Attention-driven Cross-modal Self-supervised Learning for Label-efficient Hyperspectral-LiDAR DSM Classification

Jonathan González Santiago[1], Wolfgang Gross[1], Karsten Schulz[1], Wolfgang Middelmann[1], Uwe Soergel[2]

[1]Fraunhofer IOSB, Germany
(jonathan.gonzalez-santiago, wolfgang.gross, karsten.schulz, wolfgang.middelmann)@iosb.fraunhofer.de
[2]Institute for Photogrammetry and Geoinformatics, University of Stuttgart, Germany
uwe.soergel@ifp.uni-stuttgart.de

Keywords: Multimodal pseudo-Siamese Network, Cross-modal Attention, Representation Learning, Transfer Learning, Label-efficient HS-LiDAR-based DSM Classification

Abstract

Remote sensing acquisition systems rely on a range of platforms, from drones to satellite missions, to record multimodal Earth surface data. This fact encourages the preparation of datasets with complementary properties, thereby increasing their discriminative potential. A common complementary combination is between Hyperspectral and LiDAR-generated digital surface model data. While engaging, this fusion poses challenges for specific applications. Multiple works fuse these modalities at the feature level using vector concatenation, maximization, or averaging. Although functional, these methods omit target interactions between the modalities. Another challenge in remote sensing is the quantity and quality of labels required by deep learning methods, which are expensive, error-prone, and difficult to scale. We address the challenges above by proposing a self-supervised processing framework based on cross-modal attention that effectively fuses features at multiple levels, thereby exploiting complementary information across data streams. Specifically, our method is founded on a pseudo-Siamese network that reweights each modality's features with information from the other via a mirrored cross-modal attention. The network's objective is to maximize the similarity between the feature representations of both streams. A fusion network builds a latent representation using the learned encoders and attention modules. Then, a k-Nearest Neighbor classifier categorizes each sample within the representation using ten labels per class. Our experiments show that our spatial- and channel-spatial cross-modal attention approaches outperform well-established fusion methods for label-efficient land cover classification across datasets. Our findings lay the groundwork for fusion methods that effectively exploit inter-stream data relationships to encourage complementarity.

1. Introduction

Current Remote Sensing (RS) acquisition mainly relies on satellite, airborne, and drone campaigns to continuously monitor Earth's surface dynamics. Data acquisition campaigns, carried out at specific locations, consider an appropriate sensor configuration, revisit time, and phenological seasons. The previous procedure encourages the preparation of multimodal, multitemporal, and multiscale datasets, thereby providing the data diversity necessary to study phenomena occurring on Earth's lithosphere with the help of Remote Sensing. These datasets ideally include complementary modalities, since one modality can identify objects that the other has limitations with, and vice versa. Multimodality improves their discriminative ability compared to single-modality datasets, enabling the generation of highly detailed map products. A notable example of this complementarity is the combination of HS and LiDAR DSM data. Since the HS sensor often exhibits similar spectra for scene objects belonging to different classes, the LiDAR DSM data integrates geometrical information, enabling the separation of spectrally similar categories via their distinguishable heights.
A central question is how to effectively fuse the modalities involved, enhancing their interactions to exploit their complementarity. Researchers have recognized the potential of Deep Learning (DL) in RS for multimodal image analysis, building feature extractors to fuse features at specific network depths. Considering how to fuse, common feature-level fusion strategies include concatenation, addition, averaging, and maximization. While these techniques have shown success in specific RS applications (Feng et al., 2020, Hang et al., 2020), they fundamentally fail to reinforce target interactions between the involved modalities. This deficiency limits their ability to effectively leverage complementary information, which is crucial for enhancing the system's discernment capabilities. Therefore, developing advanced fusion techniques that explicitly promote complementarity among modalities is essential.
The costly labeling is another challenge affecting the RS image interpretation. Current literature states that 5 000 labels per category are necessary for supervised classifiers to work satisfactorily (Goodfellow et al., 2016). This requirement could be manageable within Computer Vision (CV). Still, in RS, labeling is more costly because the process requires specialists who can analyze pixels' spectra to identify objects without field inspec-

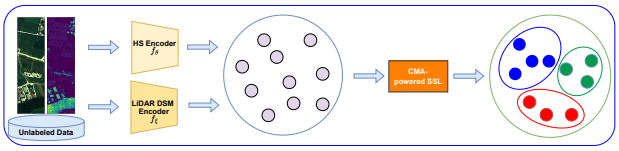


Figure 1. Our work uses unlabeled geographically paired Hyperspectral (HS)-LiDAR-generated digital surface model (LiDAR DSM) data for Cross-modal Attention (CMA) enhanced self-supervised pretraining. A fusion network employs the components learned during pretraining to project the data into a low-dimensional space. Ultimately, the k-Nearest Neighbors (k-NN) classifier assigns categories to each feature using ten labels per class.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-91-2026 | © Author(s) 2026. CC BY 4.0 License.

tion. Additionally, labeling is error-prone, tedious, and unscalable since numerous annotations are required for each newly recorded dataset.

In recent years, Self-supervised Learning (SSL) has been applied across diverse research fields, including CV, Natural Language Processing (NLP), and RS, to learn meaningful feature representations that facilitate the resolution of a specific Downstream Task (DST). Notably, it has surpassed major supervised benchmarks and made significant progress in lowering the amount of human supervision required by DL systems (Goyal et al., 2021). Considering the Land Cover Classification (LCC), SSL has enabled its accurate resolution with significantly fewer labels than Supervised Learning (SL) systems (Scheibenreif et al., 2022, González Santiago et al., 2025). Because of that, SSL represents a promising path to approach the limited labels scenario ubiquitous in RS.

Our current work addresses the challenges above by employing CMA to implement an SSL framework that jointly learns meaningful HS-LiDAR DSM features. To the best of our knowledge, our method is among the first to use CMA to effectively fuse HS-LiDAR DSM features at multiple levels in an SSL pipeline, thereby reinforcing the exploitation of complementarity. It uses fused representations with a capable learning objective that enables the acquisition of valuable features for resolving the LCC with ten labels per class. At a high level, our method comprises a pretraining phase using a pseudo-Siamese neural architecture that simultaneously processes the HS and LiDAR DSM streams as shown in Figure 1. Each data stream includes its own encoder with a designated number of convolutional blocks to extract unimodal features. It uses a mirrored CMA that finds LiDAR DSM features in the HS data stream and uses them to reweight the HS features. The same procedure applies for the HS features on the LiDAR DSM stream. After mutually reweighting the HS and LiDAR DSM features along their branches, each data stream uses an independent fully connected network to project its representations into a low-dimensional space to maximize their similarity. The pretraining learns the encoders, the CMA modules, and attention fusion networks to cluster the projected representations into distinguishable groups in feature space. The subsequent classification combines the learned encoders, CMA modules, and attention fusion networks to compute a latent representation. Then, a k-NN classifier uses it to categorize each sample after training with ten labels per class.

Our main contributions involve the implementation of mirrored CMA mechanisms to integrate complementarity between the HS and LiDAR DSM data streams, in contrast to the commonly applied fusion methods mentioned above. Additionally, our work implements an SSL framework that learns valuable representations to resolve LCC in the limited-labels regime.

