Latent Dirichlet Transformer VAE for Hyperspectral Unmixing with Bundled Endmembers

by

Giancarlo Giannetti

A thesis submitted to the
School of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Masters of Science in **Computer Science**

Faculty of Science
Ontario Tech University
Oshawa, Ontario, Canada

© Giancarlo Giannetti 2026

The University of Ontario Institute of Technology requires the Certificate of Approval (CoA) to be included as page (ii) of the PRINTED version. Check the source file on how to add the provided CoA to your thesis.

Abstract

Hyperspectral images capture rich spectral information that enables per-pixel material identification; however, spectral mixing often obscures pure material signatures. Hyperspectral unmixing aims to decompose these mixed spectra into endmember signatures and their corresponding abundances, but remains challenging due to spectral variability, nonlinear mixing effects, and the scarcity of reliable ground-truth data.

In this work, we propose a deep generative model and two "stepping stone" models for hyperspectral unmixing that integrate transformer architectures with variational autoencoders. The proposed approach leverages self-attention mechanisms to capture global spectral–spatial dependencies within image patches, while a Dirichlet latent distribution enforces the physical constraints of abundance estimation. Building on this foundation, we introduce the Latent Dirichlet Transformer VAE (LDVAE-T), which models endmembers as distributions rather than fixed spectra, enabling the representation of intrinsic spectral variability through learned mean and covariance structures.

We evaluate the proposed models on multiple benchmark datasets, including Samson, Jasper Ridge, and HYDICE Urban, and compare their performance against SOTA methods. Experimental results demonstrate improved accuracy in both abundance estimation and endmember extraction. Overall, this work presents a physically grounded and flexible framework for hyperspectral unmixing that effectively captures both global structure and material variability.

Keywords: Hyperspectral Unmixing; Remote Sensing; Vision Transformer; Variational Autoencoder

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I authorize the Ontario Tech University to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize the Ontario Tech University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Giancarlo Giannetti

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Faisal Qureshi, for his guidance, support, and encouragement throughout this research. His feedback and expertise were invaluable in helping me develop this work and complete my thesis.

I would also like to thank my parents, Wilma and Dino, for their constant support and encouragement throughout my studies. I am equally grateful to my aunt Franca for supporting me along the way.

Finally, I would like to thank my fiancé, Alexie, for her patience, encouragement, and support throughout this entire process. Her presence made the difficult moments easier and the successful moments even more meaningful.

Contents

Certificate of Approval **ii**

Abstract **ii**

Author's Declaration **iv**

Acknowledgment **v**

1 Introduction **1**

1.1 The Hyperspectral Unmixing Problem 2

1.2 Challenges: Mixing, Variability, and Limited Ground Truth 3

1.3 Motivation for Deep Generative Models and Transformers 5

1.4 Problem Statement and Research Objectives 6

1.5 Contributions of This Thesis 7

1.6 Thesis Organization 8

2 Background and Related Work **10**

2.1 Classical and Physics-Based Hyperspectral Unmixing Methods 11

2.1.1 Physics-Based Reflectance Models 11

2.1.2 NMF-Based Methods 13

2.2 Deep Learning for Hyperspectral Unmixing 15

2.2.1 Autoencoder and CNN-Based Approaches 15

2.2.2 Latent Dirichlet VAEs for Hyperspectral Unmixing 16
2.2.3 Transformer-Based Models for Hyperspectral Unmixing 17

3 Problem Definition and Methodologies **20**
3.1 Baseline Transformer-VAE for Hyperspectral Unmixing 21
3.2 Enhanced Transformer-VAE 23
3.3 Latent Dirichlet Transformer VAE with Bundled Endmembers (LDVAE-T) 25

4 Experimental Setup **30**
4.1 Datasets 30
4.1.1 Samson Dataset 32
4.1.2 Jasper Ridge Dataset 32
4.1.3 HYDICE Urban Dataset 32
4.2 Endmember Bundle Initialization 32
4.3 Evaluation Metrics 43

5 Experimental Results and Analysis **47**
5.1 Endmember Extraction Results 47
5.2 Abundance Estimation Results 51
5.3 Ablation Studies 57

