SpectralNet-X: Transformer-based Lossy Compression for Hyperspectral Satellite Data

Jannik Sheikh[1,2], Jannick Kuester[1], Wolfgang Gross[1], Andreas Michel[1], Martin Weinmann[2]

[1] Fraunhofer IOSB, Image Analysis Group, Ettlingen, Germany
[2] Institut für Photogrammetrie und Fernerkundung, Karlsruhe Institute of Technology, Karlsruhe, Germany

Keywords: Hyperspectral Data Compression, Lossy Data Compression, Remote Sensing, Deep Learning, Satellite Data, Transformer

Abstract

Hyperspectral satellite missions generate massive data volumes that are difficult to transmit and store, making effective lossy compression a key enabling technology. We propose SpectralNet-X, a transformer-based autoencoder for spectral-only compression of spaceborne hyperspectral imagery at a fixed compression ratio of 16. The encoder maps each spectrum to a low-dimensional latent code using a 1D convolutional projection followed by stacked self-attention layers with rotary position embeddings and cross-attention pooling. The decoder reconstructs full-band spectra through an upsampling stack and per-band affine calibration. To improve reconstruction fidelity and generalization, SpectralNet-X is first pretrained via masked-signal reconstruction inspired by SimMIM and then fine-tuned with a mixed objective combining mean-squared error and spectral angle mapper (SAM) terms using a scheduled weighting scheme. We evaluate SpectralNet-X on the large-scale HySpecNet–11k benchmark and in a cross-sensor transfer setting, where models trained on HySpecNet–11k are tested on PRISMA hyperspectral scenes. Compared to three compression autoencoders, SpectralNet-X achieves the lowest angular reconstruction errors while maintaining competitive distortion metrics and substantially reducing the fraction of spectra with large SAM outliers. This study evaluates learned spectral compression under a normalized post-correction setting rather than in an end-to-end operational onboard radiance-preservation pipeline. The experiments rely on L2A and radiometrically corrected products rather than raw at-sensor radiances, so narrow atmospheric absorption features and residual Fraunhofer-line effects are not represented as in a true onboard scenario. The presented results should therefore be interpreted as evidence for learned spectral reconstruction under a controlled post-correction setting, not yet as direct validation for operational onboard deployment.

1. Introduction

Hyperspectral satellite missions provide dense, contiguous spectral sampling over hundreds of bands. This enables fine-grained material discrimination for applications in disaster management (Krekeler et al., 2023), land cover classification (Vali et al., 2020), and surveillance (Gross et al., 2022). As new constellations mature and revisit intervals shrink, downlink and storage budgets increasingly become the dominant bottlenecks (Villafranca et al., 2012, Shaharim et al., 2022, Melián et al., 2021). Efficient compression is therefore a prerequisite for scalable hyperspectral Earth observation. It reduces onboard memory pressure, increases effective downlink throughput, and preserves spectral fidelity that is critical for downstream analysis (Christophe, 2011).

We propose SpectralNet-X, a transformer-based spectral-only autoencoder at a fixed compression ratio. It combines a 1D convolutional projection with RoPE-enhanced self-attention and cross-attention pooling with learnable queries to form compact latents. The encoder is pretrained with a SimMIM-style masked-signal task. We then fine-tune with a mixed loss that combines MSE, spectral angle, and first-order differences to prioritize spectral shape fidelity. Cross-sensor transfer is supported by SRF-based projection between a common base grid and target sensors. We evaluate on HySpecNet–11k (Fuchs and Demir, 2023) and test cross-sensor transfer on PRISMA (Loizzo et al., 2018). The present study evaluates learned spectral compression under a normalized post-correction setting using radiometrically corrected data rather than top-of-atmosphere radiances. Accordingly, the results should not be interpreted as an end-to-end validation of operational onboard radiance-preserving compression.

This paper contributes:

- a transformer-based lossy compression framework tailored to satellite hyperspectral data,
- a benchmark-based evaluation on HySpecNet–11k with metrics sensitive to spectral shape distortions,
- an analysis at a fixed compression ratio CR = 16 that compares loss functions and quantifies the impact of pretraining and fine-tuning, and
- a VNIR-only cross-sensor transfer evaluation on PRISMA without additional fine-tuning.

2. Related Work

Classical hyperspectral compression reduces redundancy through transform coding and predictive models that decorrelate spectral bands and encode residuals (Du and Fowler, 2007, Qian, 2004, Wang and Celik, 2017). These methods are efficient, hardware-friendly, and particularly strong in settings where spectral integrity is critical.

Learning-based compression learns compact representations directly from data. Spectral 1D convolutional autoencoders emphasize local smoothness and achieve solid rate-distortion trade-offs at low complexity (A1D-CAE) (Kuester et al., 2021, Kuester et al., 2023). Transformer-based models such as HyCoT and HyCASS leverage attention to model long-range spectral dependencies and report competitive results for lossy compression (Fuchs et al., 2024, Fuchs et al., 2025, Sheikh et al., 2025). Masked reconstruction pretraining improves stability for spectral

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-3-2026-279-2026

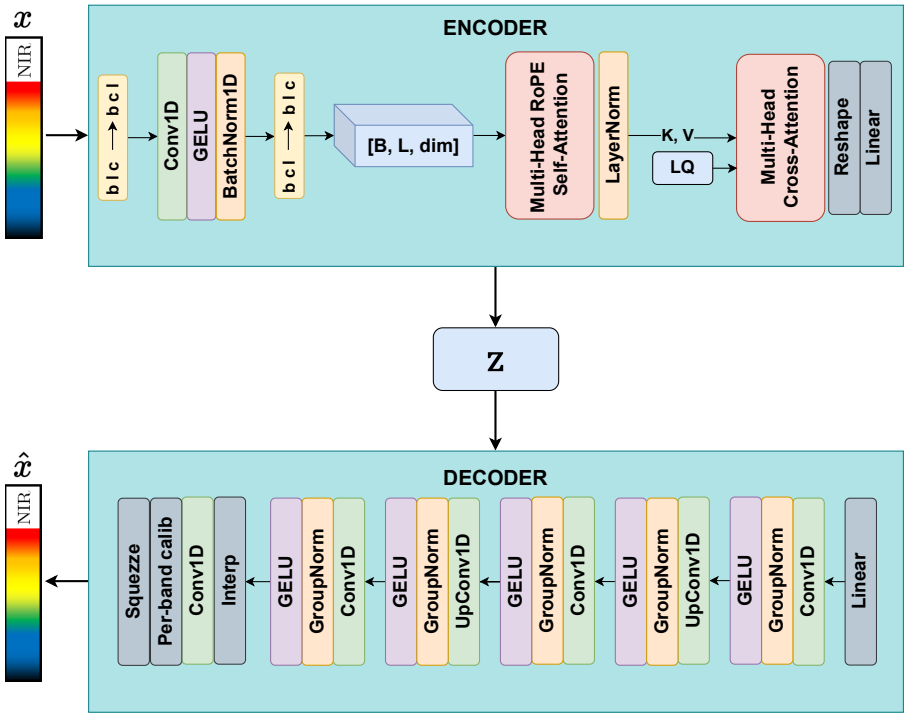

Figure 1. Overview of SpectralNet-X. Per pixel spectra are projected with a 1D convolution into an embedding sequence and encoded by Transformer blocks with RoPE. A small set of learnable queries (LQ) performs cross attention pooling to form the latent vector $\mathbf{z}$. The decoder linearly expands $\mathbf{z}$ to $[C_0, b_0]$ and applies a stack of UpConv1D blocks, followed by interpolation to the base grid length b_{base}, a 1D projection to C_{out}, and a per-band affine calibration step.