2. Related Work

2.1 Remote Sensing Data Fusion

It refers to the development of techniques, for example Machine Learning (ML) or DL methods, capable of effectively combining two or more data sources to increase information content for subsequent processing and interpretation. Depending on the level within the processing chain at which fusion occurs, Remote Sensing Data Fusion (RSDF) has traditionally operated at the pixel, feature, and decision levels. As one of the most used methods, the feature-level fusion generates an augmented information set considering the data belonging to the different sources. Specifically, it integrates multisensor data by combining spectral, spatial, textural, and structural features, primarily through stacking. The features' integration strives for an enriched scene content as input to subsequent decision steps (Dalla Mura et al., 2015). The rise of DL has led to DL-based fusion, in which researchers build Deep Neural Networks (DNNs) to combine extracted features (Hong et al., 2021). More specifically, it consists of fusing features extracted at different depth levels of the neural architecture. This integration is typically performed using concatenation, weighted summation, convolution, and element-wise addition or multiplication (Samadzadegan et al., 2025). Multiple scientific contributions (Hong et al., 2021, Hang et al., 2020) have shaped this line of research. They have benefited from the feature-extraction capabilities of neural networks, marking the onset of applying DL to multimodal data fusion. Most works aggregate the extracted features via concatenation, maximization, or averaging, achieving accurate results within their application fields. However, these operations limit appropriate cross-modal interaction between modalities, thereby compromising the exploitation of their complementary properties.

2.2 Discriminative Self-supervised Learning

It refers to employing a DNN that uses unlabeled data and an appropriate learning objective to learn feature representations with high semantic meaning necessary for the subsequent DST. As SSL subcategory, Discriminative Self-supervised Learning (DSSL) uses distinguishing signals between images or groups of them to learn feature representations (Oquab et al., 2023). In CV, DSSL emerged after having successes in instance classification methods, where (Dosovitskiy et al., 2014) trained a Convolutional Neural Network (CNN) using unlabeled data and a discriminative objective, yielding features that outperform previous unsupervised methods in visual object recognition. Since then, the field has focused on building powerful feature extractors for SSL (Caron et al., 2021) and scaling up model training to learn feature representations leading to foundation models (Oquab et al., 2023, Reed et al., 2022). Ultimately, it continuously pushes the boundaries of scale to create universal vision backbone models (Siméoni et al., 2025).

Within RS, DSSL has focused on learning meaningful feature representations via CL (Stojnic and Risojevic, 2021), clustering (Liu et al., 2022), knowledge distillation (Jain et al., 2022), and also through Masked Autoencoding (MAE) (Reed et al., 2022). These developments have led to significant advances in LCC, semantic segmentation, change-, and object detection.

2.3 Label-efficient Remote Sensing Image Interpretation

Label efficiency in RS refers to advanced DL techniques that resolve a target DST with a minimal amount of manually labeled data. Its core idea is to overcome the bottleneck of creating large, extensively labeled datasets, thereby addressing the disadvantage mentioned in Section 1. The concept behind label efficiency has motivated several works within RS, including the use of a multimodal CL collaborative strategy (Jia et al., 2023), a multiview CL approach (Xue et al., 2022), and the utilization of a multimodal transformer for long-range context modeling (Zhang et al., 2023). Most of the works in this area employ CL to achieve effective representation learning. However, CL requires the explicit preparation of negative samples, which again compromises scalability.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-91-2026 | © Author(s) 2026. CC BY 4.0 License.

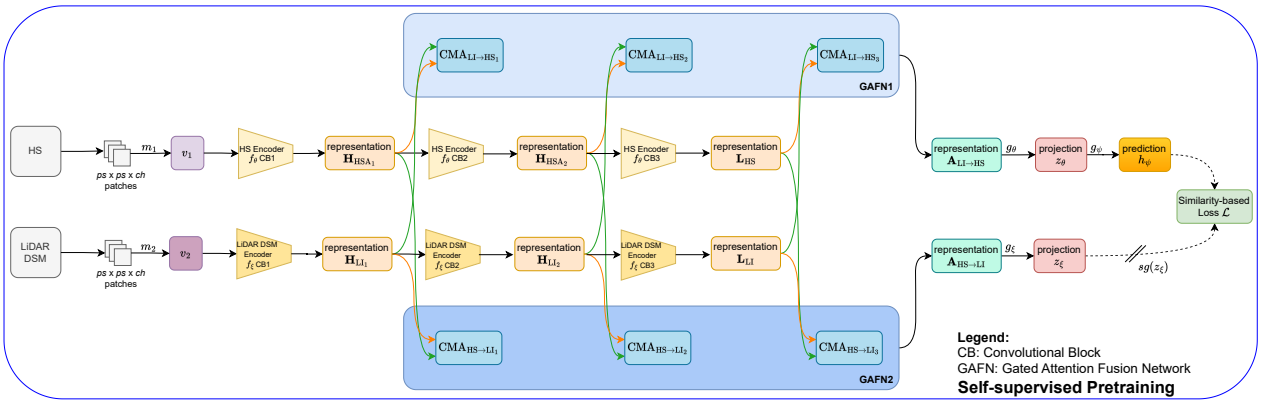


Figure 2. Overview of the proposed CMA-powered self-supervised pretraining.

3. Methodology

This section describes each main phase of our method. It outlines the data preprocessing, the CMA-powered self-supervised pretraining, and the supervised training.

3.1 Data Preprocessing

Our method takes as input the HS data cube $\mathbf{X}_{\text{HS}} \in \mathbb{R}^{\text{H}\times\text{W}\times\text{B}}$, where H × W is the spatial extent and B the number of bands. When many outliers are present, the HS data are robustly scaled by clipping values to the 0.1% and 99.9% percentiles. These percentile bounds define a range that reduces the influence of extreme values. We normalize its values using min-max normalization, scaling them to $[0, 1]$ to ensure adequate functioning of the ML pipeline. Similarly, this stage regards a normalized Digital Surface Model (DSM) derived from LiDAR data, having the form $\mathbf{X}_{\text{LI}} \in \mathbb{R}^{\text{H}\times\text{W}}$ and the same spatial dimensions as the HS cube. Given the absence of significant outliers, we only apply the min-max-normalization on the LiDAR DSM data.

3.2 CMA-powered Self-supervised Pretraining

This section describes the feature extraction, reweighting, and fusion, followed by the chosen SSL strategy.

3.2.1 Encoders' Architecture Figure 2 provides an overview of the proposed pretraining. A pseudo-Siamese network represents the foundation of our study, comprising the encoders f_θ and f_ξ for HS and LiDAR DSM data, respectively. It is a pseudo-Siamese architecture that provides flexibility by learning separate parameters for each branch. The HS encoder consists of three convolutional blocks. The first reads the number of HS bands using a three-dimensional (3D) convolution and an Exponential Linear Unit (ELU), converting the input into the intermediate representation $\mathbf{H}_{\text{HS}_1} \in \mathbb{R}^{\text{F}\times\text{C}\times\text{H}\times\text{W}}$, where F, C, H, and W are the number of feature maps, hidden channel dimensions, and the reduced height and width of the data tensor, respectively. Additionally, this first block transforms $\mathbf{H}_{\text{HS}_1}$ into a 2D representation using a learnable 3D adaptation block that summarizes relevant spectral patterns into the 2D representation denoted as $\mathbf{H}_{\text{HSA}_1}$. The second and third blocks also consist of 3D convolutions and ELU activations. Our work uses the ELU function because it resolves the dying Rectified Linear Unit (ReLU) challenge (Prince, 2023). The second convolutional block takes $\mathbf{H}_{\text{HS}_1}$ to generate a second intermediate representation $\mathbf{H}_{\text{HS}_2}$ and its 2D representation $\mathbf{H}_{\text{HSA}_2}$. Subsequently, this phase takes $\mathbf{H}_{\text{HS}_3}$, acquired after the last convolutional block, and produces its respective 2D representation $\mathbf{L}_{\text{HS}}$. The LiDAR DSM encoder also contains three convolutional blocks, each consisting of a 2D convolution, a Batch Normalization (BN), and an ELU layer. Similar to the HS encoder, the initial convolutional block processes the single-channel input, resulting in $\mathbf{H}_{\text{LI}_1} \in \mathbb{R}^{\text{F}\times\text{H}\times\text{W}}$ that is the intermediate representation used to generate $\mathbf{H}_{\text{LI}_2}$. This encoder results in the representation $\mathbf{L}_{\text{LI}}$.