6 Conclusion and Future Work **59**
6.1 Limitations and Future Research Directions 60

Bibliography **61**

List of Tables

2.1 Related works overview of existing and referenced works. 12

4.1 SAD between mean of all GT pure pixels in the dataset and provided GT spectra for HYDICE Urban. 42

4.2 SAD between mean of all GT pure pixels in the dataset and provided GT spectra for Jasper Ridge. 42

4.3 SAD between mean of all GT pure pixels in the dataset and provided GT spectra for samson 42

5.1 SAD Scores for Endmember Extraction on Samson Dataset 49

5.2 SAD Scores for Endmember Extraction on Jasper Ridge Dataset 49

5.3 SAD Scores for Endmember Extraction on HYDICE Urban Dataset. *SSF-Net used a five-endmember version of HYDICE Urban for their experiments. 51

5.4 RMSE Scores for Abundance Estimation on Samson Dataset 52

5.5 RMSE Scores for Abundance Estimation on Jasper Ridge Dataset 52

5.6 RMSE Scores for Abundance Estimation on HYDICE Urban Dataset. *SSF-Net used a five-endmember version of HYDICE Urban for their experiments. 52

5.7 Ablation study results for patch size 3/5/7, number of attention heads 4/8/16, and segment size 20/24/32. Note, for best results, number of heads should divide segment size. All run for 500 epochs 58

5.8 Abundance estimation (RMSE) and endmember extraction (SAD) average across our three models. 58

List of Figures

3.1 Overview. Transformer based encoder takes a hyperspectral image patch $\mathcal{P}(\mathbf{x})$ and construct $\mathbf{x}_{\text{latent}}$, $\hat{\boldsymbol{\alpha}}$ and abundances $\hat{\mathbf{z}}$. The MLP decoder takes $\hat{\mathbf{z}}$ and reconstructs pixel $\mathbf{x}$. 21

3.2 Overview. Transformer based encoder takes a hyperspectral image patch $\mathcal{P}(\mathbf{x})$ and construct $\mathbf{x}_{\text{latent}}$, $\hat{\boldsymbol{\alpha}}$ and abundances $\hat{\mathbf{z}}$. The transformer decoder takes **Memory** and $\hat{\mathbf{z}}$ and reconstructs pixel $\mathbf{x}$. 23

3.3 Overview. Transformer based encoder takes a hyperspectral image patch $\mathcal{P}(\mathbf{x})$ and construct $\mathbf{x}_{\text{latent}}$, $\hat{\boldsymbol{\alpha}}$ and abundances $\hat{\mathbf{z}}$. The decoder takes $\mathbf{x}_{\text{latent}}$ and $\hat{\mathbf{z}}$ and reconstructs pixel $\mathbf{x}$. The decoder also constructs endmembers $\hat{\mathbf{e}}_k$. 25

4.1 An RGB approximation of the Samson dataset 33

4.2 Visualization of the train/test split used for the experiments on Samson . 34

4.3 An RGB approximation of the Jasper Ridge dataset 35

4.4 Visualization of the train/test split used for the experiments on Jasper Ridge 36

4.5 An RGB approximation of the HYDICE Urban dataset 37

4.6 Visualization of the train/test split used for the experiments on HYDICE Urban 38

4.7 Normalized spectra of the pure pixels and the corresponding distribution mean for each endmember in the Samson dataset. In each subplot, the individual pure-pixel spectra are shown in grey and the distribution mean is shown in colour. The value n denotes the number of identified pure pixels. 39

4.8 Normalized spectra of the pure pixels and the corresponding distribution mean for each endmember in the Jasper Ridge dataset. In each subplot, the individual pure-pixel spectra are shown in grey and the distribution mean is shown in colour. The value n denotes the number of identified pure pixels. 39

4.9 Normalized spectra of the pure pixels and the corresponding distribution mean for each endmember in the HYDICE Urban dataset. In each subplot, the individual pure-pixel spectra are shown in grey and the distribution mean is shown in colour. The value n denotes the number of identified pure pixels. 40

