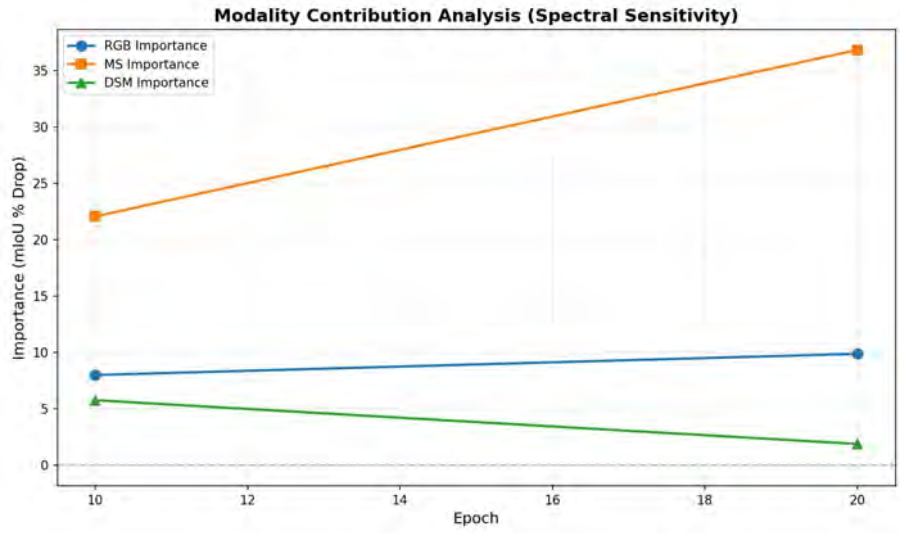


Figure 11. Modality Contribution Analysis. The network relies heavily on the MS stream (orange) to extract discriminative LULC features, while the 'GatedDSMFusion' module dynamically suppresses the highly ambiguous nDSM topography (green).

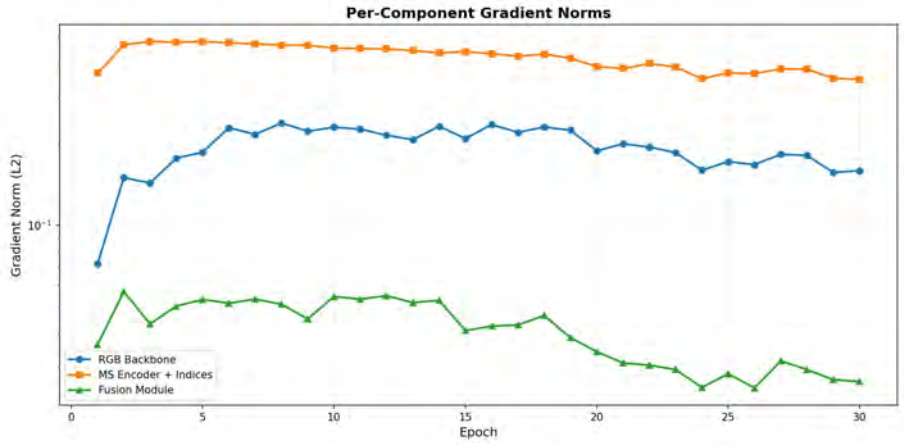


Figure 12. Log-scale tracking of L_2 gradient norms. The heavily frozen SAM RGB backbone exhibits minimal updates, while the MS encoder absorbs the massive gradients required to learn spectral feature extraction from scratch.

Plateau Scheduling and Dynamic Regularization:
The optimization process was governed by a 'ReduceLROnPlateau' scheduler (factor: 0.5, patience: 5) rather than a fixed step-decay. As seen in Figure 13, the learning rate halved precisely at Epoch 20. This triggered because the validation loss plateaued around Epoch 15. The precise reduction in learning rate correlates directly with the stabilization of gradient norms across all sub-modules, allowing the optimizer to take smaller, refined steps into local minima for the final 10 epochs.

Finally, to enforce multi-modal reliance and prevent the network from collapsing into a single-sensor dependency, dynamic modality dropout was applied. While individual modalities faced dropout rates of 10% to 20%, the cumulative logic structure of the training loop dictated that approximately 40% of all batches experienced at least one form of targeted sensor masking (Figure 14). This aggressive, compound regularization forced the cross-attention module to build highly resilient representations regardless of fluctuating sensor availability.