sequences (Scheibenreif et al., 2023), and spectral-angle-aware objectives help preserve spectral shape. More broadly, recent foundation-model approaches in remote sensing and hyperspectral imaging, such as RingMo and HyperSIGMA, demonstrate the utility of transformer-based backbones and large-scale self-supervised pretraining for learning transferable representations from high-dimensional remote-sensing data (Sun et al., 2022, Wang et al., 2025). While these works focus on general-purpose representation learning for downstream interpretation tasks rather than rate-distortion-optimized lossy compression, they provide additional motivation for employing related architectural components in hyperspectral compression, where the encoder must likewise capture informative spectral dependencies in a compact latent space.

Our approach builds on this learning-based paradigm. We use RoPE to encode relative spectral positions, learnable-query cross-attention pooling to obtain compact latent representations at a fixed compression ratio, SimMIM-style pretraining for stable initialization, and a mixed loss with a spectral-angle term. Cross-sensor transfer is addressed through SRF-based projection between a common base grid and target sensors. Baselines and training protocols are aligned to ensure fair comparison and are detailed in Section 4.

3. Methodology

Hyperspectral images (HSIs) consist of hundreds of contiguous spectral bands per pixel, providing rich spectral information at the cost of high dimensionality and, consequently, substantial storage and transmission overhead. For compression, we aim to learn a mapping from spectra represented on a common base wavelength grid to a low-dimensional latent space

$$f_\theta : \mathbb{R}^{b_{\text{base}}} \to \mathbb{R}^d \tag{1}$$

where b_{base} is the number of spectral bands on the common base grid and $d \ll b_{\text{base}}$ is the dimensionality of the compressed latent space. The decoder

$$g_\phi : \mathbb{R}^d \to \mathbb{R}^{b_{\text{base}}} \tag{2}$$

reconstructs the spectral signal on the base grid. Importantly, we focus exclusively on the spectral domain, ignoring spatial correlations between neighboring pixels. Each spectral vector $\mathbf{x} \in \mathbb{R}^{b_{\text{base}}}$ is first processed by a 1D convolutional projection, which acts as a local spectral filter and yields translation-invariant embeddings. We denote the resulting projected input sequence by $\mathbf{h}_0 \in \mathbb{R}^{b \times d_{\text{proj}}}$. To capture long-range spectral dependencies beyond local convolutional filters, we employ a transformer encoder with multi-head self-attention (MHSA) and Rotary Positional Embeddings (RoPE) (Su et al., 2024). Multi-head self-attention models content-dependent interactions between spectral elements, but by itself does not encode their relative order along the wavelength axis. Consequently, it cannot explicitly distinguish whether two bands are nearby or widely separated in the spectrum. Related interactive self-attention formulations

have been used in transformer-based hyperspectral image analysis (Yang et al., 2023). Unlike fixed or learnable absolute positional embeddings commonly used in Vision Transformers (ViTs) (Dosovitskiy et al., 2020), RoPE encodes relative spectral positions by applying complex rotations to query and key vectors during attention. This formulation allows the model to naturally generalize to unseen sequence lengths, preserves the notion of relative distance between spectral bands, and improves extrapolation across variable spectral resolutions.

After N layers of spectral self-attention, the encoder produces contextualized embeddings

$$\mathbf{S} = \text{Transformer Encoder}(\mathbf{h}_0), \quad \mathbf{S} \in \mathbb{R}^{b \times d_{\text{proj}}}. \tag{3}$$

A set of Q learnable query vectors attends to the encoder outputs via multihead scaled dot-product cross-attention. Unlike self-attention (Vaswani et al., 2017), where queries, keys, and values all originate from the same sequence, here the queries are learned parameters, whereas keys and values are taken from the contextualized spectral embeddings $\mathbf{S}$. The attended outputs are concatenated and linearly projected to form the latent embedding:

$$\mathbf{z} = \mathbf{W}_o \ \text{concat}_{i=1}^{Q} \ \text{CrossAttn}(\mathbf{q}_i, \mathbf{S}, \mathbf{S}), \tag{4}$$

where $\mathbf{q}_1, \ldots, \mathbf{q}_Q$ are the Q learnable query vectors and $\mathbf{W}_o$ is a learned output projection matrix.

The decoder maps a latent vector $\mathbf{z}$ to a base spectral grid of length b_{base} with C_{out} output channels. First, the latent is linearly expanded and reshaped into an initial feature map with C_0 channels over a short sequence:

$$\mathbf{h}_d = \text{reshape}(\mathbf{W}_e \mathbf{z}) \in \mathbb{R}^{C_0 \times b_0}, \quad b_0 \cdot 2^{N_{\text{up}}} \geq b_{\text{base}}, \tag{5}$$

where $\mathbf{W}_e$ is a learned projection and b_0 is a small starting length (e.g., $b_0 = 16$). The resolution is then increased through a stack of upsampling blocks. The number of blocks N_{up} is chosen such that $b_0 \cdot 2^{N_{\text{up}}} \geq b_{\text{base}}$. After the stack, a final linear interpolation adjusts the sequence length exactly to b_{base}, and a 1D projection maps features to C_{out} output channels, yielding a preliminary reconstruction $\tilde{\mathbf{x}}_{\text{base}}$ on the base grid. Each upsampling block applies:

- Conv1D → GroupNorm → GELU,
- Upsample×2 (linear) → Conv1D → GroupNorm → GELU,
- Conv1D → GroupNorm → GELU.

To absorb mild, band-specific biases, we append a lightweight per-band affine calibration layer,

$$\hat{\mathbf{x}}_{\text{base}} = \mathbf{a} \odot \tilde{\mathbf{x}}_{\text{base}} + \mathbf{c}, \tag{6}$$

where $\odot$ denotes element-wise multiplication, and $\mathbf{a}$ and $\mathbf{c}$ are learnable per-band scale and bias parameters initialized to 1 and 0, respectively. This layer is intended only to absorb residual band-specific biases in the reconstructed normalized signal. It is not a substitute for physical radiometric calibration, and its use does not by itself ensure preservation of absolute at-sensor radiances or full correction of band-dependent calibration errors. Consequently, the present study does not establish radiometric fidelity for quantitative downstream analyses in the original physical domain. GroupNorm is used throughout to stabilize training. Our model architecture is visualized in Figure 1.