4.10 Locations of the identified pure pixels for each endmember in the Samson dataset. The corresponding pure pixels are highlighted over a grayscale approximation of the hyperspectral scene. 44

4.11 Locations of the identified pure pixels for each endmember in the Jasper Ridge dataset. The corresponding pure pixels are highlighted over a grayscale approximation of the hyperspectral scene. 45

4.12 Locations of the identified pure pixels for each endmember in the HYDICE Urban dataset. The corresponding pure pixels are highlighted over a grayscale approximation of the hyperspectral scene. 46

5.1 Comparison between the extracted endmember spectra and the corresponding ground-truth mean spectra for the Samson dataset. 48

5.2 Comparison between the extracted endmember spectra and the corresponding ground-truth mean spectra for the Jasper Ridge dataset. 49

5.3 Comparison between the extracted endmember spectra and the corresponding ground-truth mean spectra for the HYDICE Urban dataset. . . 50
5.4 Ground-truth and predicted abundance maps for each endmember in the Samson dataset. 53
5.5 Ground-truth and predicted abundance maps for each endmember in the Jasper Ridge dataset. 54
5.6 Ground-truth and predicted abundance maps for each endmember in the HYDICE Urban dataset. 55

Chapter 1

Introduction

Hyperspectral imaging (HSI) in a remote sensing setting is an imaging technique that captures detailed spectral information across dozens or even hundreds of narrow wavelength bands. Unlike traditional RGB or multispectral imaging, which measure only a small number of broad bands, hyperspectral sensors record fine reflectance profiles that enable precise material discrimination. Each pixel in a hyperspectral image corresponds to a high-dimensional spectrum whose shape can be used to infer the physical and chemical properties of the materials present within the observed scene.

HSI has become an essential tool in a wide range of scientific and industrial applications due to its ability to capture detailed spectral information for material identification [1]. In environmental monitoring, hyperspectral data is used to assess vegetation health, monitor water and air quality, detect invasive species, and track land-use changes [29]. In geology and mineral exploration, spectral signatures enable the identification of soil types, mineral deposits, and surface composition [27]. Precision agriculture, infrastructure inspection, and military surveillance similarly rely on hyperspectral sensors for their ability to detect subtle spectral variations and identify materials that cannot be distinguished using conventional imaging systems [25, 26].

Despite these advantages, hyperspectral imaging is constrained by the inherent trade-

off between spectral and spatial resolution. Airborne and satellite-based systems often capture high-quality spectral information but at spatial resolutions where each pixel covers a relatively large area on the ground. This leads to **mixed pixels**, where multiple distinct materials contribute to the observed spectrum of each individual pixel. As a result, the reflectance vector at a given location rarely corresponds to a single pure material, but rather to a mixture of several, weighted by their relative proportions within the pixel.

Understanding and resolving these mixtures is essential to many downstream tasks in remote sensing. However, accurately decomposing a mixed spectrum into its component materials is challenging due to noise, nonlinear interactions, illumination changes, and inherent spectral variability within each material class. These challenges motivate the need for advanced modeling techniques capable of making effective use of the rich spectral content offered by hyperspectral sensors while accounting for physical constraints and real-world variability.

1.1 The Hyperspectral Unmixing Problem

Hyperspectral sensors capture hundreds of narrow, contiguous spectral bands, allowing each pixel to be represented by a high-dimensional reflectance vector. In an ideal scenario, this vector corresponds to a single material whose spectrum uniquely identifies it. However, hyperspectral imagery acquired from airborne or satellite platforms typically suffers from limited spatial resolution; a single pixel frequently covers multiple materials such as vegetation, soil, water, or man-made structures. As a result, the measured spectrum is a mixture of the spectral signatures of all materials present within the pixel's footprint.

Hyperspectral unmixing (HU) is the process of decomposing these mixed spectra into two components:

1. **Endmembers**: the spectral signatures of the pure materials present in the scene.

2. **Abundances**: the proportion of each endmember contributing to the observed pixel.