3.2.2 Cross-modal Attentions Attention-driven fusion selectively weights features across modalities, dynamically focusing on the most relevant parts of the data during processing. We implement four CMA mechanisms to leverage attention, thereby strengthening interactions between the HS and LiDAR DSM streams and enforcing complementarity during pretraining. To that end, we implement the respective CMA mechanism on both branches after each encoder's convolutional block $\text{CB}_\bullet$ has computed its respective representation at each depth of the network. For the upcoming clarification, we adopt the notation $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ for queries, keys, and values, respectively, as introduced by (Vaswani et al., 2017).
The CMA-powered self-supervised pretraining starts by sampling batches of unlabeled HS and LiDAR DSM patches, creating the standard SSL views v_1 and v_2 from the sampled modalities m_1 and m_2. These views are then processed by the first CB_1 of the HS encoder f_θ and the LiDAR DSM encoder f_ξ. They result in feature representations $\mathbf{H}_{\text{HSA}_1}$ and $\mathbf{H}_{\text{LI}_1}$, respectively. Then, our $\mathbf{CMA}_{\text{LI}\to\text{HS}_1}$ considers $\mathbf{H}_{\text{LI}_1}$ as the query tensor $\mathbf{Q}$ and $\mathbf{H}_{\text{HSA}_1}$ as the key tensor $\mathbf{K}$, which also represents the value tensor $\mathbf{V}$. We apply the same procedure simultaneously on the LiDAR DSM branch, where $\mathbf{Q}$ and $\mathbf{K} = \mathbf{V}$ are $\mathbf{H}_{\text{HSA}_1}$ and $\mathbf{H}_{\text{LI}_1}$, respectively. We apply the previous feature extraction and reweighting further to both branches at the two upcoming network depth levels.

3.2.3 Gated Cross-modal Attention Inspired by (Arevalo et al., 2020), we first implement the Gated Cross-modal Attention (GCMA). We choose to implement it because its cross-modal gate determines how much of the cross-modal information should be retained. Furthermore, it facilitates adaptive cross-modal information integration by learning to control the contribution of each input stream before producing the final feature representation. The tensors $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ are contained in $\mathbb{R}^{\text{B}\times\text{C}\times\text{H}\times\text{W}}$, where B, C, H × W represent the batch size, the number of channels, and the reduced spatial dimensions, respectively. Equation (3) materializes our GCMA.

$$\mathbf{G}_{\text{cat}} = \mathbf{cat}_1([\mathbf{Q}, \mathbf{K}, \mathbf{V}]) \tag{1}$$

$$\mathbf{G}_{\text{sig}} = \mathbf{gate}(\mathbf{G}_{\text{cat}}) \tag{2}$$

$$\mathbf{CMA}_{\bullet\to\bullet} = \mathbf{conv}(\mathbf{G}_{\text{cat}} \odot \mathbf{G}_{\text{sig}}) \tag{3}$$

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-91-2026 | © Author(s) 2026. CC BY 4.0 License.

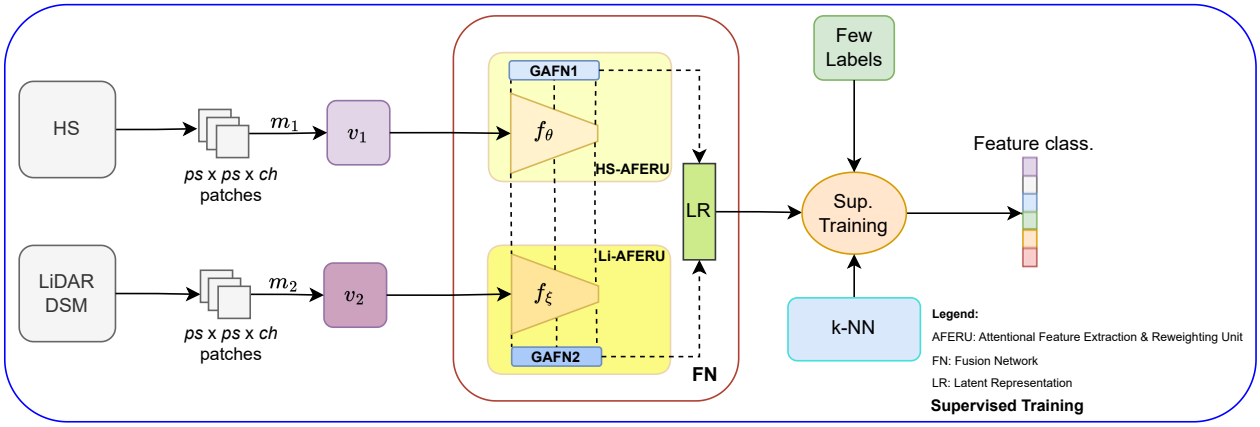


Figure 3. Overview of the supervised training. It relies on two Attentional Feature Extraction & Reweighting Units (AFERUs) wrapped in a Fusion Network (FN) that creates joint latent representations used by the k-NN classifier to categorize each test feature representation with ten labels per class.

In Equation (1), the GCMA uses the $\mathbf{cat}_1(\bullet)$ operation to concatenate the $[\mathbf{Q}, \mathbf{K}, \mathbf{V}]$ tensors along the first dimension, the channel dimension. It results in $\mathbf{G}_{\mathbf{cat}}$ contained in $\mathbb{R}^{\mathrm{B}\times 3\mathrm{C}\times \mathrm{H}\times \mathrm{W}}$. It is then processed by the $\mathbf{gate}(\bullet)$ function that is a lightweight convolutional subnetwork containing two 2D convolutions with a BN and GELU activation in-between. The subnetwork culminates in a sigmoid activation function, which produces a gating signal $\mathbf{G}_{\text{sig}}$ with the exact dimensions as $\mathbf{G}_{\text{cat}}$. The $\mathbf{G}_{\text{sig}}$ values are in the range $[0, 1]$, allowing it to act as a dynamic, input-dependent mask. Subsequently, Equation (3) shows the element-wise multiplication between $\mathbf{G}_{\text{cat}}$ and $\mathbf{G}_{\text{sig}}$. This operation selectively reweights the feature representation by amplifying salient features, where $\mathbf{G}_{\text{sig}} \approx 1$, and suppressing irrelevant ones, where $\mathbf{G}_{\text{sig}} \approx 0$. Ultimately, the result of the previous operation undergoes a final 1×1 $\mathbf{conv}(\bullet)$. It combines the reweighted information by projecting the feature map from 3C channels back to the original C channels, yielding the respective output $\mathbf{CMA}_{\bullet\to\bullet}$.