To ensure cross-sensor compatibility, spectra are aligned to a common base wavelength grid before encoding and projected to the target sensor grid after decoding. On the input side, sensor spectra are mapped to the base grid via 1D linear interpolation with boundary extrapolation. The decoder reconstructs a spectrum on the base grid, applies the lightweight per-band affine calibration described above, and then performs sensor-specific projection using the target sensor's spectral response function (SRF). Specifically, we compute

$$\hat{\mathbf{x}}_{\text{tgt}} = \mathbf{K} \hat{\mathbf{x}}_{\text{base}}, \tag{7}$$

where $\mathbf{K} \in \mathbb{R}^{b_{\text{tgt}} \times b_{\text{base}}}$ is a row-normalized kernel whose i-th row is a Gaussian centered at the i-th target band and parameterized by the band's full width at half maximum (FWHM). All resampling occurs only at inference time, enabling adaptation to different sensors without modifying training.

Before compression-oriented fine-tuning, the encoder is pretrained using a SimMIM-style masked spectral reconstruction task, as illustrated in Figure 2. During this stage, a proportion of spectral bands is randomly masked and replaced with a learnable mask token, with the masking ratio gradually reduced over training. A lightweight linear reconstruction head is trained to predict the missing values, and the objective is to minimize the L1 error over masked positions.

The complete training pipeline comprises three stages. First, the encoder is pretrained with the SimMIM-style (Scheibenreif et al., 2023) masked reconstruction objective. Second, the full autoencoder is fine-tuned end-to-end using a mean squared error (MSE) reconstruction loss. Third, we switch to a mixed spectral loss combining MSE, spectral angle, and first-order difference terms to further refine spectral fidelity. The full training schedule and hyperparameters are reported in Section 4.

Design Rationale

Transformers for spectral modeling. Hyperspectral signatures are sequential in nature, with meaningful long-range dependencies across distant spectral bands. Convolutional filters capture only local correlations, whereas self-attention in transformers allows the model to flexibly relate spectral bands regardless of distance. Related interactive self-attention formulations have also been explored in transformer-based hyperspectral image analysis (Yang et al., 2023). The use of RoPE further enhances spectral modeling by encoding relative spectral order directly within the attention mechanism, which is critical for preserving fine-grained spectral structures (Su et al., 2024).

Cross-Attention Pooling with Learnable Queries. Compression requires condensing long spectral sequences into compact codes. Instead of using a fixed pooling operation we adopt cross-attention pooling with a set of learnable queries. Each query attends to the encoder output sequence and extracts complementary spectral content, and their aggregated representations are projected into the latent space. The idea of employing learnable queries to selectively extract information stems from DETR (Carion et al., 2020), where such queries are used in the decoder for object detection. In our case, they serve as spectral "slots" that summarize the sequence into a compact latent representation.

SimMIM-style pretraining. Training deep spectral encoders from scratch can be unstable, especially when the available training data are limited relative to model capacity. We therefore adopt a SimMIM-style masked spectral reconstruction task for

encoder pretraining (Scheibenreif et al., 2023), as illustrated in Figure 2. By forcing the model to recover randomly masked spectral bands from partial information, the encoder learns both local smoothness and long-range dependencies. This pretraining stage serves as a strong initialization, leading to improved convergence and downstream reconstruction.

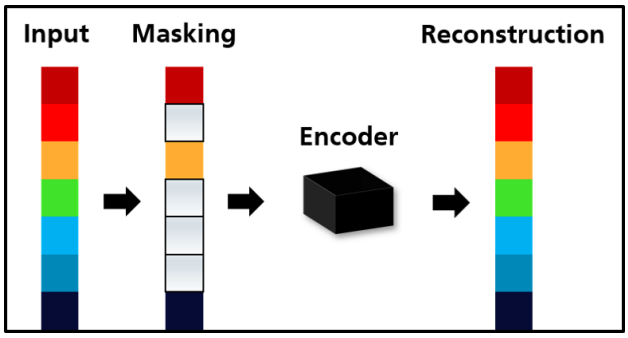

Figure 2. SimMIM pretraining overview. A subset of spectral bands is masked based on a masking ratio. The encoder processes the visible bands and learns to predict the missing ones in a self-supervised setup. A lightweight linear decoder performs the reconstruction, so most of the modeling capacity is in the encoder.

4. Experimental Setup

We conduct all experiments at a fixed compression ratio $CR = 16$ on HySpecNet–11k and evaluate cross–sensor transfer on PRISMA data. HySpecNet–11k provides 11,483 non–overlapping 128×128 EnMAP patches with contiguous spectra from the visible to short–wave infrared domain. L2A products exclude strong water–vapour bands and come with predefined train/validation/test splits (Fuchs and Demir, 2023). PRISMA is used only for out–of–domain evaluation. In the present study, we restrict the transfer experiment to the PRISMA VNIR detector range from 420 nm to 1010 nm with b=63 bands. This first cross–sensor analysis focuses on the VNIR range in order to evaluate transfer behavior in a controlled setting. PRISMA also provides SWIR measurements, but these are not included here. The reported cross–sensor results should therefore be interpreted as VNIR-only and do not yet cover transfer behavior for SWIR-specific absorption structures. The data are radiometrically corrected and georeferenced at 30 m ground sampling distance, and each scene contains 1000×1000 pixels (Loizzo et al., 2018). The six PRISMA test images used for cross-sensor evaluation span diverse scenes and different complexity levels (land, water, ice, and mixed environments) and are visualized in Figure 3. Prior to training and evaluation, zero spectral values in PRISMA that cause vertical stripes are replaced by the mean of their non–zero spectral neighbours, and the value range of all hyperspectral datasets (HySpecNet–11k and PRISMA) is scaled to $[0, 1]$ via min–max normalization with ϵ=10^{-8}. This normalization places all inputs on a common numerical scale, which is commonly used in deep learning to facilitate stable optimization during training (Goodfellow et al., 2016). Although the reconstructed outputs can in principle be mapped back to the original value range if the corresponding normalization parameters are retained, the present evaluation is carried out entirely in normalized signal space. Consequently, absolute radiometric values and physically calibrated signal magnitudes are not assessed directly, and quantitative conclusions should therefore be restricted to comparative reconstruction quality under this normalized protocol.

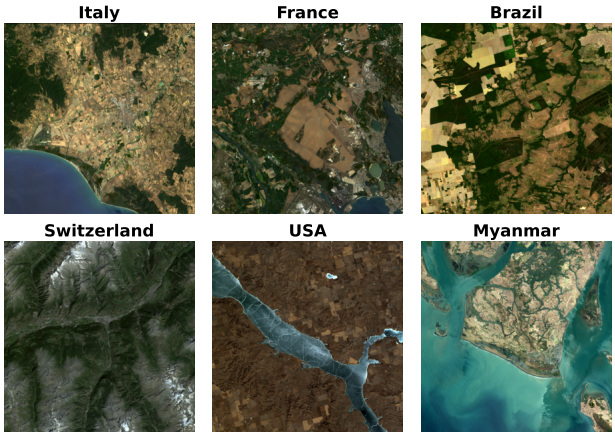

Figure 3. The six PRISMA test datasets visualized in RGB composition with bands $R = 32$, $G = 21$, and $B = 10$. For visualization purposes, the Switzerland image is gamma-corrected.