Most unmixing formulations begin with the **linear mixing model** (LMM), which assumes that each pixel spectrum is a linear combination of endmember spectra. This corresponds to a physically interpretable model where materials contribute linearly and where abundances must satisfy two key constraints: non-negativity and sum-to-one. Despite its simplicity, the LMM has served as the foundation for unmixing research.

However, real-world conditions, such as illumination changes, sensor noise, and nonlinear interactions between materials introduce substantial deviation from the ideal linear model. These challenges make accurate endmember estimation difficult and lead to uncertainty in abundance prediction. Addressing these issues requires models capable of capturing spectral variability and spatial structure while adhering to the physical constraints intrinsic in unmixing. This requirement motivates the exploration of probabilistic deep learning architectures, which provide opportunities to incorporate both expressive modeling and principled constraints into the unmixing process.

1.2 Challenges: Mixing, Variability, and Limited Ground Truth

While hyperspectral imaging provides rich spectral information, several fundamental challenges complicate the task of accurately interpreting and decomposing the data. The most prominent of these challenges arise from the mixed-pixels, spectral variability within material classes, and the scarcity of reliable ground-truth annotations for both endmembers and abundances.

A first major challenge stems from **spectral mixing**, which occurs because the spatial

resolution of hyperspectral sensors is often insufficient to isolate pure materials within a single pixel. In many remote sensing scenarios, a pixel may cover vegetation mixed with soil, man-made structures, shadows, or water.

A second challenge is **spectral variability**, the phenomenon where the spectral signature of a single material class changes depending on illumination, atmospheric conditions, moisture content, sensor noise, or geometric factors such as viewing angle and surface orientation. Even within a single dataset, the same material can exhibit considerable spectral diversity (see Figures 4.9, 4.8, 4.7). Traditional unmixing approaches often assume a single fixed spectral signature per material, which is insufficient for representing this natural variability. Failing to account for variability leads to inaccurate endmember estimates and biases in abundance predictions.

A final major challenge involves the **limited availability of ground-truth data**. Unlike tasks such as classification or segmentation, hyperspectral unmixing rarely benefits from dense labels. Ground-truth abundances are expensive and impractical to obtain at scale, and ground-truth endmember signatures are often derived through laboratory measurements or estimation techniques rather than being directly observed. Many datasets provide a single reference spectrum for each class, which does not fully reflect real-world spectral variability. Even when "pure pixels" exist, they may account for only a small fraction of the data. This scarcity of supervision makes it difficult to train robust data-driven models without incorporating additional constraints, priors, or inductive biases.

Together, these factors, mixed pixels, spectral variability, and limited ground truth, highlight the need for unmixing frameworks that are both physically grounded and model-flexible. They motivate the exploration of generative models and transformer-based architectures that can encode global spectral–spatial relationships, represent uncertainty explicitly, and integrate priors that reflect the physical structure of the unmixing problem.

1.3 Motivation for Deep Generative Models and Transformers

Deep learning has emerged as a powerful tool for hyperspectral unmixing due to its ability to learn nonlinear spectral–spatial representations directly from data. Traditional approaches, whether based on geometric intuition, sparse coding, or matrix factorization, frequently struggle to capture complex reflectance interactions or represent the inherent variability within material classes. In contrast, neural architectures offer a flexible framework for modeling high-dimensional spectra and can incorporate inductive biases that reflect the physics of hyperspectral imaging.

Among deep generative models, Variational Autoencoders (VAEs) are especially appealing because they learn structured latent distributions instead of point estimates. By carefully selecting an appropriate latent distribution, such as the Dirichlet, VAEs can encode these constraints directly into their generative framework, producing physically meaningful abundance estimates while still benefiting from powerful representation learning.

Transformers have also shown strong performance in hyperspectral tasks [11, 8]. Their self-attention mechanisms allow them to capture long-range dependencies across both spectral channels and spatial neighborhoods, relationships that CNNs and MLPs often fail to model adequately. Since hyperspectral pixels can be spectrally similar yet spatially distant, and since spectral bands may exhibit nonlocal correlations, transformer-based architectures are well suited to extract meaningful structure from hyperspectral patches.

