The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B3-2026
XXV ISPRS Congress 2026 "From Imagery to Understanding", Commission III, 4–11 July 2026, Toronto, Canada

Fusion of PlanetScope SuperDove and Orthorectified Aerial Images for Tree-Level Stress Monitoring in Boreal Forests

Xiaoyang Zhao[1,*], Langning Huo[1,*]

[1] Department of Forest Resource Management, Swedish University of Agricultural Sciences, 90183 Umea, Sweden.

Keywords: Image Fusion, Implicit Neural Representation, Spectral Preservation, Spatial Feature Extraction, Heterogeneous Remote Sensing Image.

Abstract

Detecting early-stage vegetation stress at the individual tree scale is a pivotal remote sensing application. The "green shoulder" band at 530 nm serves as a key signal for early stress detection due to its sensitivity to carotenoid changes. However, existing remote sensing systems often struggle to simultaneously capture fine-scale canopy structures and stress-sensitive spectral data, making heterogeneous fusion a promising topic. Unlike mainstream supervised methods that rely on prescribed degradation models and high-quality samples, an unsupervised blind fusion framework based on Implicit Neural Representation and low-rank decomposition is proposed in this paper. Guided by orthorectified aerial images, the framework performs per-band super-resolution on PlanetScope SuperDove data to achieve a 0.16-meter resolution. It employs Sinusoidal Representation Networks to learn a continuous joint implicit representation of spatio-spectral information, effectively modeling the non-linear relationship between canopy structure and spectral response.To mitigate high-dimensional feature redundancy during heterogeneous data fusion, low-rank decomposition is integrated to reduce computation overhead. Experimental results show that the proposed method can fuse heterogeneous images effectively, providing a solid solution with practical guidance for subsequent early stress monitoring at the individual tree level.

1. Introduction

Frequent forest pest and disease outbreaks have caused large-scale canopy degradation and tree mortality across Sweden, making early-stage vegetation stress monitoring at the individual tree scale a focal point of remote sensing research. Studies indicate that stressed vegetation exhibits significant spectral variations in the "green shoulder" band at 530 nm, specifically characterized by enhanced carotenoid absorption (Huo et al., 2025). These signals often emerge weeks or even months before visible canopy decline, providing as a vital early warning sign for stress monitoring. However, the inherent trade-off exists between spatial and spectral resolution in current sensors hinders the simultaneous acquisition of high-resolution spectra and detailed canopy structure (Vu et al., 2025). Specifically, although PlanetScope SuperDove possesses the capability to capture the sensitive bands, its 3-meter spatial resolution triggers mixed pixel effect and spectral aliasing across different ground components. In contrast, 0.16-meter orthorectified aerial images (orthophotos) clearly delineate individual tree structures but lack the necessary spectra required for stress detection (Huo et al., 2024). Consequently, the spatial or spectral limitations of single-source imagery severely restrict the feasibility of early forest stress monitoring at the individual tree level (Algiriyage et al., 2022).

Remote sensing image fusion aims to integrate complementary information from multi-source observations to generate data characterized by both high spatial resolution and spectral fidelity. Currently, fully supervised deep learning is the dominant paradigm. These methods typically employ known or pre-defined point spread functions (PSF) and spectral response functions (SRF) to model the observation process. By simulating degradation, they construct training pairs from high-resolution references and degraded observations to learn the mapping toward high-fidelity representations. Representative frameworks include those based on convolutional neural networks (CNNs), Transformers, and Generative Adversarial Networks (GANs). For instance, Shao et al. (Shao and Cai, 2018) develop a dual-branch CNN fusion framework that separately extracts spatial and spectral features from panchromatic (PAN) and multispectral images (MSIs). By incorporating residual learning, this approach effectively correlates multi-resolution information to produce high-fidelity fused imagery. To balance performance and efficiency, Wang et al. (Wang et al., 2023) create a lightweight attention-based fusion network. It utilizes partial shallow residual blocks, dynamic feature distillation mechanism, and a multi-layer attention fusion strategy to enhance super-resolution reconstruction while minimizing computational complexity. Song et al. (Song et al., 2022) propose a multi-layer feature fusion GAN, which combines a U-Net architecture with attention mechanisms. It simultaneously achieves global distribution modeling and local change characterization.