To quantify information loss introduced by compression, we report the peak signal–to–noise ratio (PSNR), the spectral angle (SAM), and the structural similarity index measure (SSIM). For PSNR, we use the implementation from (Detlefsen et al., 2022). PSNR is computed per spectral band on the $[0, 1]$ normalized data and averaged over bands and images (higher is better). SAM is computed as the mean angle in degrees between the input and reconstructed spectra, which captures spectral shape distortions (lower is better) (Kruse et al., 1993). To account for spatial structure, SSIM is evaluated per band on 2D windows using the standard `scikit-image` implementation and then averaged across bands and images (higher is better) (van der Walt et al., 2014). PSNR and SSIM are computed on min–max normalized data and are therefore interpreted strictly in normalized signal space. Since the same normalization is applied within the evaluation protocol, both metrics remain suitable for comparative assessment of reconstruction quality across methods. However, they do not quantify preservation of absolute radiometric values or physically calibrated signal magnitudes in the original domain. SAM, by contrast, is invariant to uniform scaling and is therefore particularly informative for assessing spectral-shape preservation under the present evaluation protocol.

SpectralNet-X (ours) is trained in three stages. First, we perform SimMIM-style pretraining of the encoder for 200 epochs. The masking ratio is linearly decreased from 80% to 20%, with a Gaussian-initialized mask token of standard deviation 0.02. We optimize using AdamW (Loshchilov and Hutter, 2017) with learning rate 1×10^{-3}, weight decay 1×10^{-2}, and a cosine scheduler with $T_{\max} = 200$ and 10 warmup epochs. We train the encoder with an L_1 loss computed only on the masked bands and use the same objective to select the best checkpoint for fine-tuning:

$$\mathcal{L}_{\text{mask}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left| \hat{x}_{\text{base},i} - x_i \right|, \tag{8}$$

where $\mathcal{M}$ denotes the set of masked band indices. In the second stage, we fine-tune the full autoencoder for 200 epochs using mean squared error (MSE) as the reconstruction loss. We use AdamW with separate learning rates for the pretrained encoder and the remaining modules (decoder and new norms), namely $\text{lr}_{\text{enc}} = 5 \times 10^{-4}$ and $\text{lr}_{\text{other}} = 1 \times 10^{-3}$ with weight decay 5×10^{-3}, and apply a cosine scheduler with $T_{\max} = 200$ and 5 warmup

epochs.

In the third stage, we initialize from the best MSE checkpoint and optimize a mixed spectral objective

$$\mathcal{L}_{\text{mix}} = \underbrace{\|\hat{\mathbf{x}}_{\text{base}} - \mathbf{x}\|_2^2}_{\mathcal{L}_{\text{MSE}}} + \beta \underbrace{\left(1 - \cos(\hat{\mathbf{x}}_{\text{base}}, \mathbf{x})\right)}_{\mathcal{L}_{\text{SA}}} + \lambda_d \underbrace{\left\|\Delta\hat{\mathbf{x}}_{\text{base}} - \Delta\mathbf{x}\right\|_2^2}_{\mathcal{L}_{\text{deriv}}}, \tag{9}$$

where $\cos(\hat{\mathbf{x}}_{\text{base}}, \mathbf{x}) = \frac{\hat{\mathbf{x}}_{\text{base}}^\top \mathbf{x}}{\|\hat{\mathbf{x}}_{\text{base}}\|_2 \|\mathbf{x}\|_2}$ is the spectral-angle cosine (averaged over the batch), and $\Delta\mathbf{x} = \mathbf{x}_{2:} - \mathbf{x}_{1:-1}$ denotes first-order spectral differences along the wavelength axis. The spectral-angle weight is linearly ramped as

$$\beta_t = \beta_{\max} \cdot \min\left(1, \frac{t+1}{E_{\text{warm}}}\right), \quad \beta_{\max} = 0.7,\ E_{\text{warm}} = 124. \tag{10}$$

We selected $\beta_{\max} = 0.7$ empirically. Further increases did not improve performance and this value was the breakpoint. The warm-up length $E_{\text{warm}} = 124$ was determined by gradually extending the ramp, with improvements saturating at 124. Upon switching to $\mathcal{L}_{\text{mix}}$ we freeze the encoder for the first two epochs and apply an LR cooldown factor of 0.3 during the first ten epochs. The learning rate for the encoder is set to $\text{lr}_{\text{enc}} = 5 \times 10^{-4}$, while the learning rate for other components is $\text{lr}_{\text{other}} = 1 \times 10^{-3}$, with a weight decay of 5×10^{-3}. In the final phase we fix β at 0.7, enable the derivative regularizer with $\lambda_d = 5 \times 10^{-3}$, reduce the learning rates to $\text{lr}_{\text{enc}} = 2 \times 10^{-4}$ and $\text{lr}_{\text{other}} = 5 \times 10^{-4}$, and continue training for additional epochs.

Training of SpectralNet-X is performed with a batch size of 3072 across 2 H100 GPUs. The encoder comprises 6 Transformer layers with model dimension $d_{\text{model}} = 512$, feed-forward dimension $d_{\text{ff}} = 1024$, and 6 self-attention heads. Latent pooling uses Multi-Query Attention with 24 query tokens and 2 query heads.

We compare our model with three different compression models. A1D-CAE is a pixelwise one–dimensional convolutional autoencoder that operates along the spectral axis only, using strided 1D convolutions for encoding and transposed convolutions for decoding, which emphasizes local spectral smoothness while ignoring spatial redundancy (Kuester et al., 2023). HyCoT is a transformer–based autoencoder that tokenizes spectra, models long–range spectral dependencies with self–attention, and compresses them into a compact latent through a projection head, followed by a lightweight decoder (Fuchs et al., 2024). HyCASS is a configurable transformer autoencoder with an adjustable compression adapter; it can combine pixelwise spectral encoding with stacked transformer stages and allows setting the latent dimensionality to meet a target compression ratio. In our spectral focus we use its spectral pathway and fix the latent to achieve $CR = 16$ (Fuchs et al., 2025).