Together, these observations motivate the integration of transformers, variational autoencoders, and Dirichlet latent spaces, forming the foundation of the models developed throughout this thesis. The combination enables learned representations that respect physical constraints, incorporate global context, and adapt to spectral variability, addressing several limitations of existing hyperspectral unmixing methods.

1.4 Problem Statement and Research Objectives

Despite significant progress in hyperspectral unmixing, current models do not address the intragroup variability of endmembers referred to as the **bundled endmember** problem. To solver this problem, endmembers can be instead represented as a distribution with both mean and covariance rather then a fixed spectrum.

The core problem addressed in this thesis is therefore:

To develop deep generative models that accurately estimate both endmember distributions and abundance vectors while incorporating global spectral–spatial structure and respecting the physical constraints of hyperspectral unmixing.

To address this overarching goal, the thesis pursues the following specific research objectives:

1. Design a baseline Transformer-VAE model for hyperspectral unmixing that leverages self-attention to capture spectral–spatial dependencies within local patches and a Dirichlet-enforced latent layer to enforce physical mixing constraints.

2. Develop an improved Transformer-VAE model by replacing the standard MLP decoder with a carefully constructed transformer based decoder to more accurately reconstruct full hyperspectral image patches and extract endmember spectral information.

3. Introduce bundled endmembers to the Transformer-VAE model via manufactured ground-truth Gaussian distributions created using ground-truth pure pixels within the scene.

4. Evaluate all proposed methods on benchmark datasets and compare them against classical and modern baselines in abundance estimation and endmember extraction.

5. Analyze the progression between the three models to understand how architectural and probabilistic components contribute to improved unmixing performance.

These objectives guide the methodology and structure of the thesis, culminating in the development of the LDVAE-T model and a comprehensive experimental evaluation.

1.5 Contributions of This Thesis

This thesis makes several contributions to the field of hyperspectral unmixing and deep generative modeling:

1. **A baseline Dirichlet Transformer-VAE architecture for hyperspectral unmixing.** This model demonstrates that purely attention-based encoders can be effectively applied to hyperspectral patches, capturing long-range spectral interactions that conventional CNN-based models overlook.

2. **An enhanced Transformer-VAE with improved decoder structure.** Through refined tokenization strategies and architectural adjustments, this model addresses minor limitations of the baseline and provides a stronger foundation for abundance estimation.

3. **A novel Latent Dirichlet Transformer VAE (LDVAE-T) with bundled endmembers.** This method introduces Dirichlet-distributed abundance latents and a bundled endmember decoder that models each material as a distribution, rather than a single fixed spectrum. The decoder predicts both the mean and structured covariance for each endmember, enabling the model to represent spectral variability more naturally.

4. **Extensive experimental validation across three benchmark datasets.** The thesis presents detailed evaluations on the Samson, Jasper Ridge, and HYDICE

Urban datasets, providing comparisons with classical NMF-based methods, CNN-based autoencoders, Dirichlet-based VAEs, and transformer-based unmixing models.

5. **A progression analysis across the three proposed models.** The thesis analyzes how each architectural improvement: tokenization, latent modeling, endmember distribution modeling, contributes to the final performance gains, providing insight for future work in transformer-based unmixing models.

1.6 Thesis Organization

This thesis is organized into nine chapters, reflecting the research progression from foundational concepts to the development and evaluation of the proposed model.

- **Chapter 1** introduces hyperspectral imaging, outlines the key challenges associated with spectral mixing and variability, and presents the motivation, problem statement, and main contributions of this work.
- **Chapter 2** provides background on hyperspectral imaging, classical unmixing techniques, deep learning-based approaches, and the theoretical foundations behind variational autoencoders and transformer architectures.
- **Chapter 3** presents the Three models, including their respective architectures, performance characteristics and design rationale.
- **Chapter 4** outlines the datasets, preprocessing strategies, and experimental setup used to evaluate all models, including details on patch extraction, initialization of endmember bundles, and evaluation metrics.
- **Chapter 5** presents quantitative and qualitative results across the Samson, Jasper Ridge, and HYDICE Urban datasets. This includes abundance estimation, end-

member extraction, ablation studies, and comparisons with state-of-the-art methods.