While fully supervised methods perform well on simulated datasets, their efficacy depends on the availability of paired training samples. In practice, obtaining high-resolution ground-truth data that strictly aligns with low-resolution observations is often costly or even infeasible, particularly in large-scale, multitemporal, or cross-sensor scenarios (Lian et al., 2025). This reliance on supervised labels limits their generalization in real-world applications. To address this, unsupervised fusion methods have gained considerable attention and achieved certain progress. Unsupervised methods exploit inher-

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-1443-2026 |

ent statistical characteristics of the data and incorporate observation consistency constraints, introducing self-supervised signals by predicting of low-resolution observations, thereby learning fusion mappings directly from the data itself (Zhao et al., 2025). Yu et al. (Yu et al., 2024) take advantage of spectral-spatial collaborative constraints and group convolution enhancement module to mitigate radiometric differences, while employing a 3D attention dynamic convolution kernel to boost feature extraction. Similarly, Wu et al. (Wu et al., 2025) develops a spatiotemporal fusion network, which constructs a coarse-to-fine dual-mode complementary mechanism. By combining iterative optimization with prior constraints, this approach effectively models spatiotemporal variations to achieve high-quality reconstruction.

Despite the advantages given above, existing unsupervised methods face several critical limitations. Most rely on a mandatory consistency assumption, presupposing that the degradation process can be modeled through simple parametric functions. However, such assumptions are often difficult to satisfy in real-world heterogeneous multi-source remote sensing data fusion tasks, leading to spectral distortions when processing heterogeneous data scenarios. Therefore, there remains a pressing need to enhance fusion accuracy and generalization when processing multi-source data characterized by significant radiometric and geometric discrepancies.

Since explicitly modeling the degradation process in real-world observations is challenging, blind fusion has emerged as a promising direction for addressing heterogeneous problems. These methods relax rigid assumptions on PSF and SRF by treating degradation as an unknown latent variable. In the absence of prior degradation models, they simultaneously estimate the observation process and the high-resolution target within a unified optimization framework. This approach enables adaptive modeling of actual observations without relying on sensor-specific priors. Li et al. (Li et al., 2023) propose an unrolled network incorporating coordinate encoding to accurately estimate degradation parameters. Similarly, Cao et al. (Cao et al., 2024) integrate Transformer and CNNs with a multi-layer cross-feature attention mechanism to facilitate hyperspectral-multispectral fusion. Furthermore, Bacca et al. (Bacca et al., 2025) introduce a middle output deep image prior method, which reconstructs fused images at intermediate network layers while jointly learning a nonlinear degradation model.

Given the highly ill-posed nature of the fusion problem, current blind fusion methods typically incorporate learnable degradation modules or additional prior constraints to narrow the solution space. Nevertheless, managing the degrees of freedom within this space remains difficult, often resulting in spectral drift in the fused results. Balancing the introduction of effective structural constraints with the maintenance of model flexibility is a central challenge in blind fusion research. Currently, unsupervised blind fusion for heterogeneous remote sensing images is still in its early stages of development.

Recently, implicit neural representation (INR) has made significant progress in computer vision (Chen et al., 2021). By employing continuous functions to describe signal distributions, INR transforms discrete observations into mappings of a latent continuous field, enabling the capture of complex structures and high-frequency details. Unlike traditional networks, INR's representation is not constrained by input sampling resolution and supports querying at arbitrary coordinates. In remote sensing data processing, INR can unify multi-source, multi-modal, and spatiotemporally inconsistent data into a continuous field, treating heterogeneous observations as samples under varying conditions (Zhu et al., 2025). This characteristic offers a promising solution for blind fusion, achieving high-precision results without prior knowledge of sensor parameters (Xu et al., 2025). Furthermore, applying INR to unsupervised blind fusion still faces several challenges. First, extreme resolution disparities demand sophisticated extraction and learning of high-frequency information, without explicit priors, any loss or misestimation may lead to blurring or structural distortion. Second, the continuous spatial sampling of INR involves a large parameter training space, making the optimization prone to local optima or overfitting. Third, if the network fails to decouple spatial structure from spectral information, the results become susceptible to spectral distortion and color shifts, posing a significant threat to spectral fidelity.

To address these challenges, an unsupervised blind fusion framework for heterogeneous remote sensing images is proposed. By formulating the fusion task as an inverse problem in the continuous domain, the framework leverages a coordinate-driven INR to learn the underlying spatial and spectral distributions, thereby circumventing the need for explicit PSF and SRF modeling. To further enhance the representations of key features, a low-rank decomposition mechanism is integrated into the framework. The primary contributions of this work are threefold:

(1) A novel INR framework integrated with low-rank decomposition is proposed to advance implicit continuous representation learning. By performing dimension-wise decoupling and recombination of spatial coordinate encodings and spectral feature modulation, the framework captures the complex correlations within high-dimensional spatial-spectral tensors with high parameter efficiency. This approach overcomes the spectral representation bottlenecks inherent in traditional discrete learning mechanisms, enabling high-fidelity feature reconstruction even at ultra-high scaling factors.

(2) An edge-enhanced spatial-spectral attention encoder is designed to extract and fuse heterogeneous features. This encoder incorporates a Sobel-based edge detection module to provide explicit spatial structure priors. By integrating an attention mechanism with simplified residual learning, the encoder dynamically synergizes high-frequency details from orthophotos with spectral information from SuperDove imagery. This strategy effectively alleviates semantic inconsistencies between spatial structures and spectral attributes, ensuring coherent feature integration during the fusion of heterogeneous data.