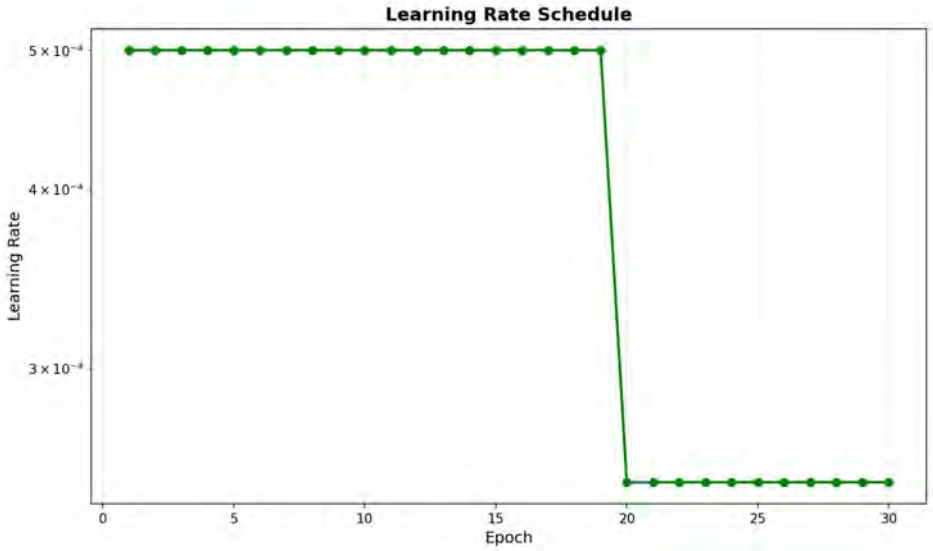


Figure 13. Learning rate schedule under the 'ReduceLROnPlateau' policy. The rate halves at Epoch 20 following the validation loss plateau near Epoch 15.

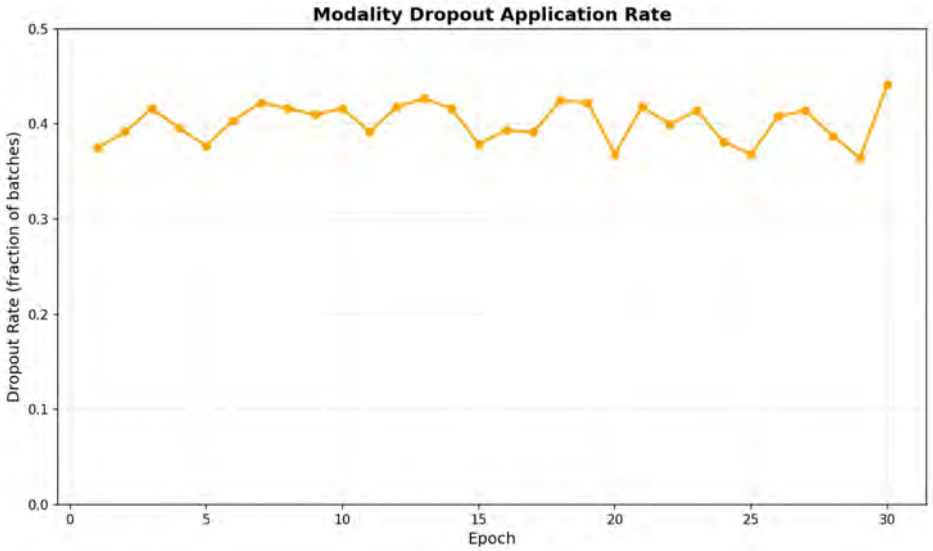


Figure 14. Per-batch modality dropout incidence during training. Individual modalities are masked at 10–20%, with roughly 40% of all batches experiencing at least one targeted sensor masking event.

The different native resolutions of the input data represent a potential limitation. Resampling the 86 cm multispectral imagery to the finer common grid enables its integration with the high-resolution inputs but does not generate additional spatial detail. The multispectral branch should therefore be interpreted as providing spatially coarser spectral context, while the RGB/nDSM inputs provide finer information on object structure and boundaries. Interpolation can also smooth spectral transitions, and small residual co-registration errors may affect predictions near narrow or fragmented land-cover elements. The dual-encoder architecture partially addresses this discrepancy by processing the spatial and spectral information through dedicated branches before feature fusion; however, explicit multiresolution fusion at native sensor resolutions remains a relevant direction for future research.