- **Chapter 6** concludes the thesis by summarizing the main findings, discussing the implications of the proposed models, and outlining potential directions for future research.

Chapter 2

Background and Related Work

Hyperspectral imaging captures reflectance across hundreds of narrow, contiguous spectral bands, enabling fine-grained material discrimination that surpasses the capabilities of multispectral sensors [2, 12]. Each material exhibits a distinct spectral signature; however, due to limited spatial resolution, most pixels represent mixtures of multiple materials. The classical Linear Mixing Model (LMM) assumes that each pixel spectrum is a straightforward combination of endmember signatures weighted by abundance fractions that satisfy non-negativity and sum-to-one constraints [17]. Although the LMM is widely used, real-world reflectance often involves nonlinear interactions such as multiple scattering, intimate mixing, and illumination-dependent effects, making unmixing challenging [28]. Additionally, spectral signatures of the same material vary due to illumination geometry, atmospheric effects, and intrinsic material differences, a phenomenon known as endmember variability [34]. These challenges motivate the development of complex unmixing models capable of capturing both spatial–spectral structure and intrinsic spectral variability.

Under the LMM assumption, hyperspectral unmixing methods can be broadly categorised into physics-based [13, 30, 15, 7], classical matrix factorisation-based [35, 32, 14, 36], and learning-based approaches [6, 24, 3, 16, 8, 11, 22, 10, 5, 33, 31, 9]. Learning-

based approaches can be further categorised as MLP autoencoder-based, CNN-based, Transformer-based, and Latent Dirichlet Variations of the aforementioned. Table 2.1 provides a summary of all works discussed in this section.

2.1 Classical and Physics-Based Hyperspectral Unmixing Methods

Classical hyperspectral unmixing methods provide the foundation for many modern approaches by formalizing the relationship between observed spectra, endmember signatures, and material abundances. These methods typically rely on either physically motivated models of light–material interaction or matrix factorization frameworks that impose interpretable constraints on the unmixing process. Although they remain valuable due to their interpretability and strong theoretical grounding, their assumptions often limit their ability to capture the nonlinear, variable, and spatially complex nature of real hyperspectral scenes.

2.1.1 Physics-Based Reflectance Models

Physics-based methods for hyperspectral unmixing rely on models that describe how light interacts with materials. The vast majority of these approaches derive directly from radiative transfer theory and optical scattering physics.

Hapke's classical BRDF model [13] provides the foundational description of light interacting with particulate surfaces, but is difficult to invert and requires detailed material parameters. Hyperspectal unmixing methods, such as the mineralogical unmixing approach of Sun and Lucey [30], apply Hapke's radiative transfer inversion to estimate mineral abundances and compositional parameters, though at significant computational cost and with limited applicability to complex real-world scenes. Drumetz et al. [7] bridge physics-based and data-driven deep learning unmixing approaches by showing that the

Table 2.1: Related works overview of existing and referenced works.