3.2.4 Channel Cross-modal Attention Influenced by the Squeeze-and-Excitation (SE) networks (Hu et al., 2018), we introduce the Channel Cross-modal Attention (CCMA), which adapts the SE concept to a cross-modal context and explicitly incorporates the interaction between the $(\mathbf{Q},\mathbf{K},\mathbf{V})$ tensors. We employ this mechanism as it learns to identify the most informative modality channels and suppress redundant ones. Moreover, the CCMA dynamically recalibrates the channel features of the target modality $\mathbf{Q}$ based on information aggregated from the source modality $\mathbf{K} = \mathbf{V}$, leading to improvements in the representation quality. The CCMA consists of three main stages: squeeze, cross-modal excitation, and recalibration. The tensors $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ are within $\mathbb{R}^{\mathrm{B}\times \mathrm{C}\times \mathrm{H}\times \mathrm{W}}$. In the squeeze step, we apply a Global Average Pooling (GAP) to each tensor independently, compressing their spatial dimension and producing vectors $(\mathbf{q}, \mathbf{k}, \mathbf{v}) \in \mathbb{R}^{\mathrm{B}\times \mathrm{C}}$. In the second stage, we use three independent Multilayer Perceptrons (MLPs) to project each vector into lower-dimensional representations $\mathbf{q}_{\text{ld}}$, $\mathbf{k}_{\text{ld}}$, and $\mathbf{v}_{\text{ld}}$. The previous step reduces model complexity and improves generalization. We then compute the attention $\mathbf{a}$ via a factorized trilinear interaction implemented with element-wise products, as depicted in Equation (4). The similarity between $\mathbf{q}_{\text{ld}}$ and $\mathbf{k}_{\text{ld}}$ is first captured through the first product and then used to reweight $\mathbf{v}_{\text{ld}}$. This interaction models complex, second-order relationships between modalities. In Equation (5), the attention vector $\mathbf{a}$ is passed through a projection MLP, a sigmoid activation, and a reshape operation. The activation normalizes the gains to $[0, 1]$. This process results in $\mathbf{G} \in \mathbb{R}^{\mathrm{B}\times \mathrm{C}\times 1\times 1}$. Then, Equation (6) computes $\mathbf{CMA}_{\bullet\to\bullet}$ via an element-wise multiplication that adaptively reweights the significance of each feature channel in the target modality.

$$\mathbf{a} = (\mathbf{q}_{\text{ld}} \odot \mathbf{k}_{\text{ld}}) \odot \mathbf{v}_{\text{ld}} \tag{4}$$

$$\mathbf{G} = \mathbf{proj}(\mathbf{a}) \tag{5}$$

$$\mathbf{CMA}_{\bullet\to\bullet} = \mathbf{Q} \odot \mathbf{G} \tag{6}$$

3.2.5 Spatial Cross-modal Attention We use the work by (Deng et al., 2022) as a reference for implementing our Spatial Cross-modal Attention (SCMA). We decide to implement the SCMA because it emphasizes the spatial dimension of the feature representations involved, thereby focusing on where to pay attention. Additionally, it learns where in the scene the $\mathbf{Q}$ and $\mathbf{K}$ features complement each other, rather than only performing channel reweighting. The SCMA takes the $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ tensors and applies three 2D convolutions as specified by Equations (7)-(9). We apply the first two to reduce the channel dimension of $\mathbf{Q}$ and $\mathbf{K}$ by two. Equation (9) employs a 2D 1×1 convolution, which acts as a learnable linear projection across channels.

$$\mathbf{Q}_p = \mathbf{q\_conv}(\mathbf{Q}) \tag{7}$$

$$\mathbf{K}_p = \mathbf{k\_conv}(\mathbf{K}) \tag{8}$$

$$\mathbf{V}_p = \mathbf{v\_conv}(\mathbf{V}) \tag{9}$$

$$\mathbf{Q}_p = \mathbf{Q}_p \cdot \mathrm{D}^s \tag{10}$$

Subsequently, the processing scales $\mathbf{Q}_p$ using Equation (10), where D is the number of input channels divided by the number of heads and s is a learnable scalar value. Then, the procedure arranges $(\mathbf{Q}_p, \mathbf{K}_p, \mathbf{V}_p) \in \mathbb{R}^{\mathrm{B}\times \mathrm{C}\times \mathrm{H}\times \mathrm{W}}$ into a data representations of the form $(\mathbf{Q}_p, \mathbf{K}_p, \mathbf{V}_p) \in \mathbb{R}^{\mathrm{E}\times \mathrm{S}\times \mathrm{D}}$, where E is the product between the batch size and the number of heads, S denotes the length of the sequence that represents the size of the flattened image patch, and D is the number of channels divided by the number of heads. The SCMA core functionality is the scaled dot-product CMA. Considering Equation (11), it takes the $(\mathbf{Q}_p, \mathbf{K}_p)$ tensors to compute their dot product, finding similar spatial features on $\mathbf{K}_p$. It then scales the result by $\sqrt{\mathrm{D}} \cdot \tau$, where τ is a learnable temperature parameter that regulates the sharpness of the CMA. Ultimately, it applies a softmax function, resulting in a multihead attention tensor $\mathbf{M} \in \mathbb{R}^{\mathrm{E}\times \mathrm{S}\times \mathrm{D}}$, where each head attends to different parts of the input. We

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-91-2026 | © Author(s) 2026. CC BY 4.0 License.

use all the heads to achieve robust representations that capture the complementary information required for the DST. By using Equation (12), the SCMA multiplies $\mathbf{M}$ and $\mathbf{V}_p$ element-wise to reweight $\mathbf{V}_p$.

$$\mathbf{M} = \mathbf{softmax}\left(\frac{\mathbf{Q}_p \odot \mathbf{K}_p^T}{\sqrt{\mathrm{D}} \cdot \tau}\right) \tag{11}$$

$$\mathbf{R} = \mathbf{M} \odot \mathbf{V}_p \tag{12}$$

$$\mathbf{CMA}_{\bullet\to\bullet} = \mathbf{norm}(\mathbf{conv}(\mathbf{rearrange}(\mathbf{R}))) \tag{13}$$

Finally, the module uses Equation (13) to arrange $\mathbf{R}$ to the original spatial dimensions $\mathbb{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$. It then sequentially applies a 1×1 convolution and a normalization to produce the final attended representation.

3.2.6 Channel-Spatial Cross-modal Attention Inspired by CBAM (Woo et al., 2018), we implement the Channel-Spatial Cross-modal Attention (CSCMA), which follows Equation (14) to sequentially apply the CCMA and the SCMA on the respective input features.

$$\mathbf{CMA}_{\bullet\to\bullet} = \mathbf{scma}(\mathbf{ccma}([\mathbf{Q}, \mathbf{K}, \mathbf{V}])) \tag{14}$$

We implement this technique to compute complementary attentions. The CCMA focuses on discovering the most relevant features along the channel dimension. In contrast, the SCMA determines which spatial locations within the source feature map should receive attention based on the network's objective.

3.2.7 Fusion of multilevel cross-modal attentions As a reminder, we apply each implemented CMA on the features obtained after the three main convolutional blocks of each encoder. This process yields three $\mathrm{CMA}_{\mathrm{LI}\to\mathrm{HS}}$ reweighted tensors on the HS branch and three $\mathrm{CMA}_{\mathrm{HS}\to\mathrm{LI}}$ on the LiDAR DSM stream. To fuse the three reweighted CMA tensors on each branch, we can simply sum or concatenate them along the channels dimension. These operations are parameter-free, straightforward to implement, and computationally efficient. However, they exhibit several limitations that are counterproductive to our research objectives. First, they assume both modalities contribute equally everywhere, treating each reweighted tensor as equally important. Moreover, they are hard operations that, by default, can increase the effect of one modality or information redundancy, thereby facilitating conflicting signals during training. Lastly, they are incapable of dynamically adapting to input and learning the fusion. To overcome the drawbacks above, we distill information from multilevel representations by implementing an adaptive, learnable Gated Attention Fusion Network (GAFN). Considering in Figure 2, for instance, $\mathrm{CMA}_{\mathrm{LI}\to\mathrm{HS}_1}$, $\mathrm{CMA}_{\mathrm{LI}\to\mathrm{HS}_2}$, and $\mathrm{CMA}_{\mathrm{LI}\to\mathrm{HS}_3}$, we first summarize their spatial dimensions using a GAP since it provides an adequate balance between salient and less informative features, giving a robust foundation for the upcoming steps. Subsequently, our GAFN takes the three reweighted pooled features and projects them using three 2D 1×1 convolutions, to handle their different input dimensions and map them into a unified projection space. They are then concatenated along the channels dimension to obtain $\mathbf{T}_{\mathrm{cat}} \in \mathbb{R}^{\mathrm{B}\times 3\mathrm{D}}$, where B and D represent the batch size and unified projection dimension, respectively. The GAFN processing also stacks each projected feature along the channels' dimension, obtaining $\mathbf{T}_{\mathrm{stk}} \in \mathbb{R}^{\mathrm{B}\times 3\times\mathrm{D}}$.Subsequently, it passes $\mathbf{T}_{\mathrm{cat}}$ to a linear subnetwork that contains two MLPs, an ELU between them, and culminates with a softmax layer that creates a gating signal denoted as $\mathbf{G}_{\mathrm{sig}} \in \mathbb{R}^{\mathrm{B}\times\mathrm{D}\times 1}$. Equation (15) shows the learnable weighted sum guiding the fusion, where N denotes the number of reweighted features, which in our case is three.