6. Conclusions

This study presented a novel multi-modal deep learning framework for high-resolution LULC mapping in complex alpine environments using sparsely annotated aerial imagery. The proposed architecture combines a SAM-based RGB encoder with a dedicated multispectral encoder, integrates spectral indices and DSM information through cross-modal fusion, and incorporates a deterministic post-processing strategy to improve boundary delineation. The proposed workflow was specifically designed to address the challenges of alpine landscapes, where strong topographic variability, heterogeneous vegetation, seasonal snow cover, and limited annotated data often reduce the effectiveness of conventional remote sensing approaches.

The experimental evaluation demonstrated that the proposed framework successfully learned robust semantic representations despite relying exclusively on weak supervision. The simultaneous decrease in training and validation losses throughout the optimization process, together with the parallel increase in validation mIoU, confirms that the proposed architecture generalizes well without exhibiting significant overfitting. The final validation mIoU of approximately 75% and the high agreement between training and validation metrics indicate that the network effectively exploits complementary spatial and spectral information to generate reliable semantic segmentation across previously unseen areas.

The current work post-hoc diagnostics revealed that while the SAM provides an exceptionally robust spatial prior, multispectral data remains the overwhelmingly dominant signal for accurate alpine LULC segmentation, whereas relative topological features (nDSM) provide only marginal disambiguation. The architecture achieved highly robust generalization across 13 of 14 evaluated complex alpine typologies (including snow, lakes, and distinct meadow categorizations), with validation mIoU scores consistently tracking training performance.

By detailing internal network dynamics such as the delayed learning onset for morphologically ambiguous classes and the necessity of aggressive modality dropout this work establishes a transparent and highly resilient foundational baseline for future instance-level environmental monitoring in rapidly changing, high-altitude Alpine environments.

6.1. Broader Impact Through Public–Private Collaboration

Beyond its technical contribution, this study was made possible by the "Test Before Invest" initiative, an EU funding mechanism that covers the full cost of piloting emerging digital services that a public body would not normally be in a position to procure or justify on its own. For ARPA VdA, this meant the opportunity to test an approach with no equivalent among established commercial offerings, and to help shape, from within, an operational workflow suited to its actual monitoring needs rather than one dictated by whatever a vendor happens to sell.

What made the collaboration work was arguably less about the technology itself and more about timing: the people who understand the monitoring problem and the people able to build a solution for it were brought together before the architecture had been fixed, not after. Public agencies hold the ground-level understanding of what a given land-cover class means operationally and what decisions a resulting map is meant to support; industrial partners hold the capacity to turn that understanding into a working system. Neither side could have produced this framework in isolation. As high-resolution aerial monitoring becomes a more routine part of environmental agency work, this kind of early, close cooperation between demand and supply seems more likely to yield genuinely usable services than a conventional client–vendor handoff arranged after the fact.

6.2. Limitations

A more technical limitation deserves to be stated plainly. SAM's ViT-H backbone is large and, on paper, should carry every advantage a transformer can offer, yet that scale comes at a real cost to spatial resolution. Patch-based tokenization suits natural images at moderate resolution reasonably well, but it does not translate as cleanly to centimetre-level aerial imagery, and the mismatch only sharpens once training has to proceed from sparse rather than dense labels [55,56]. Evidence from outside this study's domain even suggests that scaling a SAM backbone up buys comparatively little: one recent evaluation found only a 1.27% mIoU gain from ViT-H over ViT-B despite roughly four times the inference cost [57], which is broadly consistent with what we observed here the backbone recognised most of the fourteen classes in this study, but a purpose-built convolutional stream might have preserved fine boundary detail more cheaply than simply relying on a larger transformer. A genuine hybrid CNN–ViT design, rather than a transformer used more or less as-is, may well close part of this gap, a point we return to below.

The multispectral stream was weighted more heavily in the fusion module for a related reason: transformers in this setting still struggled to make full use of sparse labels, even where the underlying spectral signal was rich, a pattern also reported for CNN-based segmentation trained on incomplete ground truth [58,59]. Before manual annotation began, an unsupervised k-means pass over the imagery was used to surface candidate spectral clusters, which domain experts at ARPA VdA then used to narrow down and define the classification scheme eventually adopted in this study. That ordering, clustering first and defining classes second, is itself a partial answer to a constraint this project faced directly: although field expertise was present throughout the labelling process, understanding label accuracy per class on these new classes was still limited within the duration of this work. We see the pipeline described in this paper as one workable response to that constraint, rather than a limitation to be designed around later.