Method	Supervision	Category	Venue	Year	Benchmarks Used
Hapke BRDF [13]	Unsupervised	Physics-based	Journal of Geophysical Research: Solid Earth	1981	N/A
Sun–Lucey RTM [30]	Unsupervised	Physics-based	Journal of Geophysical Research: Planets	2021	LSSC lunar soil spectra; Lunar Prospector global maps
ELMM [7]	Unsupervised	Physics-based	IEEE Geoscience and Remote Sensing Letters	2019	-
DDM [15]	Unsupervised	Physics-based	ECCV	2020	Feely, Gulfport, Gulfport synthetic, TES Martian
SGSNMF [32]	Unsupervised	NNMF	IEEE Transactions on Geoscience and Remote Sensing	2017	HYDICE Urban, Cuprite, UAV-BORNE, Synthetic Data
MRSNMF [21]	Unsupervised	NNMF	IEEE Transactions on Geoscience and Remote Sensing	2013	Cuprite, Synthetic
TV-RSNMF [14]	Unsupervised	NNMF	IEEE Transactions on Geoscience and Remote Sensing	2017	HYDICE Urban, Cuprite, Synthetic Data
NMF-QMV [36]	Unsupervised	NNMF	IEEE Geoscience and Remote Sensing Letters	2022	Samson, Jasper Ridge, Cuprite, Synthetic Data
DeepGUn [3]	Unsupervised	MLP autoencoder-based	IEEE Transactions on Computational Imaging	2019	Samson, Jasper Ridge, Houston, Synthetic data
CNNAEU [24]	Unsupervised	CNN-based	IEEE Transactions on Geoscience and Remote Sensing	2021	Samson, Jasper Ridge, Houston, Apex
LDVAE [22]	Supervised	Latent Dirichlet MLP	IEEE Transactions on Geoscience and Remote Sensing	2024	Samson, HYDICE Urban, Cuprite, OnTech-HSI-Syn-21
SpACNN-LDVAE [5]	Supervised	Latent Dirichlet CNN	IGARSS	2024	Samson, HYDICE Urban, Cuprite, OnTech-HSI-Syn-21
DeepTrans-HsU [11]	Unsupervised	Transformer-based	IEEE Transactions on Geoscience and Remote Sensing	2022	Samson, Apex, Washington DC Mall
UnDAT [8]	Unsupervised	Transformer-based	IEEE Transactions on Geoscience and Remote Sensing	2023	Samson, Jasper Ridge, Apex, Synthetic Data
SSF-Net [31]	Unsupervised	Transformer-based	IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing	2024	Samson, Jasper Ridge, HYDICE Urban, Synthetic Data
SCEELA [10]	Supervised	Transformer-based	ISPRS Open Journal of Photogrammetry and Remote Sensing	2025	Samson, Jasper Ridge, HYDICE Urban
SSLT-Net [9]	Unsupervised	Transformer-based	IEEE Geoscience and Remote Sensing Letters	2025	Samson, Jasper Ridge, Apex

Extended Linear Mixing Model arises as a first-order approximation of Hapke's theory, thereby grounding illumination-induced spectral variability in physical principles. Other approaches, such as Janiczek et al.'s differentiable dispersion model [15], reformulate radiative-transfer equations into fully differentiable pipelines that can be optimized with gradient descent, providing a bridge between physics-based models and deep learning techniques.

Despite these advances, physics-based methods remain constrained by the need for material-specific physical parameters and by limited scalability, making them complementary to but distinct from modern data-driven unmixing systems.

2.1.2 NMF-Based Methods

Non-Negative Matrix Factorization (NMF) remains one of the most widely used data-driven approaches for hyperspectral unmixing. NMF decomposes a hyperspectral image into a set of endmember spectra and corresponding abundance maps under the constraint that both matrices are non-negative. This aligns naturally with the physics of reflectance and material proportions, making NMF an effective baseline for low-label or unsupervised unmixing tasks. However, the basic NMF formulation relies solely on spectral information and assumes that each endmember can be represented by a single fixed signature. In practice, hyperspectral data exhibit significant spectral variability due to illumination changes, sensor noise, and material heterogeneity, limiting the effectiveness of classical NMF.

To mitigate these shortcomings, numerous extensions to NMF have been proposed. One major direction involves incorporating *spatial information*. Spectral–spatial NMF models such as Zhang et al. [35] integrate neighbourhood structure into the factorization process, encouraging abundance maps that vary smoothly across homogeneous regions while preserving local spatial continuity. This improves robustness to noise and helps maintain spatial coherence in the estimated abundances.

A related line of work focuses on *group-based spatial structure*. Wang et al. [32] introduce Spatial Group Sparsity Regularized NMF (SGSNMF), which groups pixels into superpixels or local clusters and enforces sparsity within each group. This allows the model to preserve material boundaries and capture localized spatial structure more effectively than pixel-wise NMF.

Other extensions incorporate *geometric structure* through manifold regularization. Lu et al. [21] propose Manifold Regularized Sparse NMF (MRSNMF), which assumes that hyperspectral data lie on a lower-dimensional manifold and uses Laplacian regularization to preserve local geometric relationships. This helps discriminate between materials with similar spectra by respecting the intrinsic geometry of the dataset.