(3) A multi-dimensional collaborative optimization strategy is developed to jointy constrain the fusion process. By comprehensively integrating spatial fidelity, spectral consistency terms with a regularization term,

this strategy stabilizes the optimization process and effectively narrows the solution space of the ill-posed inverse problem. It achieves spatial precision while maximizing spectral fidelity, ensuring robust performance in unsupervised scenarios.

The remainder of this paper is organized as follows: Section 2 introduces the architecture of the proposed framework and provides technical details for each component. Section 3 presents the experimental results, including a description of the study areas and data sources, followed by a comprehensive evaluation and visualization of the fusion performance. Finally, Section 4 concludes the paper and discusses potential future research directions.

2. Methodology

This chapter elaborates on the overall architecture of the proposed framework and its end-to-end pipeline for the collaborative processing of multi-source information. First, the edge-enhanced spatial-spectral attention encoder is introduced, detailing how it leverages edge priors and residual learning mechanisms to extract robust high-frequency details. Next, the core contribution — the INR module coupled with low-rank decomposition - is highlighted, with a focus on its ability to achieve high-fidelity modeling with minimal parameters. Finally, the loss function is formulated to demonstrate how the model converges toward the optimal solution under unsupervised conditions.

2.1 Overall Architecture

The proposed network follows an “encoding–fusion–implicit representation” design paradigm, with the overall architecture illustrated in Figure 1. It primarily consists of the following core components:

(1) Edge-enhanced spatial-spectral attention encoder

Firstly, a dual-branch feature encoder is employed to extract complementary information from heterogeneous data. It consists of a Sobel-enhanced spatial feature extractor and a lightweight spectral feature extractor, designed to capture salient structural details from high-resolution orthophotos and rich spectral attributes from SuperDove imagery, respectively. Specifically, the spatial extractor integrates a parameter-free Sobel branch, which explicitly bolsters boundary definitions and structural fidelity without expanding the model’s training scale.

Next, an attention-guided spatial-spectral fusion module is designed to achieve adaptive integration of heterogeneous features. By leveraging an attention mechanism, the module effectively filters out redundant information while accentuating salient spatial structures and spectral details. The resulting fused feature then serves as a global context, providing essential guidance for the subsequent implicit query module.

(2) Low-rank Based Implicit Continuous Fusion

This module serves as the core point, formulating the high-magnification fusion task as the learning of implicit neural functions within a continuous coordinate system. It leverages a SIREN to encode sub-pixel offsets, accurately capturing high-frequency spatial textures through its inherent periodic activation. Inspired by low-rank decomposition, the module synergizes spatial coordinate encoding with spectral feature modulation via an element-wise Hadamard product, achieving an efficient, low-rank coupling of cross-modal information with minimal parameterization. Furthermore, the framework incorporates feature unfolding to encapsulate local context and employs a local ensemble strategy for area-weighted prediction, ensuring smooth and spatially coherent reconstructions across neighboring feature points.

2.2 Edge-enhanced spatial-spectral attention encoder

To mitigate the computational burden and overfitting risks associated with excessively deep architectures, the network leverages two feature encoders based on lightweight residual blocks. These blocks, comprising standard convolutional layers, ReLU activations, and skip connections, are designed to minimize parameter count while ensuring efficient gradient propagation. Furthermore, to incorporate fine-grained details from orthophotos as a structural prior, a fixed Sobel filtering layer is integrated. This non-trainable component explicitly reinforces edge-aware learning, providing the network with robust geometric constraints from the outset.

For the input orthophotos $X \in R^{B\times C\times H\times W}$, a convolution operation is first performed using predefined kernels to calculate the horizontal gradient G_x and the vertical gradient G_y.

$$G_x = \sum_{c=1}^{C} K_x * X_c, \quad G_y = \sum_{c=1}^{C} K_y * X_c \tag{1}$$

where B represents the batch size, K_x and K_y represent the Sobel operators in the horizontal and vertical directions, respectively, and $*$ denotes the convolution operation. X_c refers to the c-th channel of X. The corresponding Sobel kernels K_x and K_y are defined as follows.

$$\mathbf{K}_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad \mathbf{K}_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{2}$$

Subsequently, the gradient magnitude $\mathbf{M}$ is calculated.

$$\mathbf{M} = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2 + \epsilon} \tag{3}$$

where $\epsilon = 10^{-6}$ is a smoothing term designed to prevent numerical instability. Finally, to eliminate the influence of magnitude variations between different samples, Min-Max normalization is applied.