All models are trained on the HySpecNet–11k *hard* split; train, validation, and test sets are adopted exactly as defined in the dataset paper (Fuchs and Demir, 2023). We fix the compression ratio to $CR = 16$ by selecting latent dimensionality and quantization parameters accordingly. For HyCoT and HyCASS, we rely on the official reference implementations provided by the authors and follow their training pipelines unless stated otherwise (Fuchs et al., 2024, Fuchs et al., 2025, Remote Sensing Image Analysis (RSIM) Group, 2025b, Remote Sensing Image Analysis (RSIM) Group, 2025a). In our environment, these reference implementations were executed on a single A100 GPU. For HyCoT, the batch size is set to 5 to prevent out–of–memory errors with our hardware configuration, all other settings follow the authors' recommendations. The convolutional baseline A1D-CAE is trained with Adam, large pixelwise batches, and up to 200 epochs; a learning–rate scheduler `ReduceLROnPlateau` is applied with *factor* 0.9, *patience* 6, *threshold* 10^{-5} (*relative*), *min_lr* set to the implementation default, and *eps* 10^{-8}. The best checkpoint per model is selected by validation SA and used for all metrics. Importantly, HyCoT and HyCASS do not provide a sensor–transfer setting in their released implementations and are therefore not considered for the cross–sensor experiment. In the cross–sensor evaluation, we report results for A1D-CAE and SpectralNet-X only.

5. Results

We present PSNR in dB, SAM in degrees, and SSIM in $[0, 1]$, and report the standard deviation as $\pm\sigma$.

Stage	PSNR (dB) ↑	SAM ↓	SSIM ↑
1	**40.336 ± 4.950**	3.761 ± 5.564	**0.959 ± 0.058**
2	39.597 ± 5.620	3.405 ± 4.966	0.954 ± 0.067
3	39.914 ± 5.661	**3.378 ± 5.006**	0.955 ± 0.064

Table 1. Performance of different training stages of SpectralNet-X over 1140 test images. Best values per metric are highlighted in **bold**.

Stage-wise performance is summarized in Table 1. Each stage is evaluated on the same 1140 test images. PSNR is bandwise on normalized inputs. SAM is the mean angle between input and reconstruction. SSIM is computed per band on 2D windows and then averaged.

Table 2 compares all models on the full test set. All models operate at a fixed compression ratio CR = 16. Metrics are computed with the same pipeline to ensure consistency. We include standard deviations to show the spread across scenes.

Model	PSNR (dB) ↑	SAM ↓	SSIM ↑
SpectralNet-X	39.914 ± 5.661	**3.378 ± 5.006**	0.955 ± 0.064
A1D-CAE	39.100 ± 7.342	5.332 ± 7.481	0.941 ± 0.098
HyCoT	**41.027 ± 5.458**	4.099 ± 7.030	**0.956 ± 0.067**
HyCASS	39.578 ± 4.955	4.667 ± 7.189	0.935 ± 0.104

Table 2. Reconstruction performance of all models over all 1140 test images. Best mean values per column are highlighted in **bold**. SAM (in degrees) is the most relevant metric for spectral fidelity.

Figure 4 shows the distribution of per-pixel SAM errors over all test images. We use three bins. SAM up to 3°. SAM between 3° and 10°. SAM above 10°. Counts are shown on a logarithmic scale. The binning reflects typical thresholds for spectral fidelity. SAM $\leq 3°$ indicates a good reconstruction (Licciardi and Chanussot, 2018), motivating these spectral-fidelity thresholds.

Since a subset of test images contains severe artifacts, as shown in Figure 7, we also report a filtered analysis in Table 3. The subset contains images with mean SAM less than or equal to 3°. We highlight only the image count N to focus on the number of reconstructions that meet the SAM threshold. PSNR, SAM, and SSIM are provided for context.

Cross-sensor results on PRISMA are shown in Table 4. We evaluate each scene independently and report per-scene metrics and the overall mean.

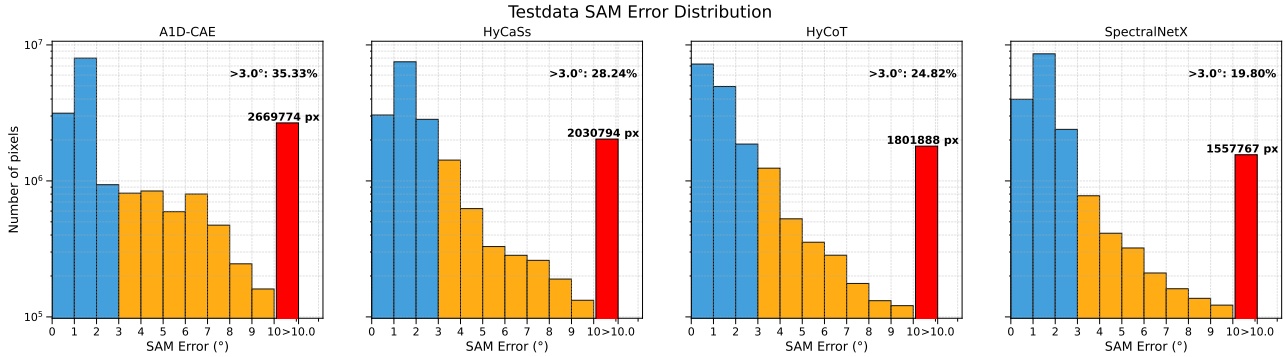

Figure 4. Distribution of per-pixel spectral angle mapper (SAM) errors for all test images and models. For each model, the histogram shows the number of pixels (log scale) for different SAM error ranges: blue bars indicate pixels with SAM $\leq 3°$ (the most relevant for high spectral fidelity), orange bars indicate SAM between 3° and 10°, and red bars indicate pixels with SAM $> 10°$. The total number and percentage of pixels with SAM $> 3°$ are annotated for each model.

To facilitate reproducibility, we follow a single metric implementation across all experiments. PSNR uses TorchMetrics. SSIM uses scikit-image with default windowing. SAM is the mean spectral angle in degrees.

Model	N ↑	PSNR (dB) ↑	SAM ↓	SSIM ↑
SpectralNet-X	**897**	41.454 ± 4.365	1.441 ± 0.582	0.977 ± 0.037
A1D-CAE	708	**42.651 ± 3.774**	1.410 ± 0.510	**0.994 ± 0.004**
HyCoT	822	42.559 ± 4.399	**1.285 ± 0.651**	0.982 ± 0.031
HyCASS	761	40.959 ± 3.802	1.571 ± 0.618	0.981 ± 0.032

Table 3. Count and quality metrics for images with spectral angle $\leq 3°$. Best values per metric are highlighted in **bold**. SpectralNet-X reconstructs the highest number of images meeting the SAM threshold.

Figure 6 displays representative reference and reconstructed spectra in normalized signal space for two example pixels, a vegetation spectrum and a rock spectrum. While all methods accurately reproduce the general spectral shape, the enlarged views show local differences in the reconstruction in selected wavelength ranges. These qualitative examples complement the aggregated metrics by illustrating how reconstruction errors manifest directly in the spectral domain.