6.3. Future Work

The results reported here are best read as one snapshot of a design that continues to be iterated on. The most immediate open question is architectural: whether a transformer-only spatial backbone remains the right long-term choice for this task, or whether a genuine hybrid convolutional–transformer design would retain boundary detail better under the same weak-supervision regime. Related work adapting SAM through auxiliary CNN encoders and low-rank adaptation for aerial land-cover mapping points in this direction [60], and benchmarking a pure-transformer configuration against such a hybrid under identical sparse-label conditions is the comparison we consider most useful to pursue next.

That comparison cannot really be separated from the hardware it has to run on. A larger backbone or an additional adapter branch is only worth adopting if it fits inside the GPU memory actually available for training, and in a weakly supervised setting where labelled data, not compute, is usually the scarcer resource, spending a fixed budget on a bigger transformer instead of a leaner hybrid CNN branch is a genuine trade-off rather than a formality. Future iterations of this framework will treat available compute as a design constraint from the outset, rather than as an afterthought once an architecture has already been chosen, so that whichever configuration is eventually selected is one an environmental agency could realistically train and maintain without specialised infrastructure.

Author Contributions: Conceptualization, P.D., L.O. and U.M.d.C.; methodology, D.S.M.R., L.O. and G.F.; software, D.S.M.R.; validation, P.D., L.O. and U.M.d.C.; formal analysis, D.S.M.R. and L.P.; investigation, L.O. and G.F.; resources, D.S.M.R., L.O. and U.M.d.C.; data curation, D.S.M.R. and L.P.; writing—original draft preparation, D.S.M.R.; writing—review and editing, P.D., D.S.M.R.

and U.M.d.C.; supervision, P.D.; project administration, L.O.; funding acquisition, P.D. and L.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was done under the CHEDIH project (Circular Health European Digital Innovation Hub), operating as a 'Test Before Invest' initiative for the health and agrifood sector. CHEDIH is a partnership co-financed by the European Union and MIMIT under the EU Digital Europe Programme and the Italian PNRR Next Generation EU (M4C2 – Investment 2.3). It operates under Grant Agreement No. 101083745 and the relevant Italian funding provisions. CHEDIH partners act as service providers and operate in compliance with applicable EU and national regulations, including the DNSH principle, GBER requirements, and de minimis rules.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The aerial multispectral imagery, orthoproducts, DEM/DSM, and derived spectral index stacks are not publicly deposited due to proprietary restrictions and are available from the corresponding author upon reasonable request subject to a data sharing agreement. The deep learning segmentation model code will be made available upon publication via a Zenodo repository with a DOI, which will be provided in the final published article.

Acknowledgments: During the preparation of this manuscript, the authors used Claude (Anthropic; Claude Opus and Claude Sonnet models, accessed through Claude Code (Opus 5) and Google Gemini (Gemini Pro 3.1) to assist with limited editorial tasks, including drafting and refining selected text passages, formatting citations, and generating or editing selected schematic figures. The AI tools were not used to generate, collect, analyze, or interpret the scientific data presented in this study. All datasets, experiments, analyses, and scientific conclusions are entirely the authors' own work. Every AI-assisted output was carefully reviewed, edited, and verified by the authors, including independent verification of all cited references against their original publications.