$$\mathbf{E} = \frac{\mathbf{M} - \min(\mathbf{M})}{\max(\mathbf{M}) - \min(\mathbf{M}) + \epsilon} \tag{4}$$

Next, the HSI encoder focuses on modeling spectral correlations and performs progressive upsampling until it is spatially aligned with the orthophoto features. To integrate the resulting spatial and spectral deep features,

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-1443-2026 |

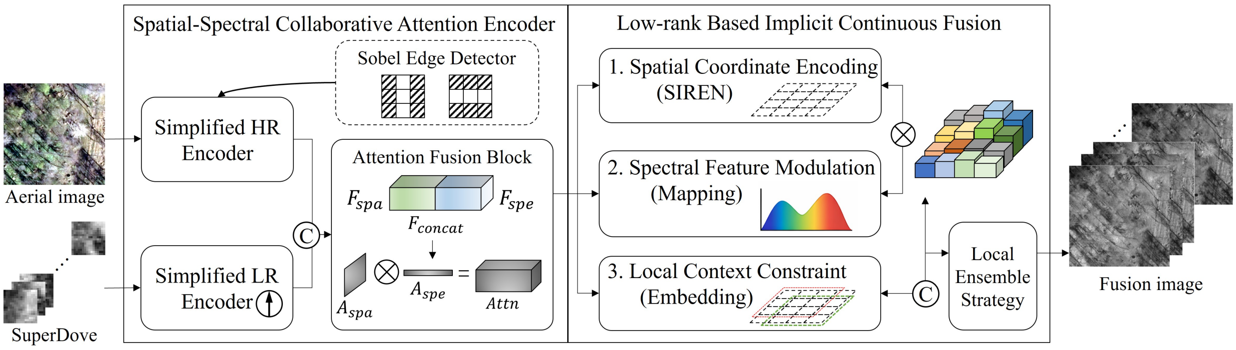

Figure 1. Overall architecture.

a lightweight attention fusion module is designed, enabling the network to adaptively emphasize spatial locations and spectral features that contribute more significantly to the reconstruction task. The spatial attention map A_{spa} is formulated as follows:

$$F_{concat} = \text{Concat}(F_{spa}, F_{spe}) \tag{5}$$

$$\mathbf{A}_{spa} = \sigma(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{1\times1}(F_{concat})))) \tag{6}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, and F_{concat} represents the concatenated spatial and spectral deep features. Similarly, the spectral attention map $\mathbf{A}_{spe}$ is derived. The final fused features are obtained through element-wise multiplication for attention modulation.

$$F_{fuse} = F_{concat} \odot A_{spa} \odot A_{spe} \tag{7}$$

2.3 Low-rank Based Implicit Continuous Fusion

The fusion task is formulated as a continuous mapping problem over a 2D coordinate domain $\Omega = [-1, 1]^2 \subset \mathbb{R}^2$, aimed at learning a continuous function f_θ. For any normalized spatial coordinate $z_q \in \Omega$ on the target image, the model predicts the corresponding continuous spectral vector $\hat{I}(z_q)$ by querying the feature maps in conjunction with coordinate encoding:

$$\hat{I}(z_q) = f_\theta(F_{spe}, F_{spa}, z_q) \tag{8}$$

To bypass the heavy computational burden of complex non-linear mappings between high-dimensional features, the decoding process is decomposed into a low-rank coupling mechanism. This approach involves spatial basis encoding synergized with spectral coefficient modulation. By factorizing the high-dimensional latent space, the model effectively reduces redundant degrees of freedom, enabling a more efficient and stable reconstruction of the spatial-spectral continuous field. The single implicit query is formulated as:

$$\hat{I}_i(z_q) = \mathcal{P}\left(\underbrace{\Phi(z_q - z_i^*; \theta) \odot \Psi(F^*_{i,spe}; \phi)}_{\text{Low-rank Coupled Feature}} \,\|\, \underbrace{\Gamma(F^*_{i,fused}; \psi)}_{\text{Local Context}} \right) \tag{9}$$

where $\hat{I}_i(z_q)$ represents the local spectral intensity predicted based on the i-th neighboring anchor z_i^*. The relative displacement $z_q - z_i^*$ provides sub-pixel geometric guidance, while $F^*_{i,spe}$ and $F^*_{i,fused}$ are feature vectors sampled at the grid center z_i^* from the spectral and fused feature maps, respectively. The output is integrated via feature concatenation $\|$ and a final linear projection $\mathcal{P}$.

Spatial Coordinate Encoding Branch: $\Phi(z_q - z_i^*; \theta)$: This branch employs a SIREN to generate the spatial basis tensor. By leveraging the inherent sensitivity of periodic sine functions to infinitesimal coordinate variations, this branch constructs the spatial backbone of the implicit field. It defines the analytical precision of the mapping from coordinate space to complex geographic structures, establishing the representational foundation for capturing sub-pixel, high-frequency textures.