$$\mathbf{A}_{\mathrm{LI}\to\mathrm{HS}} = \sum_{n=1}^{N} \mathbf{T}_{\mathrm{stk}(:,n,:)} \odot \mathbf{G}_{\mathrm{sig}(:,n,1)} \tag{15}$$

This phase results in tensors $\mathbf{A}_{\mathrm{LI}\to\mathrm{HS}}$ and $\mathbf{A}_{\mathrm{HS}\to\mathrm{LI}}$, which represent the input for the upcoming stage.

3.3 Self-supervised learning strategy

We employ the work of (Chen and He, 2020) as the SSL strategy. Considering Figure 2, the projectors g_θ and g_ξ are two independent MLP networks that project $\mathbf{A}_{\mathrm{LI}\to\mathrm{HS}}$ and $\mathbf{A}_{\mathrm{HS}\to\mathrm{LI}}$ into low-dimensional spaces $\mathbf{z}_\theta$ and $\mathbf{z}_\xi$, respectively. The pseudo-Siamese network stops the gradient for $\mathbf{z}_\xi$ and uses a predictor MLP g_ψ on the other branch to predict representation $\mathbf{z}_\xi$. Then, the network stops the gradient for $\mathbf{z}_\theta$ and predicts its representation from $\mathbf{z}_\xi$. The previous process ensures a symmetric setting that prevents representation collapse under the influence of the stop-gradient operation (Chen and He, 2020). The model feeds $\mathbf{p}_1 = g_\psi(\mathbf{z}_\theta)$ and $\mathbf{z}_\xi$ into a negative cosine similarity function as denoted in Equation (16).

$$\mathbf{N}_{\mathrm{cs}}(\mathbf{p}_1, \mathbf{z}_\xi) = -\frac{\mathbf{p}_1}{||\mathbf{p}_1||_2} \cdot \frac{\mathbf{z}_\xi}{||\mathbf{z}_\xi||_2} \tag{16}$$

Equation (17) defines the symmetric loss function, determining what and how the network should learn.

$$\mathcal{L} = \frac{1}{2}\mathbf{N}_{\mathrm{cs}}(\mathbf{p}_1, \mathbf{z}_\xi) + \frac{1}{2}\mathbf{N}_{\mathrm{cs}}(\mathbf{p}_2, \mathbf{z}_\theta) \tag{17}$$

3.4 Fusion of learned representations

The current phase uses both learned modalities' encoders, CMA modules, and GAFNs to build a HS and a LiDAR DSM Attentional Feature Extraction & Reweighting Unit (AFERU). As depicted in the supervised training in Figure 3, we use these AFERUs to encapsulate the functionality of each encoder, its respective multilevel CMA modules, and GAFN. A superordinate FN wraps the functionality of the AFERUs, freezing each of them for the subsequent steps. The FN then stacks tensors $\mathbf{A}_{\mathrm{LI}\to\mathrm{HS}}$ and $\mathbf{A}_{\mathrm{HS}\to\mathrm{LI}}$ along the batch dimension and averages the stacked tensor. We implement the FN to aggregate local features into high-level semantic representations, thereby improving the network's ability to accurately distinguish features. We use both modalities and the FN to generate the Latent Representations (LRs) $\mathbf{R}_{\mathrm{tr}}$ for training and $\mathbf{R}_{\mathrm{ts}}$ for testing, which serve as input to the next phase.

3.5 Classification

We use k-NN for non-parametric classification. For each inquired sample, the classifier assigns the most frequent category among its nearest neighbors. We choose the k-NN classifier because it addresses the study's research questions and converges quickly, all without the need for additional hyperparameter search.

4. Experimental Setup

The current section details the characteristics of the employed datasets, describes the experimental phases of our study, and explains how we evaluate the classification results.

Table 1. Quantitative results on the Trento dataset as percentages (%).

Class	Supervised			Standard SSL			Specialized SSL			
	SVM HS	SVM HS-LiDAR DSM	Late Fusion	SimSiam AVG	SimSiam CAT	SimSiam MAX	Channel CMA	Gated CMA	Spatial CMA	Channel-Spatial CMA
Trees	89 ± 4	90 ± 4	94 ± 4	91 ± 6	92 ± 6	91 ± 7	96 ± 3	95 ± 2	96 ± 3	98 ± 1
Buildings	73 ± 13	97 ± 2	93 ± 3	98 ± 1	98 ± 1	96 ± 2	96 ± 1	97 ± 1	97 ± 3	98 ± 2
Ground	95 ± 1	94 ± 2	86 ± 3	95 ± 1	96 ± 1	95 ± 1	95 ± 2	93 ± 6	94 ± 4	97 ± 1
Woods	73 ± 4	98 ± 1	98 ± 2	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Vineyard	66 ± 4	67 ± 3	90 ± 5	99 ± 0	100 ± 0	99 ± 0	99 ± 1	99 ± 0	99 ± 1	100 ± 0
Roads	54 ± 25	85 ± 3	85 ± 6	91 ± 1	90 ± 1	91 ± 1	92 ± 3	90 ± 2	91 ± 3	90 ± 4
AA	75 ± 2	88 ± 1	91 ± 2	96 ± 1	96 ± 1	96 ± 1	96 ± 1	96 ± 1	96 ± 1	$\mathbf{97 \pm 0}$
F1 Score	71 ± 2	84 ± 1	93 ± 1	97 ± 1	97 ± 1	97 ± 1	$\mathbf{98 \pm 0}$	$\mathbf{98 \pm 0}$	$\mathbf{98 \pm 0}$	$\mathbf{98 \pm 0}$

Table 2. Quantitative results on the 2013 IEEE GRSS DF Contest dataset as percentages (%).