Conflicts of Interest: Authors D.S.M.R., L.O., L.P. are employed by the company DigiSky s.r.l. Authors U.M.D.C. and G.F. are affiliated with ARPA Valle d'Aosta, which benefited from a service provided by DigiSky s.r.l., financially supported by the Test before Invest program. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funding sponsor (Test before Invest program) had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Belcore, E.; Piras, M.; Wozniak, E. Specific Alpine Environment Land Cover Classification Methodology: Google Earth Engine Processing for Sentinel-2 Data. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2020**, *43*, 663–670. [CrossRef]
2. Remondino, F.; Nocerino, E.; Toschi, I.; Menna, F. A Critical Review of Automated Photogrammetric Processing of Large Datasets. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2017**, *XLII-2/W8*, 191–199.
3. Adler, C.; Wester, P.; Bhatt, I.; Huggel, C.; Insarov, G.; Morecroft, M.; Muccione, V.; Prakash, A.; Alcántara-Ayala, I.; Allen, S.K.; et al. Cross-Chapter Paper 5: Mountains. In *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Pörtner, H.-O., Roberts, D.C., Tignor, M.M.B., Poloczanska, E., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., et al., Eds.; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2022; pp. 2273–2318. [CrossRef]
4. Alpine Convention; Alpine Biodiversity Board. *Cross-Cutting Biodiversity Issues: Report 2024*; Permanent Secretariat of the Alpine Convention: Innsbruck, Austria, 2024; pp. 1–30. Available online: https://www.alpconv.org/fileadmin/user_upload/Organisation/TWB/ABB/Cross-cutting_biodiversity_issues_Report_2024.pdf (accessed on 2 September 2026).
5. Qin, R.; Liu, T. A Review of Landcover Classification with Very-High Resolution Remotely Sensed Optical Images—Analysis Unit, Model Scalability and Transferability. *Remote Sens.* **2022**, *14*, 646. [CrossRef]
6. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 4015–4026.

7. Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 27672–27683. [CrossRef]
8. Zhang, Y.; Ru, L.; Wu, K.; Yu, L.; Liang, L.; Li, Y.; Chen, J. SkySense V2: A Unified Foundation Model for Multi-Modal Remote Sensing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 9136–9146. [CrossRef]
9. Wu, K.; Zhang, Y.; Ru, L.; Dang, B.; Lao, J.; Yu, L.; Luo, J.; Zhu, Z.; Sun, Y.; Zhang, J.; et al. A Semantic-Enhanced Multi-Modal Remote Sensing Foundation Model for Earth Observation. *Nat. Mach. Intell.* **2025**, *7*, 1235–1249. [CrossRef]
10. Bi, H.; Feng, Y.; Tong, B.; Wang, M.; Yu, H.; Mao, Y.; Chang, H.; Diao, W.; Wang, P.; Yu, Y.; et al. RingMoE: Mixture-of-Modality-Experts Multi-Modal Foundation Models for Universal Remote Sensing Image Interpretation. *arXiv* **2025**, arXiv:2504.03166. [CrossRef]
11. Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A.J.; Hanna, J.; Borth, D.; Papoutsis, I.; Le Saux, B.; Camps-Valls, G.; Zhu, X.X. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. *arXiv* **2024**, arXiv:2403.15356. [CrossRef]
12. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [CrossRef] [PubMed]
13. Zhang, P.; Peng, B.; Lu, C.; Huang, Q. ASANet: Asymmetric Semantic Aligning Network for RGB and SAR Image Land Cover Classification. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 574–587. [CrossRef]
14. Zhang, X.; Yokoya, N.; Gu, X.; Tian, Q.; Bruzzone, L. Local-to-Global Cross-Modal Attention-Aware Fusion for HSI-X Semantic Segmentation. *arXiv* **2024**, arXiv:2406.17679. [CrossRef]
15. Fan, J.; Shi, Z.; Du, Y.; Zhuang, C. HR-MM SegFormer: Enhancing Land Use and Land Cover Semantic Segmentation through Transformer-Based Multisource Remote Sensing Feature Fusion. *Environ. Model. Softw.* **2026**, *197*, 106848. [CrossRef]
16. Hu, J.M.; Shean, D. Improving Mountain Snow and Land Cover Mapping Using Very-High-Resolution (VHR) Optical Satellite Images and Random Forest ML. *Remote Sens.* **2022**, *14*, 4227. [CrossRef]
17. Florath, J.; Keller, S.; Abarca-del-Rio, R.; Hinz, S.; Staub, G.; Weinmann, M. Glacier Monitoring Based on Multi-Spectral and Multi-Temporal Satellite Data: A Case Study for Classification with Respect to Different Snow and Ice Types. *Remote Sens.* **2022**, *14*, 845. [CrossRef]
18. Spooner, C.; Scheck-Wenderoth, M.; Cacace, M.; Anikiev, D. How Alpine Seismicity Relates to Lithospheric Strength. *Int. J. Earth Sci.* **2022**, *111*, 1129–1148. [CrossRef]
19. Zhang, Y.; Han, F.; Zhou, M.; Hou, Y.; Wang, S. Application of Deep Learning in Glacier Boundary Extraction: A Case Study of the Tomur Peak Region, Tianshan, Xinjiang. *Sustainability* **2025**, *17*, 3678. [CrossRef]
20. Peel, M.C.; Finlayson, B.L.; McMahon, T.A. Updated World Map of the Köppen–Geiger Climate Classification. *Hydrol. Earth Syst. Sci.* **2007**, *11*, 1633–1644. [CrossRef]
21. Prieur, C.; Rabatel, A.; Thomas, J.-B.; Farup, I.; Chanussot, J. Machine Learning Approaches to Automatically Detect Glacier Snow Lines on Multi-Spectral Satellite Images. *Remote Sens.* **2022**, *14*, 3868. [CrossRef]
22. Kaushik, S.; Singh, T.; Bhardwaj, A.; Joshi, P.K.; Dietz, A.J. Automated Delineation of Supraglacial Debris Cover Using Deep Learning and Multisource Remote Sensing Data. *Remote Sens.* **2022**, *14*, 1352. [CrossRef]
23. Feyisa, G.L.; Meilby, H.; Fensholt, R.; Proud, S.R. Automated Water Extraction Index (AWEI): A New Technique for Water Mapping. *Remote Sens. Environ.* **2014**, *152*, 657–667. [CrossRef]
24. Adams, M.S.; Fromm, R.; Graf, A.; Bühler, Y.; Stoffel, A.; Boesch, R.; Ginzler, C. *Investigating the Potential of Low-Cost Remotely Piloted Aerial Systems for Monitoring the Alpine Snow Cover (RPAS4SNOW)*; Austrian Academy of Sciences Press: Vienna, Austria, 2016. [CrossRef]
25. CHEDIH Project Consortium. CHEDIH: Changing European Highland Environment: Dynamics, Interactions, and Hazards; European Commission Horizon Europe Project. Available online: https://www.chedih.eu/wall (accessed on 5 May 2026).
26. Bouhennache, R.; Bouden, T.; Taleb-Ahmed, A.; Cheddad, A. A New Spectral Index for the Extraction of Built-Up Land Features from Landsat 8 Satellite Imagery. *Geocarto Int.* **2019**, *34*, 1531–1551. [CrossRef]
27. Rikimaru, A.; Roy, P.S.; Miyatake, S. Tropical Forest Cover Density Mapping. *Trop. Ecol.* **2002**, *43*, 39–47.
28. Gitelson, A.A.; Gritz, Y.; Merzlyak, M.N. Relationships between Leaf Chlorophyll Content and Spectral Reflectance and Algorithms for Non-Destructive Chlorophyll Assessment in Higher Plant Leaves. *J. Plant Physiol.* **2003**, *160*, 271–282. [CrossRef] [PubMed]
29. As-syakur, A.R.; Adnyana, I.W.S.; Arthana, I.W.; Nuarsa, I.W. Enhanced Built-Up and Bareness Index (EBBI) for Mapping Built-Up and Bare Land in an Urban Area. *Remote Sens.* **2012**, *4*, 2957–2970. [CrossRef]
30. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a Green Channel in Remote Sensing of Global Vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
31. Xu, H. A New Index for Delineating Built-Up Land Features in Satellite Imagery. *Int. J. Remote Sens.* **2008**, *29*, 4269–4276. [CrossRef]