Spectral Feature Modulation Branch: $\Psi(F^*_{i,spe}; \phi)$: This branch projects the localized spectral feature $F^*_{i,spe}$ to derive a spectral coefficient tensor. Acting as modulation weights, these coefficients dictate the activation intensity of different spectral bands across the spatial basis. Through an element-wise Hadamard product $\odot$ with the spatial basis, the model simulates complex spatio-spectral nonlinear interactions while maintaining an extremely low-rank parameterization.

Local Context Constraint Branch: $\Gamma(F^*_{i,fused}; \psi)$: This branch captures textural trends, local gradients, and semantic priors within the neighborhood of the query point through feature unfolding, yielding the local context tensor. By incorporating these spatial dependencies, the branch enforces semantic coherence among reconstructed pixels, seamlessly “aligning” generated values with their local neighborhoods. This mechanism effectively mitigates isolated artifacts and spectral discontinuities typically inherent in coordinate-based implicit functions, ensuring a spatially continuous and structurally consistent output.

To eliminate boundary discontinuities arising from the discrete sampling of feature maps, the model incorporates an ensemble strategy inspired by the Local Implicit Image Function (LIIF). For each query coordinate z_q, the model identifies its four nearest feature centers $z_i^*(i = 1, \ldots, 4)$ in the latent space. Independent implicit predictions $\hat{I}_i(z_q)$ are then performed for each of these neighboring anchors. To ensure spatial smoothness, their contributions are modulated by weights w_i, derived from the relative conformal area between z_q and each respective center. The final reconstructed value $I(z_q)$ is synthesized through an area-weighted aggregation process:

$$I(z_q) = \sum_{i=1}^{4} w_i \cdot \hat{I}_i(z_q) \tag{10}$$

where the ensemble weights $w_i = \frac{S_i}{S_{total}}$ are calculated based on the rectangular area formed by z_q and the diagonal anchor z_i^*. By leveraging this area-based weighting scheme, the model ensures strict mathematical continuity and seamless transitions across grid boundaries, effectively suppressing the artifacts typical of coordinate-based implicit functions.

2.4 Loss Function

The proposed multi-dimensional collaborative optimization strategy integrates spatial fidelity, spectral consistency, and regularization terms into a unified optimization framework. This joint optimization achieves an effective balance between the preservation of spatial structural details and the maintenance of spectral accuracy.

(1) Spatial Fidelity Term

To ensure that the fusion results maintain spatial consistency with the high-resolution orthophotos, a spatial gradient loss $\mathcal{L}_{grad}$ is introduced. Gradient information effectively characterizes the spatial structural details and topographic edge features. By constraining the differences in both horizontal and vertical directions, the model is forced to capture sharp transitions and textures. The spatial gray-scale representations for the fused output $\hat{I}$ and the reference orthophoto X are first derived:

$$\hat{I}_{gray} = \frac{1}{C_I} \sum_{c=1}^{C_I} \hat{I}^c, \quad X_{gray} = \frac{1}{C_X} \sum_{c=1}^{C_X} X^c \tag{11}$$

where C_I and C_X denote the number of spectral channels for the predicted image and the orthophoto, respectively. The spatial gradient loss is then defined as:

$$\mathcal{L}_{grad} = \frac{1}{N} \sum_{b=1}^{B} \left(\sum_{i=1}^{H} \sum_{j=1}^{W-1} \left| \nabla_x \hat{I}_{gray}^{(b,i,j)} - \nabla_x X_{gray}^{(b,i,j)} \right| + \sum_{i=1}^{H-1} \sum_{j=1}^{W} \left| \nabla_y \hat{I}_{gray}^{(b,i,j)} - \nabla_y X_{gray}^{(b,i,j)} \right| \right) \tag{12}$$

where B represents the batch size, H and W are the spatial dimensions, and N serves as the normalization factor. ∇_x and ∇_y denote the horizontal and vertical gradient operators, respectively.

(2) Spectral Consistency Term

This term is designed to ensure that the spectral signatures of the fusion results remain consistent with the original SuperDove imagery. Since the fused output $\hat{I}$ has a higher spatial resolution than the SuperDove data Y, $\hat{I}$ is first processed via average pooling to match the spatial dimensions of Y. The downsampled result is denoted as $\hat{I}_{down}$. To quantify the spectral distortion, the Spectral Angle Mapper (SAM) loss is employed:

$$\mathcal{L}_{SAM} = \frac{1}{N_{down}} \sum_{b,i,j=1}^{B,h,w} arccos \left(\frac{\langle \hat{I}_{down}^{(b,i,j)}, Y^{(b,i,j)} \rangle}{\|\hat{I}_{down}^{(b,i,j)}\|_2 \cdot \|Y^{(b,i,j)}\|_2 + \epsilon} \right) \tag{13}$$

where h and w represent the height and width of the low-resolution SuperDove image Y, respectively, and $N_{down} = B \times h \times w$ is the total number of pixels at the low-resolution scale. ϵ is a small constant introduced to ensure numerical stability. By minimizing the angular coordinate between the predicted and reference spectral vectors, this loss term constrains the model to preserve the intrinsic physical attributes of the land covers.