Class	Supervised			Standard SSL			Specialized SSL			
	SVM HS	SVM HS-LiDAR DSM	Late Fusion	SimSiam AVG	SimSiam CAT	SimSiam MAX	Channel CMA	Gated CMA	Spatial CMA	Channel-Spatial CMA
Healthy grass	84 ± 1	85 ± 5	92 ± 7	94 ± 5	94 ± 5	94 ± 5	92 ± 5	92 ± 5	82 ± 9	87 ± 7
Stressed grass	71 ± 10	74 ± 8	79 ± 10	53 ± 12	57 ± 14	58 ± 15	89 ± 6	88 ± 3	80 ± 5	79 ± 11
Synthetic grass	98 ± 1	96 ± 5	96 ± 3	100 ± 0	99 ± 1	99 ± 1	99 ± 0	99 ± 2	99 ± 1	99 ± 1
Trees	76 ± 32	89 ± 5	81 ± 7	95 ± 2	95 ± 2	95 ± 2	92 ± 3	93 ± 3	96 ± 3	96 ± 2
Soil	82 ± 1	71 ± 9	96 ± 3	97 ± 2	97 ± 3	96 ± 2	88 ± 8	91 ± 4	98 ± 2	96 ± 2
Water	84 ± 0	80 ± 4	75 ± 7	78 ± 3	79 ± 4	78 ± 3	71 ± 9	72 ± 6	84 ± 6	79 ± 7
Residential	15 ± 23	79 ± 9	49 ± 4	67 ± 9	68 ± 10	65 ± 8	88 ± 4	89 ± 1	91 ± 3	93 ± 4
Commercial	16 ± 10	59 ± 10	68 ± 16	75 ± 5	74 ± 4	74 ± 5	71 ± 5	69 ± 2	77 ± 10	79 ± 7
Road	77 ± 11	80 ± 13	71 ± 12	79 ± 3	82 ± 3	82 ± 4	74 ± 5	70 ± 2	83 ± 6	80 ± 5
Highway	3 ± 4	22 ± 12	65 ± 9	59 ± 5	58 ± 6	57 ± 6	59 ± 2	58 ± 5	75 ± 7	77 ± 9
Railway	40 ± 20	54 ± 12	74 ± 10	65 ± 5	67 ± 5	65 ± 5	72 ± 8	73 ± 7	92 ± 2	96 ± 1
Parking Lot 1	7 ± 15	6 ± 6	66 ± 13	78 ± 4	78 ± 3	78 ± 4	68 ± 10	70 ± 9	82 ± 6	90 ± 9
Parking Lot 2	6 ± 6	28 ± 5	72 ± 6	86 ± 3	84 ± 2	85 ± 2	75 ± 11	72 ± 8	94 ± 4	93 ± 3
Tennis Court	95 ± 3	96 ± 2	92 ± 4	99 ± 1	99 ± 1	99 ± 1	96 ± 3	96 ± 2	100 ± 1	100 ± 0
Running Track	97 ± 2	95 ± 5	97 ± 2	96 ± 1	96 ± 1	96 ± 1	91 ± 3	92 ± 3	93 ± 8	95 ± 6
AA	57 ± 3	68 ± 1	78 ± 3	81 ± 1	82 ± 1	81 ± 1	82 ± 2	82 ± 1	88 ± 1	89 ± 2
F1 Score	53 ± 4	66 ± 1	77 ± 3	79 ± 1	80 ± 1	79 ± 1	$\mathbf{81 \pm 1}$	$\mathbf{81 \pm 1}$	$\mathbf{87 \pm 1}$	$\mathbf{89 \pm 2}$

4.1 Datasets

Trento It depicts a rural area in southern Trento, Italy. The HS cube comprises 63 spectral bands ranging from 400 to 980 nm. The image spatial extension is 600×166 pixels, with a Ground Sampling Distance (GSD) of 1 m. The normalized LiDAR DSM has the same spatial extension and contains one channel with heights per pixel. The ground truth consists of six land cover classes distributed across 30 214 labels (Xu et al., 2017).

2013 IEEE GRSS DF Contest The second dataset focuses on the University of Houston campus in the USA and its surroundings. The HS data cube contains 144 spectral bands ranging from 380 to 1050 nm. Its spatial extension is 349×1905 pixels with a GSD of 2.5 m. Likewise, the LiDAR DSM has the same spatial size and resolution. It has been normalized via the Digital Terrain Model (DTM) computation developed by (Bulatov et al., 2014). Its ground truth contains 15 land-cover classes distributed across 15 029 labels (Debes et al., 2014).

4.2 Self-supervised Pretraining

Our pretraining considers the entire unlabeled dataset, following the established practice in prior SSL work (Chen and He, 2020, Jia et al., 2023). It samples batches of 512 data patches, where each patch measures $11 \times 11 \times$ B, where B corresponds to the HS part of the respective dataset. For the LiDAR DSM data, B is set to 1. The modality encoders f_θ and f_ξ independently process the multimodal data from the first up to the third convolutional block, where $\mathrm{C} = [64, 128, 256]$, respectively. The feature map dimensions increase in width across each block because this configuration yielded the best results during experimentation. As we specify in Sec. 3.2.1, our encoders yield a HS and LiDAR DSM feature representation per depth, creating at the end six feature tensors. In our first experiment, we compute the GCMA for each encoder depth and branch, combining the reweighted features at multiple levels using GAFN1 and GAFN2. The rest of our experiments follow the same procedure, but use respectively the CCMA, SCMA, and CSCMA, where the latter two employ four heads. During each experiment, the aggregated results of the GAFN, $\mathbf{A}_{\mathrm{LI}\to\mathrm{HS}}$ and $\mathbf{A}_{\mathrm{HS}\to\mathrm{LI}}$, are taken by g_θ and g_ξ to project them into $\mathbf{z}_\theta$ and $\mathbf{z}_\xi$, respectively. The predictor g_ψ interchangeably predicts the features of the other branch from the current branch and vice versa, yielding the tensors $\mathbf{p}_1$ and $\mathbf{p}_2$. The objective function $\mathcal{L}$ takes $\mathbf{p}_1$ and $\mathbf{p}_2$ and maximizes the similarity between each branch's representations during stochastic gradient descent-based training. Using the described learning strategy, we expect the network to build semantically meaningful clusters in feature space, enabling the use of the learned encoders, CMA modules, and GAFNs to perform classification with as few labels as needed.

4.3 Baselines

We establish a comparison framework for our methods, defining the category Supervised to which two configurations of the SVM (Fan et al., 2008) algorithm and the Late Fusion (Hong et al., 2021) belong to, as depicted in Tables 1 and 2. We compute two linear SVM classifications using the preprocessed HS and LiDAR DSM data. The first is executed on HS data only, while the second is conducted on a concatenated HS-LiDAR DSM tensor along their channel dimension. We select the SVM for comparison because it is a well-known algorithm that provides a simple reference performance floor. Additionally, we use the public codebase of the Late Fusion for its implementation, as it is a widely used DL technique based on concatenation. We employ its published hyperparameters, optimizer, and subnetwork initializations during training (Hong et al., 2021).
Subsequently, we define the umbrella term Standard SSL and use the architecture of our encoders f_θ and f_ξ to train three models through SSL under the SimSiam scheme (Chen and He, 2020). After SSL training, we freeze the learned encoders and use them to encode the bimodal data and generate their latent feature representations. We then combine these representations using Concatenation (CAT), Maximization (MAX), and Averaging (AVG) along the channels dimension, yielding the corresponding inputs for the classification. We have selected these techniques because they are practical, well-established fusion methods that deliver accurate results across various RS applications. Lastly, we define the category Specialized SSL as superordinate for our developed techniques.