6. Discussion

Table 1 shows that adding a mixed loss on top of standard MSE finetuning (Stage2) markedly improves spectral fidelity, substantially reducing SAM without degrading PSNR. Stage3, initialized from Stage2 (itself derived from Stage1), consolidates these improvements and achieves the best mean SAM and SSIM with competitive PSNR, reflecting the cumulative benefit of the multi-stage pipeline.

As shown in Table 2, HyCoT achieves the highest PSNR and SSIM values, while SpectralNet-X obtains the lowest mean SAM, the most relevant metric for spectral fidelity. This suggests that, although some models may achieve higher overall signal-to-noise ratios, SpectralNet-X is more effective in preserving spectral information. To further analyze model performance at the pixel level, Figure 4 shows the distribution of per-pixel SAM errors for all test images. Here, pixels with SAM errors smaller than 3° (blue bars) represent high spectral fidelity, those between 3° and 10° (orange bars) indicate moderate errors, and those above 10° (red bars) correspond to large spectral distortions. The proportion of pixels with high SAM errors is lowest for SpectralNet-X, confirming its robustness in spectral reconstruction across most pixels.

Figure 5 shows SAM error maps in degrees for various HySpecNet–11k test cases across all models. A uniform color scale is used in all subfigures, allowing the SAM values shown to be directly compared both between models and between scenes. The first row shows a mix of different vegetation, rocks, and water. All models generally exhibit good reconstruction, but SpectralNet-X performs best in the water area at the top of the image, where the other models have a higher SAM value. In the urban view in the second row, SpectralNet-X exhibits fewer errors along building edges and at the transitions between asphalt and vegetation, and shows fewer outliers in the cast shadows. In the land surface with clouds in the third row, our method keeps errors within bright cloud cores and at cloud edges to a moderate level. The complex scene in the fourth row exhibits variations in shadow and depth, with increasingly darker areas. All models have significant difficulty ($SAM \geq 3°$) reconstructing this scene, although SpectralNet-X is able to reconstruct some information. Similar results can be observed in the fifth row. All models are able to reconstruct the structure (possibly an island), but exhibit a high SAM value in the blue region (possibly water). Of all the models, our approach is the most capable of reconstructing the scene. In the last row, the heterogeneous terrain includes craters whose steep walls cast deep shadows that create extended low-signal regions. SpectralNet-X keeps errors smaller inside the dark areas and along the bright rims and reduces spurious high-SAM spots, while HyCoT and HyCASS show increasing SAM as radiance decreases. These qualitative patterns reinforce the quantitative results and suggest that reconstruction accuracy is strongly dependent on the overall signal level. In regions with lower radiation intensity, errors tend to increase, whereas brighter and spectrally smoother regions are reconstructed more reliably. This suggests that signal-dependent noise and a reduced effective signal-to-noise ratio in regions with weak signals continue to pose a challenge for learning-based compression.

To complement the aggregate metrics, Figure 6 shows representative reference and reconstructed spectral signatures for vegetation and rock pixels in normalized signal space. Across both examples, the overall spectral shape is preserved well by all methods, whereas the zoomed views show that the remaining reconstruction errors are concentrated in localized wavelength

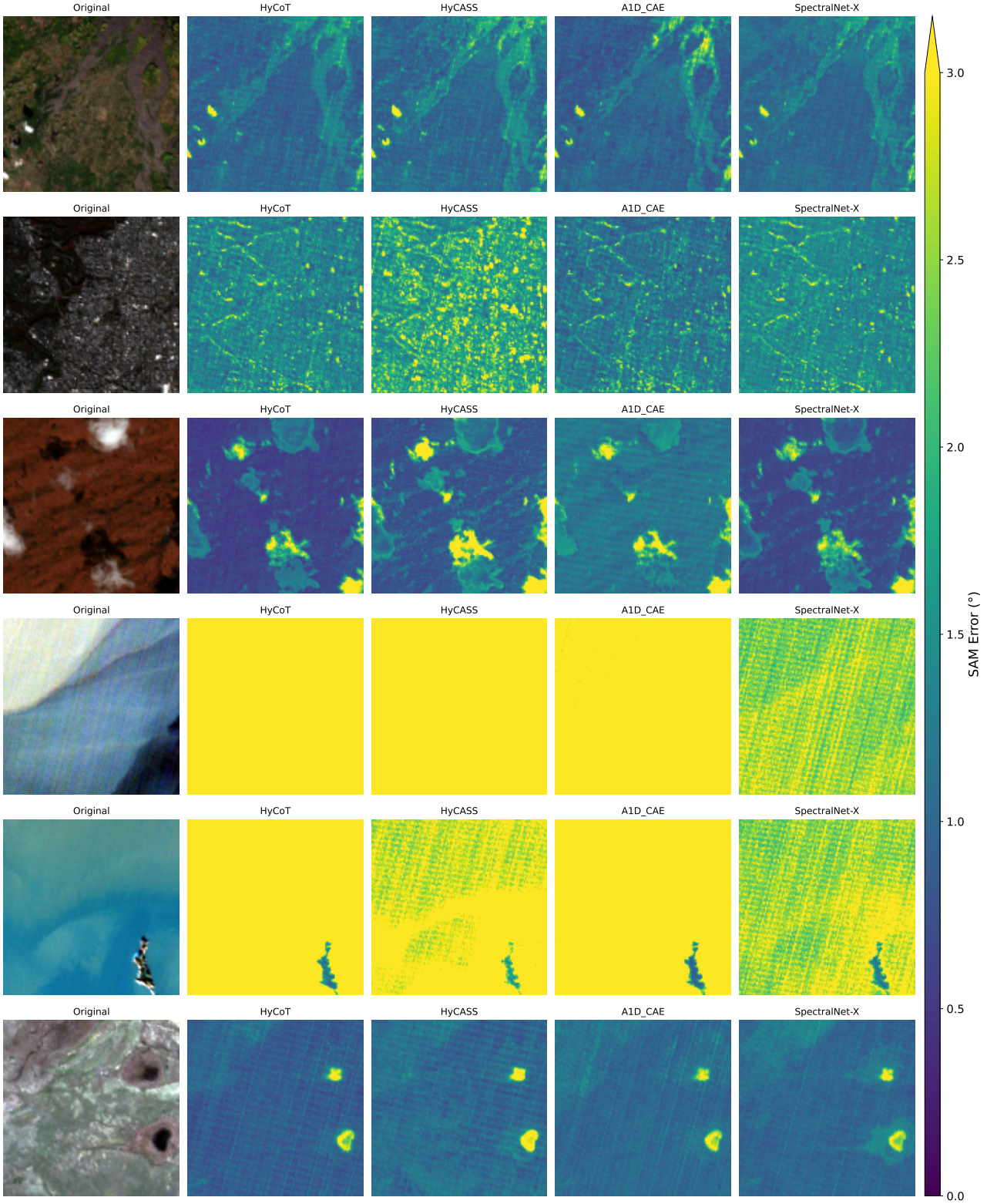

Figure 5. SAM error maps (°) across all models for different HySpecNet–11k test dataset examples. Lower values indicate higher spectral fidelity. Colors encode SAM in degrees using a common colour scale across all panels.