While the SAM loss constrains the orientation of spectral vectors, it remains invariant to their magnitudes. To prevent global radiometric bias and ensure intensity consistency, a band-mean loss $\mathcal{L}_{mean}$ is introduced to explicitly constrain the spectral magnitude of each band. This term aligns the mean intensity of the fused image with that of the SuperDove reference:

$$\mathcal{L}_{mean} = \frac{1}{B \times C} \sum_{b,c=1}^{B,C} \left| \frac{1}{h \times w} \sum_{i,j=1}^{h,w} \left(\hat{I}_{down}^{(b,c,i,j)} - Y^{(b,c,i,j)} \right) \right|_1 \tag{14}$$

where C denotes the number of spectral bands. By minimizing the absolute difference between the spatial averages of the predicted and reference bands, this term ensures that the reconstructed continuous field maintains the physical radiometric scale of the original multi-spectral data.

(3) Regularization Term

Finally, a Total Variation (TV) loss $\mathcal{L}_{TV}$ is introduced as a regularization term to suppress potential noise and promote piecewise smoothness in the reconstructed continuous field. By penalizing abrupt intensity changes between adjacent pixels, $\mathcal{L}_{TV}$ effectively mitigates high-frequency artifacts often inherent in coordinate-based neural representations. The TV loss is formulated as:

$$\mathcal{L}_{TV} = \frac{1}{N} \sum_{b,c=1}^{B,C} \Big(\sum_{i,j=1}^{H,W-1} \left| \hat{I}^{(b,c,i,j+1)} - \hat{I}^{(b,c,i,j)} \right| + \sum_{i,j=1}^{H-1,W} \left| \hat{I}^{(b,c,i+1,j)} - \hat{I}^{(b,c,i,j)} \right| \Big) \tag{15}$$

where $N = B \times C \times H \times W$ denotes the total number of elements in the high-resolution prediction. This constraint encourages the model to generate spatially coherent structures while preserving the sharp edges enforced by the spatial fidelity term.

The overall optimization objective of the proposed network is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{grad} + \mathcal{L}_{SAM} + \mathcal{L}_{mean} + \lambda \mathcal{L}_{TV}, \tag{16}$$

where λ is a very small constant.

3. Experiments

In this section, the performance of the proposed fusion method is evaluated through experiments conducted on real-world remote sensing datasets. The remainder of this section is organized as follows: First, the construction of the heterogeneous dataset and the corresponding experimental configurations are described. Subsequently, the fusion results are presented, followed by an in-depth analysis of the spectral preservation and spatial enhancement capabilities of the proposed method, particularly within complex forest environments.

3.1 Study Area and Dataset Construction

A typical forest-covered area in Sweden is selected as the research site for the experiments. Characterized by dense vegetation coverage and complex surface structures, this region poses rigorous challenges to the spatial reconstruction capabilities of fusion algorithms. The experimental dataset comprises PlanetScope SuperDove satellite imagery and high-resolution orthophotos. Specifically, the SuperDove imagery serves as the low-resolution source, containing eight multispectral bands with a spatial resolution of 3 m, providing rich spectral information including the characteristic "green-shoulder" band (530 nm). In contrast, the orthophotos act as the high-resolution reference, featuring four spectral bands with a resolution of 0.16 m to provide superior spatial structural representation.

In the data preprocessing stage, a series of essential procedures, including coordinate re-projection and normalization, are performed on both image types. To construct the sample dataset, a non-overlapping sliding window technique was employed to crop the SuperDove imagery into patches of $16 \times 16 \times 8$, with the corresponding orthophotos cropped into patches of $300 \times 300 \times 4$. Ultimately, 500 pairs of valid samples were randomly selected and partitioned into a training set (350 pairs), a validation set (100 pairs), and a test set (50 pairs) at a ratio of 7:2:1, providing standardized data support for the training and evaluation of the deep learning models.

3.2 Experimental Setup

The proposed model is implemented using the PyTorch framework with python 3.10. All experiments are conducted on a workstation equipped with an NVIDIA RTX 6000 Ada GPU. During the training phase, the network is optimized using the Adam optimizer with an initial learning rate 1×10^{-4}. A weight decay strategy is applied to prevent overfitting and enhance generalization. The batch size is set to 8, and the model is trained for a total of 150 epochs to ensure convergence.