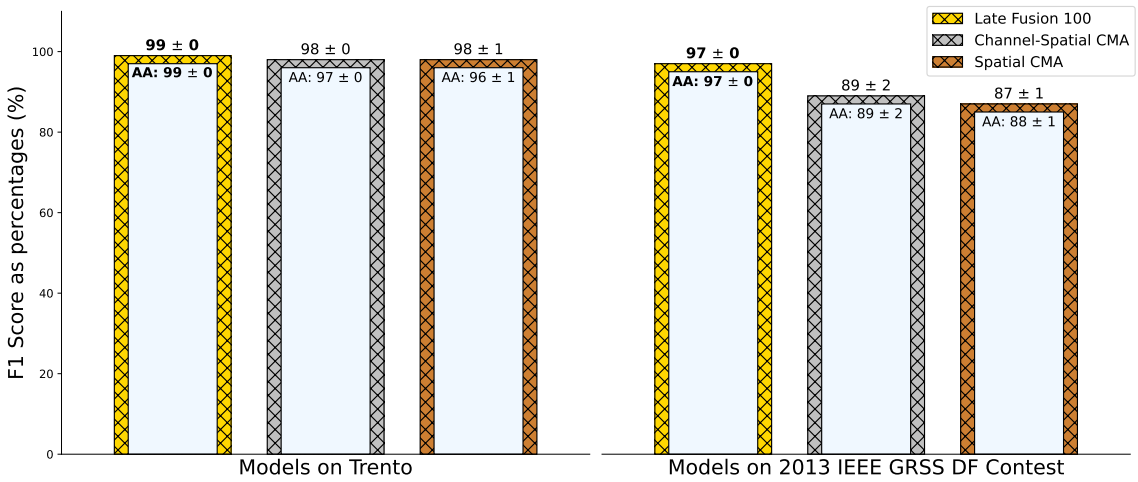


Figure 4. Upper bound comparison for the studied datasets.

4.4 Label-efficient HS-LiDAR DSM Classification

Our work is based on the premise that we have only ten labels per category for training. Because of that, we use random sampling for training and testing. During training, the classification samples ten labels per class from $\mathbf{R}_{\text{tr}}$ to fit the k-NN classifier with k=5. Subsequently, the classifier predicts on $\mathbf{R}_{\text{ts}}$, the remaining data after selecting the sample-label pairs for a given dataset. For instance, for the Trento dataset $|\mathbf{R}_{\text{tr}}| = 60$ and $|\mathbf{R}_{\text{ts}}| = 30\,354$. The same procedure is followed for the 2013 IEEE GRSS Data Fusion Contest dataset.
We also train the Late Fusion with 100 labels per class, generating the LR to be classified using the same number of labels. This setup defines the classification upper bound for each dataset, and we refer to it hereafter as Late Fusion 100.

4.5 Evaluation

We evaluate the classification performance using the F1 score (F1), Average Accuracy (AA), and per-class accuracies. We choose these metrics because they align with the objectives of our study and are widely used for assessing multilabel classification. We report their values using five distinct random seeds to ensure statistical reliability and consistency of the results.

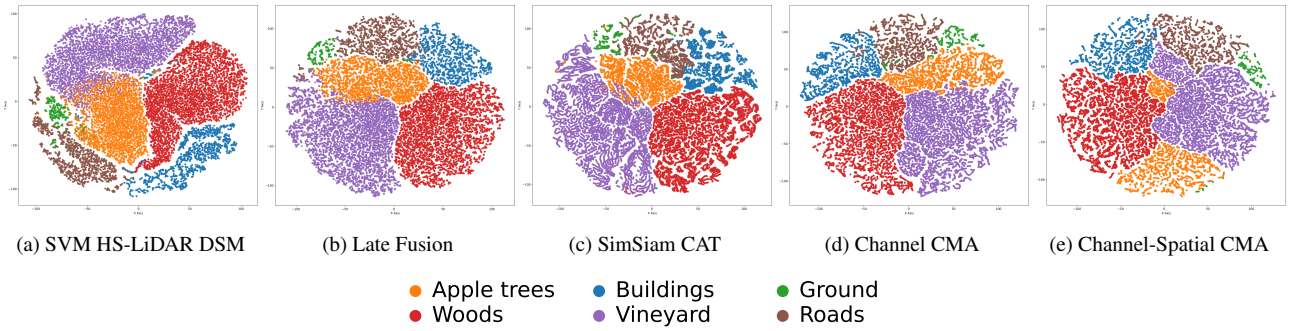
5. Results and Discussion

The Table 1 shows the quantitative results for the Trento dataset. Its F1 scores indicate that our CMA-based methods outperform all SimSiam-based feature fusion methods and the supervised methods. Specifically, each of our methods is 1 percentage point better than each of the SimSiam-based fusions and 5 more accurate than the best supervised technique. Regarding the AA values, our single techniques yield the same average and standard deviation as all the SimSiam-based fusions. They reflect accurate outcomes in which each class contributes equally. Our CSCMA yields the best results, outperforming the standard SSL techniques and the best supervised method by 1 and 6 percentage points, respectively. Our results on Trento highlight the benefits of actively deepening the interaction between the HS and LiDAR DSM branches through CMA, strengthening complementarity. The previous fact enables our current method to achieve near State-of-the-Art (SOTA) performance on the current dataset.
The Table 2 shows the quantitative results for the 2013 IEEE GRSS DF Contest dataset. Regarding its F1 scores, it evidences the superiority of our SCMA against the comparison methods. Furthermore, it also shows that the sequential use of CCMA followed by the SCMA yields the most accurate classification results, outperforming both the standard SSL and the supervised techniques. Specifically, our SCMA is 7 percentage points better than the best SimSiam-based fusion and 10 more accurate than the best of the supervised configurations. The AA value of the best SimSiam-based fusion is comparable with the AA values of the CCMA and GCMA. In contrast, our SCMA and CSCMA are 6 and 7 percentage points superior to the best SimSiam-based fusion. Our results on the current dataset again emphasize the benefits of deepening the interaction between data streams at multiple levels. Explicitly, by comparing the quality of the CCMA, the SCMA, and their influence on the CSCMA results, we find that the SCMA performs better. It relies on multi-head attention, whose heads focus on different spatial patterns of the input, and consider spatial neighborhood relationships during fusion. It learns crucial cross-modal spatial correlations, yielding more accurate classifications than CCMA, which relies on refining the importance of spectral bands. Furthermore, our best techniques considerably improve the classification accuracy for critical categories such as *Stressed grass*, *Residential*, and *Highway*. These categories pose significant challenges for the SimSiam CAT method, which struggles to reliably distinguish among them.

5.1 Comparison against the upper bound

Regarding the Trento dataset, the Late Fusion 100, representing the upper bound, outperforms our emblematic methods by 1 percentage point, as depicted on the left side of Figure 4. Considering the AA values, the upper bound method is 3 and 2 percentage points better than our SCMA and CSCMA, respectively. Considering the 2013 IEEE GRSS DF Contest dataset, the Late Fusion 100 outperforms our SCMA and CSCMA by 10 and 8 percentage points, respectively, as shown on the right side of Figure 4. Regarding the AA values, the upper bound is better than our SCMA and our best method by 9 and 8 percentage points, respectively. The superiority of the Late Fusion 100 is due to its proven neural architecture and the usage of 10 times more labels than our proposed methods. The upper bound should be treated as an ideal result unlikely to be repeatable under realistic conditions. The number of required training samples and the employed random sampling lead, on average,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-91-2026 | © Author(s) 2026. CC BY 4.0 License.



(a) SVM HS-LiDAR DSM (b) Late Fusion (c) SimSiam CAT (d) Channel CMA (e) Channel-Spatial CMA

Figure 5. Feature space classification maps for Trento with ten labels per class.

to artificially embellished results, e.g., training and test samples are direct neighbors with very little spectral-spatial variation. In the future, we will analyze the use of sophisticated sampling strategies, such as cluster sampling (Gross et al., 2019, Liang et al., 2017).