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-3-2026-279-2026 | © Author(s) 2026. CC BY 4.0 License.

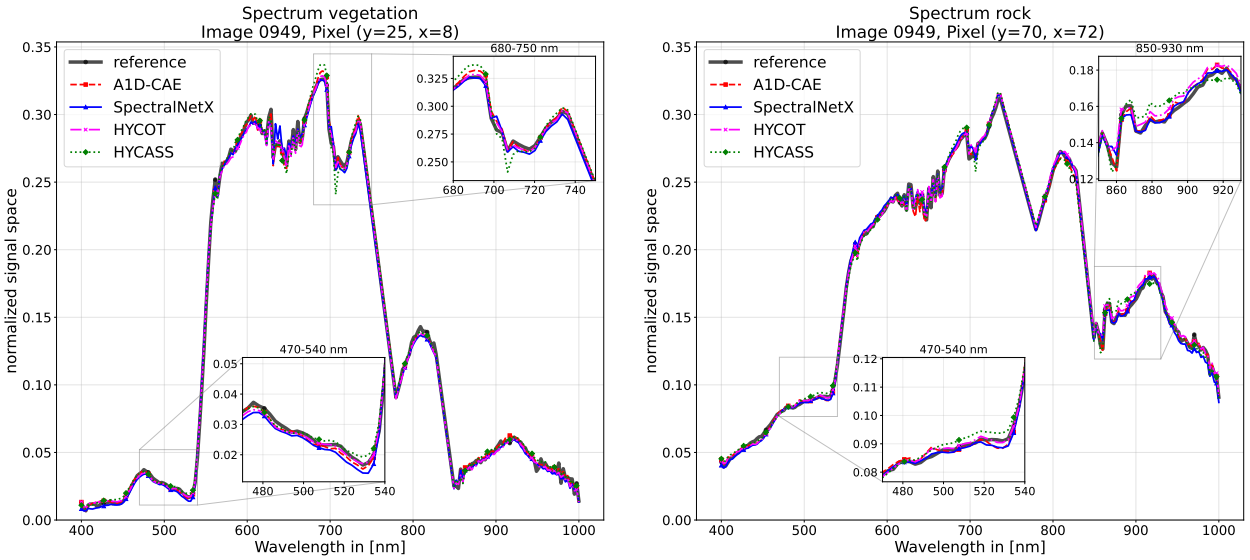

Figure 6. Example reference and reconstructed spectra in normalized signal space for two sample pixels, with vegetation spectrum on the left and rock spectrum on the right. The highlights show selected wavelength ranges with visible local differences in reconstruction.

Dataset	SpectralNet-X			A1D-CAE		
	PSNR (dB) ↑	SAM ↓	SSIM ↑	PSNR (dB) ↑	SAM ↓	SSIM ↑
1 (Italy)	**32.259**	**6.396**	**0.934**	30.140	10.872	0.915
2 (France)	**35.324**	**7.513**	0.913	32.876	12.594	**0.921**
3 (Brazil)	**27.929**	**7.260**	**0.888**	22.759	12.765	0.784
4 (Switzerland)	**35.633**	**7.838**	**0.939**	32.763	13.533	0.914
5 (USA)	30.547	**4.503**	**0.980**	**33.397**	5.994	0.970
6 (Myanmar)	**30.340**	**5.392**	**0.959**	28.808	9.509	0.919
Mean	**32.005**	**6.484**	**0.935**	30.124	10.878	0.904

Table 4. Reconstruction performance on the six PRISMA test datasets. Best values per row and metric are highlighted in **bold**.

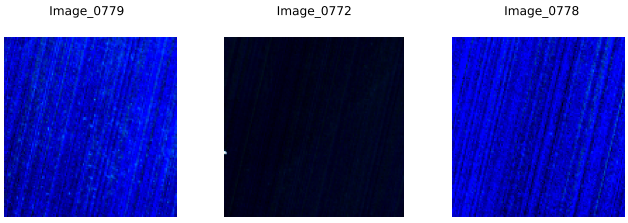

Figure 7. The provided images exhibit severe distortions and artifact patterns that make reconstruction particularly challenging.

intervals. These deviations mainly appear as mild smoothing, small amplitude differences, and slight shape distortions rather than large spectral displacements.

Figure 7 indicates that some test images exhibit significant distortion and artifacts attributable to the capture or preprocessing, which can have a disproportionately large impact on the overall evaluation. At the same time, such scenes should not be dismissed as operationally irrelevant, because visually uninformative images may still contain rare spectral signatures or small targets of practical interest. To address their effect on aggregate benchmark statistics, we also report results on a subset filtered to images with SAM smaller or equal to 3°, summarized in Table 3. SpectralNet-X reconstructs the largest number of images, indicating a higher degree of generalization and robustness. While HyCoT and A1D-CAE achieve slightly higher PSNR and SSIM values on this subset, the ability of SpectralNet-X to provide reliable reconstructions on a larger portion of the dataset is particularly noteworthy.

Cross-sensor transfer on PRISMA remains challenging. As shown in Table 4, SpectralNet-X attains moderate absolute performance, yet consistently improves spectral fidelity over A1D-CAE, with gains in PSNR/SSIM on most scenes. These results suggest that base grid alignment and SRF-based projection mitigate, but do not eliminate, domain gaps due to sensor characteristics, calibration differences, and scene-specific artifacts. Interestingly, scenes dominated by land exhibit higher SAM than ice or water. Land scenes are more spectrally complex and heterogeneous, with mixed materials and variable illumination, which makes reconstruction and alignment harder and raises SAM. In contrast, ice and water spectra are smoother and more homogeneous, so resampling and projection introduce fewer distortions, yielding lower SAM. The present cross-sensor analysis is restricted to PRISMA VNIR data, and transfer behavior in the SWIR range is not assessed here.

7. Conclusion

This work provides a broad evaluation of hyperspectral image reconstruction on the HySpecNet–11k dataset. We compare multiple architectures and training configurations with a focus on

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-3-2026-279-2026 | © Author(s) 2026. CC BY 4.0 License.

spectral fidelity, robustness, and generalization. The results show that design and configuration choices strongly affect reconstruction quality. Among the considered metrics, SAM is particularly suited to assess spectral reconstruction quality, because it directly compares the shape of the spectral signatures and is invariant to overall intensity scaling. In contrast, PSNR operates on per-band intensity differences and can reward reconstructions that smooth out high-frequency spectral features, even though such smoothing distorts the underlying spectral signature. Since many downstream hyperspectral analysis tasks (e.g. material classification or unmixing) primarily depend on the integrity of spectral signatures, SAM provides a more informative indicator of reconstruction quality than PSNR. SAM values below about 3° are typically regarded as indicative of high-fidelity spectral reconstruction (Licciardi and Chanussot, 2018). In the cross-sensor setting from HySpecNet–11k to PRISMA, absolute SAM scores are moderate, but SpectralNet-X consistently yields the best spectral fidelity and generalizes to the new sensor without additional training. The increased error is partly explained by the mismatch in spectral resolution between the two sensors, which alters the spectral signatures, yet key features are still reasonably well reconstructed. We hypothesize that brief fine-tuning on limited PRISMA data may further reduce SAM and narrow the gap to in-sensor performance.