3.3 Visual Results and Analysis

To visually evaluate the effectiveness of the proposed method, the fusion results of representative samples from the test set are analyzed. As illustrated in Figure 2, the first to third rows display the spectral bands of the high-resolution orthophoto, the PlanetScope SuperDove imagery, and the fused results, respectively.

Overall, an excellent balance between spatial detail reconstruction and spectral preservation is achieved in the fused images. Fine-grained spatial information derived from the orthophotos, such as canopy textures and road boundaries, is effectively injected into the fused output. Simultaneously, high consistency with the original SuperDove spectral characteristics is maintained, without noticeable spectral distortion or artifacts. These results demonstrate that a robust mapping across heterogeneous data sources is successfully established by the proposed model, despite significant radiometric discrepancies.

Regarding spatial detail representation, it is observed that even with a substantial resolution gap, fine-grained structural features at 0.16 m are successfully integrated into the fusion results. The clarity of object boundaries is significantly enhanced; specifically, individual tree canopy structures and land-cover edges are reconstructed with sharp boundaries and rich textural details. Such results demonstrate a superior capability in feature extraction and spatial texture injection.

In terms of spectral preservation, high stability is exhibited across various land-cover types, with a high degree of consistency maintained relative to the original PlanetScope SuperDove data. These findings indicate that robust generalization and stability are possessed by the proposed method within complex forest environments. Such radiometric integrity ensures that the fused data can support subsequent high-precision vegetation index calculations and various downstream remote sensing analysis tasks.

In summary, the experimental results demonstrate that the gap between single-source remote sensing imageries is effectively bridged in terms of spatial expressiveness and spectral completeness. By achieving the complementary advantages of heterogeneous remote sensing data, a highly reliable data augmentation solution is provided for refined, individual tree-level forest stress monitoring.

4. Conclusion

This paper addresses the bottleneck of limited joint representation of high spatial and spectral features in single-source remote sensing imagery, specifically for early forest stress monitoring at the individual tree scale. An unsupervised blind fusion framework for heterogeneous remote sensing images is proposed, leveraging implicit neural representation and low-rank tensor decomposition. By employing a SIREN, a coordinate-driven continuous reflectance field is constructed, formulating the fusion task as an inverse problem within the continuous domain. Without requiring prior knowledge of the PSF or SRF, orthorectified aerial images are utilized as guidance to super-resolve PlanetScope SuperDove multispectral data to a spatial resolution of 0.16 meters. Experimental results demonstrate that significant advantages in crown edge preservation and spectral fidelity are achieved by the proposed method, providing stable and reliable data support for capturing fine-grained critical information during the forest stress detection process.

Future research will be prioritized toward the utilization of the reconstructed high-spatial-resolution 530 nm band, in conjunction with other critical spectral channels, to

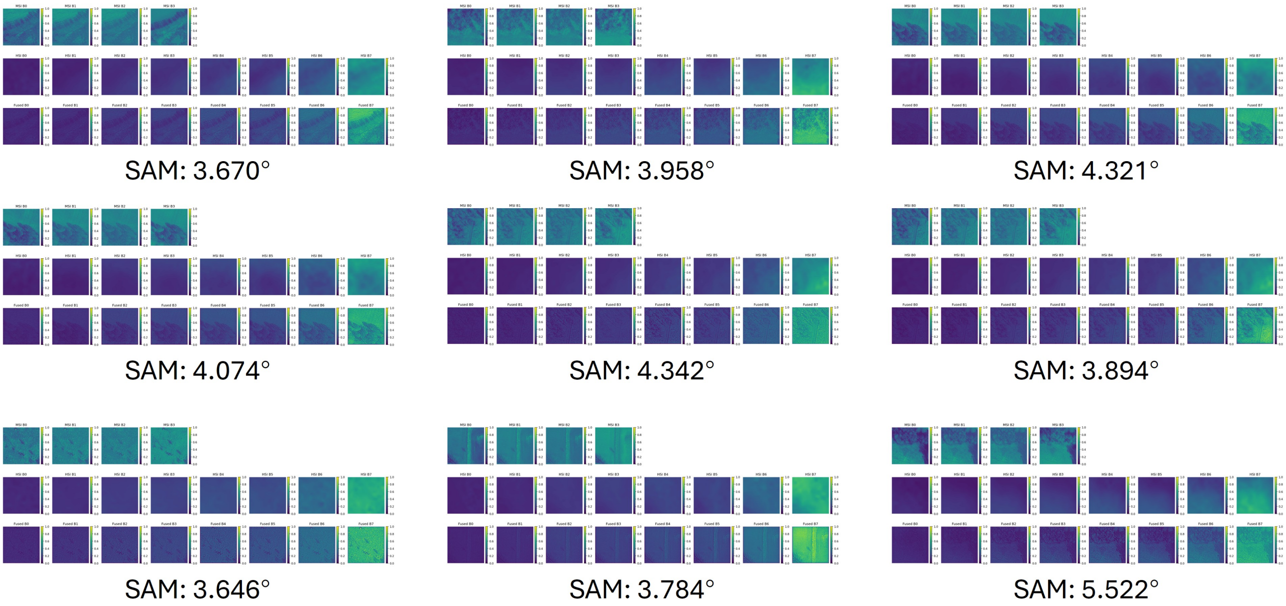

Figure 2. Visual Results.

develop targeted vegetation indices. Through this approach, an early-warning system for vegetation stress is intended to be established across the expansive forest regions of Sweden and the broader Nordic area.