32. Zha, Y.; Gao, J.; Ni, S. Use of Normalized Difference Built-Up Index in Automatically Mapping Urban Areas from TM Imagery. *Int. J. Remote Sens.* **2003**, *24*, 583–594. [CrossRef]
33. Gitelson, A.A.; Merzlyak, M.N. Spectral Reflectance Changes Associated with Autumn Senescence of *Aesculus hippocastanum* L. and *Acer platanoides* L. Leaves. Spectral Features and Relation to Chlorophyll Estimation. *J. Plant Physiol.* **1994**, *143*, 286–292. [CrossRef]
34. Tian, Y.; Chen, H.; Song, Q.; Zheng, K. A Novel Index for Impervious Surface Area Mapping: Development and Validation. *Remote Sens.* **2018**, *10*, 1521. [CrossRef]
35. Haboudane, D.; Miller, J.R.; Tremblay, N.; Zarco-Tejada, P.J.; Dextraze, L. Integrated Narrow-Band Vegetation Indices for Prediction of Crop Chlorophyll Content for Application to Precision Agriculture. *Remote Sens. Environ.* **2002**, *81*, 416–426. [CrossRef]
36. Kawamura, M.; Jayamana, S.; Tsujiko, Y. Relation between Social and Environmental Conditions in Colombo Sri Lanka and the Urban Index Estimated by Satellite Remote Sensing Data. *Int. Arch. Photogramm. Remote Sens.* **1996**, *31*, 321–326.
37. Shen, L.; Li, C. Water Body Extraction from Landsat ETM+ Imagery Using Adaboost Algorithm. In Proceedings of the 18th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010; pp. 1–4. [CrossRef]
38. Feranec, J. Chapter 2 Project CORINE Land Cover. In *European Landscape Dynamics: CORINE Land Cover Data*; CRC Press: Boca Raton, FL, USA, 2016. [CrossRef]
39. Feranec, J.; Hazeu, G.; Kosztra, B.; Arnold, S. Chapter 3 CORINE Land Cover Nomenclature. In *European Landscape Dynamics: CORINE Land Cover Data*; CRC Press: Boca Raton, FL, USA, 2016. [CrossRef]
40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
41. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*; Association for Computational Linguistics: Kerrville, TX, USA, 2018; Volume 2, pp. 464–468.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
43. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
44. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 1251–1258.
45. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2015; Volume 37, pp. 448–456.
46. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* ***2016***, *arXiv:1511.07122.*
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2017**, *30*, 5998–6008.
48. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
49. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2016**, arXiv:1606.08415.
50. Mahdavi, S.; Salehi, B.; Granger, J.; Amani, M.; Brisco, B.; Huang, W. Remote Sensing for Wetland Classification: A Comprehensive Review. *GIScience Remote Sens.* **2018**, *55*, 623–658. [CrossRef]
51. He, Y.; Li, M.; Tang, X.; Ren, P. Remote Sensing-Based Detailed Wetland Classification: A Review of Advances from 2020 to 2025. *Front. Remote Sens.* **2026**, *7*, 1852249. [CrossRef]
52. Xu, H. Modification of Normalised Difference Water Index (NDWI) to Enhance Open Water Features in Remotely Sensed Imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]
53. Van Alphen, R.; Rains, K.C.; Malservisi, R.; Rodgers, M.; Dixon, T.H. UAV-Based Wetland Monitoring: Multispectral and Lidar Fusion with Random Forest Classification. *Drones* **2024**, *8*, 113. [CrossRef]
54. Zhao, P.; Meng, R.; Xu, B.; Wu, J.; Shen, Y.; Liu, J.; Huang, B.; Yin, T. Improved Grass Species Mapping in High-Diversity Wetland by Combining UAV-Based Spectral, Textural, Geometric Measurements. *Remote Sens.* **2026**, *18*, 927. [CrossRef]
55. Gui, B.; Bhardwaj, A.; Sam, L. Evaluating the Efficacy of Segment Anything Model for Delineating Agriculture and Urban Green Spaces in Multiresolution Aerial and Spaceborne Remote Sensing Images. *Remote Sens.* **2024**, *16*, 414. [CrossRef]
56. Sahay, R.; Savakis, A. On Aligning SAM to Remote Sensing Data. In *Proceedings of the SPIE Geospatial Informatics XIV*; SPIE: Bellingham, WA, USA, 2024; Volume 13037, pp. 10–18.
57. Anchan, S.H. The Diminishing Returns of Scale: A Systematic Evaluation of Segment Anything Model Variants Across Image Characteristics. *Zenodo* **2026**. [CrossRef]
58. Hua, Y.; Marcos, D.; Mou, L.; Zhu, X.X.; Tuia, D. Semantic Segmentation of Remote Sensing Images with Sparse Annotations. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8006305. [CrossRef]

59. Maggiolo, L.; Marcos, D.; Moser, G.; Serpico, S.B.; Tuia, D. A Semisupervised CRF Model for CNN-Based Semantic Segmentation with Sparse Ground Truth. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5606315. [CrossRef]
60. Xue, X.; Cheng, H.; Yang, R.; Wang, T.; He, X. Adapting Segment Anything Model to Aerial Land Cover Classification with Low-Rank Adaptation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2502605. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.