Future work will explore improved composite loss formulations to better balance spectral fidelity and distortion. It will also extend the cross-sensor evaluation to the full PRISMA spectral range, including SWIR, in order to assess whether the present findings transfer beyond the controlled VNIR-only setting and to analyze reconstruction behavior for SWIR-specific absorption structures. Beyond aggregate benchmark performance, operational deployment would require dedicated validation on atypical scenes, rare materials, narrow-band absorption features, and single-pixel or sub-pixel targets that may be underrepresented in the training data.

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. *CoRR*, abs/2005.12872. https://arxiv.org/abs/2005.12872.

Christophe, E., 2011. Hyperspectral data compression tradeoff. *Optical Remote Sensing: Advances in Signal Processing and Exploitation Techniques*, Springer, 9–29.

Detlefsen, N. S., Borovec, J., Schock, J., Harsh, A., Koker, T., Liello, L. D., Stancl, D., Quan, C., Grechkin, M., Falcon, W., 2022. Torchmetrics - measuring reproducibility in pytorch. `https://github.com/Lightning-AI/torchmetrics`.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929. https://arxiv.org/abs/2010.11929.

Du, Q., Fowler, J. E., 2007. Hyperspectral Image Compression Using JPEG2000 and Principal Component Analysis. *IEEE Geoscience and Remote Sensing Letters*, 4(2), 201-205.

Fuchs, M. H. P., Demir, B., 2023. Hyspecnet-11k: a large-scale hyperspectral dataset for benchmarking learning-based hyperspectral image compression methods. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, IEEE.

Fuchs, M. H. P., Rasti, B., Demir, B., 2024. Hycot: A transformer-based autoencoder for hyperspectral image compression. *2024 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, IEEE, 1–5.

Fuchs, M. H. P., Rasti, B., Demir, B. u., 2025. Adjustable Spatio-Spectral Hyperspectral Image Compression Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1-14.

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

Gross, W., Queck, F., Schreiner, S., Vögtli, M., Kuester, J., Mispelhorn, J., Kneubühler, M., Middelmann, W., 2022. A multi-temporal hyperspectral camouflage detection and transparency experiment. *Target and Background Signatures VIII SPIE*, 12270, 11–19.

Krekeler, M. P. S., Burke, M., Allen, S., Sather, B., Chappell, C., McLeod, C. L., Loertscher, C., Loertscher, S., Dawson, C., Brum, J. et al., 2023. A novel hyperspectral remote sensing tool for detecting and analyzing human materials in the environment: a geoenvironmental approach to aid in emergency response. *Environmental Earth Sciences*, 82(4), 1–7.

Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heidebrecht, K. B., Shapiro, A., Barloon, P., Goetz, A. F., 1993. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3), 145–163.

Kuester, J., Gross, W., Middelmann, W., 2021. 1D-CONVOLUTIONAL AUTOENCODER BASED HYPERSPECTRAL DATA COMPRESSION. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B1-2021, 15–21. https://isprs-archives.copernicus.org/articles/XLIII-B1-2021/15/2021/.

Kuester, J., Gross, W., Schreiner, S., Middelmann, W., Heizmann, M., 2023. Adaptive two-stage multisensor convolutional autoencoder model for lossy compression of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–22.

Licciardi, G., Chanussot, J., 2018. Spectral transformation based on nonlinear principal component analysis for dimensionality reduction of hyperspectral images. *European Journal of Remote Sensing*, 51(1), 375–390.

Loizzo, R., Guarini, R., Longo, F., Scopa, T., Formaro, R., Facchinetti, C., Varacalli, G., 2018. Prisma: The italian hyperspectral mission. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 175–178.

Loshchilov, I., Hutter, F., 2017. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101. http://arxiv.org/abs/1711.05101.

Melián, J. M., Jiménez, A., Díaz, M., Morales, A., Horstrand, P., Guerra, R., López, S., López, J. F., 2021. Real-Time Hyperspectral Data Transmission for UAV-Based Acquisition Platforms. *Remote Sensing*, 13(5), 850.

Qian, S.-E., 2004. Hyperspectral data compression using a fast vector quantization algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8), 1791-1798.

Remote Sensing Image Analysis (RSIM) Group, 2025a. Hycass: Reference implementation. https://git.tu-berlin.de/rsim/hycass/-/tree/main?ref_type=heads. Technische Universität Berlin. Accessed 2025-11-12.

Remote Sensing Image Analysis (RSIM) Group, 2025b. Hycot: Reference implementation. https://git.tu-berlin.de/rsim/hycot/-/tree/main?ref_type=heads. Technische Universität Berlin. Accessed 2025-11-12.

Scheibenreif, L., Mommert, M., Borth, D., 2023. Masked vision transformers for hyperspectral image classification. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2166–2176.

Shaharim, N. A. N., Tan, L. C., Zainul, A. Z., Noor, N. R. M., 2022. Parallelization of CCSDS Hyperspectral Image Compression Using OpenMP. *Journal of Engineering Science*, 18(1), 1–16.

Sheikh, J., Gross, W., Michel, A., Weinmann, M., Kuester, J., 2025. Transformer-based lossy hyperspectral satellite data compression. K. Schulz, U. Michel, K. G. Nikolakopoulos, V. Gagliardi, A. C. M. Teodoro (eds), *Earth Resources and Environmental Remote Sensing/GIS Applications XVI*, 13671, International Society for Optics and Photonics, SPIE, 136710Q.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y., 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568, 127063. https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H. et al., 2022. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–22.

Vali, A., Comai, S., Matteucci, M., 2020. Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. *Remote Sensing*, 12(15), 2495.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., the scikit-image contributors, 2014. scikit-image: image processing in Python. *PeerJ*, 2, e453. https://doi.org/10.7717/peerj.453.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I., 2017. Attention is all you need. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems*, 30, Curran Associates, Inc.

Villafranca, A. G., Corbera, J., Martín, F., Marchán, J. F., 2012. Limitations of Hyperspectral Earth Observation on Small Satellites. *Journal of Small Satellites*, 1(1), 19–29.

Wang, D., Hu, M., Jin, Y., Miao, Y., Yang, J., Xu, Y., Qin, X., Ma, J., Sun, L., Li, C. et al., 2025. HyperSIGMA: Hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, H., Celik, T., 2017. Sparse Representation-Based Hyperspectral Data Processing: Lossy Compression. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5), 2036-2045.

Yang, J., Du, B., Zhang, L., 2023. From center to surrounding: An interactive learning framework for hyperspectral image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197, 145–166.