5. Acknowledgements

The project is funded by skogssällskapet (project 2024-1275 to LH). The orthophotos were provided by the Lantmäteriet (Swedish mapping, cadastral and land registration authority) and the PlanetScope images were provided by Planet Labs PBC.

References

Algiriyage, N., Prasanna, R., Stock, K., Doyle, E. E., Johnston, D., 2022. Multi-source multimodal data and deep learning for disaster response: a systematic review. *SN Computer Science*, 3(1), 92.

Bacca, J., Arcos, C., Ramírez, J. M., Arguello, H., 2025. Middle-output deep image prior for blind hyperspectral and multispectral image fusion. *Signal Processing: Image Communication*, 132, 117247.

Cao, X., Lian, Y., Wang, K., Ma, C., Xu, X., 2024. Unsupervised Hybrid Network of Transformer and CNN for Blind Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15.

Chen, Y., Liu, S., Wang, X., 2021. Learning continuous image representation with local implicit image function. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8628–8638.

Huo, L., Koivumäki, N., Näsi, R., Honkavaara, E., 2025. Carotenoids increase as an indicator of early stress of trees: estimations using green shoulder indices from hyperspectral drone data and radiative transfer model prosail. *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 4185–4188.

Huo, L., Koivumäki, N., Oliveira, R. A., Hakala, T., Markelin, L., Näsi, R., Suomalainen, J., Polvivaara, A., Junttila, S., Honkavaara, E., 2024. Bark beetle pre-emergence detection using multi-temporal hyperspectral drone images: Green shoulder indices can indicate subtle tree vitality decline. *ISPRS Journal of Photogrammetry and Remote Sensing*, 216, 200-216.

Li, J., Li, Y., Wang, C., Ye, X., Heidrich, W., 2023. BUSIFusion: Blind Unsupervised Single Image Fusion of Hyperspectral and RGB Images. *IEEE Transactions on Computational Imaging*, 9, 94-105.

Lian, Z., Zhan, Y., Zhang, W., Wang, Z., Liu, W., Huang, X., 2025. Recent advances in deep learning-based spatiotemporal fusion methods for remote sensing images. *Sensors*, 25(4), 1093.

Shao, Z., Cai, J., 2018. Remote sensing image fusion with deep convolutional neural network. *IEEE journal of selected topics in applied earth observations and remote sensing*, 11(5), 1656–1669.

Song, B., Liu, P., Li, J., Wang, L., Zhang, L., He, G., Chen, L., Liu, J., 2022. MLFF-GAN: A Multilevel Feature Fusion With GAN for Spatiotemporal Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-16.

Vu, A. D., Nguyen, K. V., Bui, B. Q., Kamel, N., 2025. A comprehensive survey of super-resolution remote sensing image datasets: Evolution, challenges, and future directions. *IEEE Access*.

Wang, H., Cheng, S., Li, Y., Du, A., 2023. Lightweight remote-sensing image super-resolution via attention-based multilevel feature fusion network. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–15.

Wu, H., Zhang, K., Li, L., Liu, X., Zhao, D., Zhu, D., Shi, S., Wu, S., Shi, Z., 2025. An unsupervised model-guided mutual-learning network for spatio-temporal fusion of satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228, 16-27.

Xu, R., Yao, M., Chen, C., Wang, L., Xiong, Z., 2025. Continuous spatial-spectral reconstruction via implicit neural representation. *International Journal of Computer Vision*, 133(1), 106–128.

Yu, H., Ling, Z., Zheng, K., Gao, L., Li, J., Chanussot, J., 2024. Unsupervised hyperspectral and multispectral image fusion with deep spectral-spatial collaborative constraint. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–14.

Zhao, L., Huang, S., Li, J., 2025. Multimodal data fusion research review: From traditional methods to unsupervised learning methods. *Proceedings of the 2025 2nd Symposium on Big Data, Neural Networks, and Deep Learning*, 249–254.

Zhu, C., Deng, S., Song, X., Li, Y., Wang, Q., 2025. Mamba collaborative implicit neural representation for hyperspectral and multispectral remote sensing image fusion. *IEEE Transactions on Geoscience and Remote Sensing*.