5.2 Classification maps in feature space

For simplicity, the current analysis considers the Trento dataset. The Figure 5 illustrates each method’s attempt to build semantically meaningful clusters in feature space. More specifically, each method projects $\mathbf{R}_{\text{ts}}$ into the aforementioned low-dimensional space to categorize the projected features. We begin by describing the CMA-based results. The diagrams 5d and 5e display high compactness and well-defined class boundaries, indicating strong class separability. Analysing overlap, we identify in 5d mixed features between the classes *Buildings* and *Roads*, *Ground* and *Roads*, and also between *Apple trees* and *Vineyard*. There are also some overlapping features on 5e, this time between the *Ground* and *Apple trees* classes, and again, between *Buildings* and *Roads*. Regarding the SimSiam CAT fusion results, we observe intra-class sparsity that is higher than in the CMA-based results. Furthermore, there are identifiable cluster boundaries, and visualization 5c exhibits class overlap among the classes mentioned above for 5d. Considering the Late Fusion on 5b, we observe no clearly distinguishable boundary between *Apple trees* and *Vineyard*. Furthermore, the method struggles to separate features from the *Ground*, *Buildings*, and *Roads* classes. Regarding 5a, we perceive a different geometry of the feature space because of the nature of the classification. The built clusters for the classes *Apple trees*, *Woods*, and *Vineyard* have moderate compactness. We can identify the classes’ boundaries; however, it is difficult to distinguish the boundary between *Apple trees* and *Vineyard*, and between *Ground* and *Roads*. Additionally, there is significant confusion between the *Apple trees* and *Vineyard* classes, and between *Ground* and *Roads*.

6. Conclusions

We present a technique for learning complementary feature representations in a low-dimensional space, thereby facilitating accurate, label-efficient HS-LiDAR DSM classifications. For that matter, our work introduces feature extractors that identify the most relevant features for the next steps. It implements four cross-modal attention techniques to deepen the interaction between the HS and LiDAR DSM branches, exploiting the intrinsic complementarity within the datasets. It also implements a learnable GAFN that dynamically weights the fused feature representation based on the pipeline’s objective. The SSL strategy leverages fused feature representations from both streams to maximize the similarity between their projections. The classification uses the trained encoders, CMA modules, and GAFNs to classify with few sample-label pairs. Ultimately, the evaluation employs an appropriate set of metrics to assess classification quality while addressing the challenges outlined at the beginning of this study.

Within the established experimental framework, our work shows that using the introduced SCMA and CSCMA mechanisms for cross-modal fusion enhances the SSL maximization strategy for building semantic feature clusters in a low-dimensional space. It also showcases that fusing feature representations at each level using the CMAs above enables the pseudo-Siamese architecture to learn high-level cross-modal interactions, thereby achieving accurate HS-LiDAR DSM classifications across datasets with ten labels per class.

A subsequent step in our study is to experiment with advanced feature extractors for the HS stream. We plan to use a transformer that integrates spatial and spectral information, exploiting its capacity to model contextual relationships and long-range dependencies.

References

Arevalo, J., Solorio, T., y Gómez, M. M., González, F. A., 2020. Gated multimodal networks. *Neural Computing and Applications*, 32, 10209 - 10228. https://api.semanticscholar.org/CorpusID:210196134.

Bulatov, D., Häufel, G., Meidow, J., Pohl, M., Solbrig, P., Wernerus, P., 2014. Context-based automatic reconstruction and texturing of 3D urban terrain for quick-response tasks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 157–170.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630-9640. https://api.semanticscholar.org/CorpusID:233444273.

Chen, X., He, K., 2020. Exploring Simple Siamese Representation Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15745-15753. https://api.semanticscholar.org/CorpusID:227118869.

Dalla Mura, M., Prasad, S., Pacifici, F., Gamba, P., Chanussot, J., Benediktsson, J., 2015. Challenges and Opportunities of Multimodality and Data Fusion in Remote Sensing. *Proceedings of the IEEE*, 103(9), 1585-1601.

Debes, C. et al., 2014. Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6), 2405-2418.

Deng, S., Liang, Z., Sun, L., Jia, K., 2022. Vista: Boosting 3d object detection via dual cross-view spatial attention. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8448–8457.

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M. A., Brox, T., 2014. Discriminative unsupervised feature learning with exemplar convolutional neural networks.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. LIBLINEAR: A library for large linear classification. *the Journal of machine Learning research*, 9, 1871–1874.

Feng, D. et al., 2020. Deep multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Transactions on Intelligent Transportation Systems*.

González Santiago, J., Gross, W., Middelmann, W., 2025. Assessment of self-supervised learning techniques for few-shot classification of joint hyperspectral and lidar dsm data. *Dreiländertagung D-A-CH 2025 Räumbezogene Bilddaten und Künstliche Intelligenz für nachhaltige Lebensräume¨*, Geschäftsstelle der DGPF;, Stuttgart, 363–373.

Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press.

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A. et al., 2021. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*.

Gross, W., Tuia, D., Soergel, U., Middelmann, W., 2019. Nonlinear feature normalization for hyperspectral domain adaptation and mitigation of nonlinear effects. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8), 5975–5990.

Hang, R., Li, Z., Ghamisi, P., Hong, D., Xia, G., Liu, Q., 2020. Classification of Hyperspectral and LiDAR Data Using Coupled CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 58, 4939-4950.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2021. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Trans. Geosci. Remote Sens.*, 59(5), 4340–4354.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jain, P., Schoen-Phelan, B., Ross, R. J., 2022. Self-Supervised Learning for Invariant Representations From Multi-Spectral and SAR Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 7797-7808. https://api.semanticscholar.org/CorpusID:248512508.

Jia, S., Zhou, X., Jiang, S., He, R., 2023. Collaborative Contrastive Learning for Hyperspectral and LiDAR Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-14. https://api.semanticscholar.org/CorpusID:257898470.

Liang, J., Zhou, J., Qian, Y., Wen, L., Bai, X., Gao, Y., 2017. On the Sampling Strategy for Evaluation of Spectral-Spatial Methods in Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 862-880.

Liu, F., Qian, X., Jiao, L., Zhang, X., Li, L., Cui, Y., 2022. Contrastive learning-based dual dynamic GCN for SAR image scene classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1), 390–404.

Oquab, M. et al., 2023. DINOv2: Learning Robust Visual Features without Supervision. *ArXiv*, abs/2304.07193. https://api.semanticscholar.org/CorpusID:258170077.

Prince, S. J., 2023. *Understanding Deep Learning*. The MIT Press, chapter Chapter 3 - Shallow neural networks, 37–38.

Reed, C. et al., 2022. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4065-4076. https://api.semanticscholar.org/CorpusID:255340927.

Samadzadegan, F., Toosi, A., Farzaneh, D., 2025. A critical review on multi-sensor and multi-platform remote sensing data fusion approaches: current status and prospects. *International Journal of Remote Sensing*, 46(3), 1327–1402. https://doi.org/10.1080/01431161.2024.2429784.

Scheibenreif, L., Mommert, M., Borth, D., 2022. Contrastive self-supervised data fusion for satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, 705–711.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M. et al., 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Stojnic, V., Risojevic, V., 2021. Self-Supervised Learning of Remote Sensing Scene Representations Using Contrastive Multiview Coding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1182-1191. https://api.semanticscholar.org/CorpusID:233240819.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. Cbam: Convolutional block attention module. *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xu, X., Li, W., Ran, Q., Du, Q., Gao, L., Zhang, B., 2017. Multisource remote sensing data classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2), 937–949.

Xue, Z. et al., 2022. Self-Supervised Feature Representation and Few-Shot Land Cover Classification of Multimodal Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-18. https://api.semanticscholar.org/CorpusID:253354759.

Zhang, Y., Xu, S., Hong, D., Gao, H., Zhang, C., Bi, M., Li, C., 2023. Multimodal Transformer Network for Hyperspectral and LiDAR Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-17.