Total variation regularization has also been used to enforce spatial smoothness. He et al. [14] propose Total Variation Regularized Reweighted Sparse NMF (TV-RSNMF), in which total variation encourages piecewise-smooth abundance maps with sharp transitions at material boundaries. The iterative reweighting scheme improves convergence stability and reduces noise sensitivity, particularly in complex or high-variance scenes.

Finally, Zhao et al. [36] introduce a hybrid approach that combines *handcrafted priors*, such as sparsity and smoothness, with *learned priors* obtained from data. This method bridges classical NMF with learned representations, allowing the model to capture richer spectral–spatial characteristics and improving generalization across different datasets and imaging conditions.

Despite these advances, NMF-based approaches still have several limitations. Most variants still assume a *linear* mixing model and cannot fully capture nonlinear light interactions typical in vegetation, intimate mineral mixtures, or environments with strong scattering. Additionally, NMF methods rely on *fixed endmember signatures* and do not model distributional spectral variability, wavelength-dependent effects, or probabilistic uncertainty. As a result, although NMF and its variants offer computational efficiency and interpretability, they remain limited in their ability to represent the full complexity

of real-world hyperspectral mixtures.

Overall, NMF-based approaches are more flexible and practical than physics-based radiative transfer models, yet less expressive than modern deep learning methods.

2.2 Deep Learning for Hyperspectral Unmixing

Deep learning methods extend hyperspectral unmixing beyond fixed physical or linear assumptions by learning spectral–spatial representations directly from data. These approaches are particularly useful for modelling nonlinear mixing effects, spatial context, and complex variability that are difficult to capture with classical methods. The following subsections review the progression from autoencoder- and CNN-based models to probabilistic Dirichlet VAE frameworks and transformer-based architectures, highlighting how each class of methods addresses key limitations in earlier unmixing approaches.

2.2.1 Autoencoder and CNN-Based Approaches

Deep learning approaches have become an important class of methods for hyperspectral unmixing due to their ability to learn nonlinear spectral–spatial relationships directly from data. Unlike classical NMF or geometric techniques, autoencoders (AEs) and convolutional neural networks (CNNs) can capture complex feature representations through learned transformations, enabling improved robustness to noise, spectral variability, and nonlinear mixing effects.

One of the early deep learning approaches for unmixing is DeepGUn, which combines a deep autoencoder architecture with vertex component analysis [3]. The model learns a latent representation of spectral mixtures while using a geometric constraint to estimate endmembers. This hybrid strategy improves over classical methods by leveraging nonlinear feature extraction.

Autoencoder-based hyperspectral unmixing was further advanced by the CNNAEU

model, which introduced convolutional layers into the encoder and decoder [24]. CNNAEU was the first method to explicitly exploit local spatial structure through CNNs, recognizing that neighbouring pixels in hyperspectral images often share similar material compositions. By replacing fully connected layers with convolutional ones, CNNAEU improves abundance estimation accuracy and reduces noise sensitivity.

In general, autoencoder and CNN-based methods represent an important step toward more expressive unmixing models. They can learn nonlinear mappings, capture spatial structure, and via variational formulations incorporate physically meaningful constraints.

2.2.2 Latent Dirichlet VAEs for Hyperspectral Unmixing

Variational Autoencoders or VAEs have emerged as a very promising direction for hyperspectral unmixing. VAEs employ distributional sampling in the latent representation of models, this allows the decoder to see more than just the examples present in training data [18]. In other domains, VAEs employ normal distributions to sample from; however, for the purposes of hyperspectral unmixing, normal distributions do not accurately represent the real-world requirements of sum-to-one and non-negativity.

The Latent Dirichlet Variational Autoencoder (LDVAE) introduced a major conceptual shift by incorporating a Dirichlet latent distribution into the unmixing pipeline [22]. The Dirichlet prior naturally enforces the real-world sum-to-one and non-negativity constraints on abundances, providing a physically meaningful latent space and enabling the use of probabilistic modeling of mixture proportions [4]. LDVAE showed that learned priors and variational inference could outperform purely deterministic autoencoders. Nonetheless, its encoder and decoder are based on multilayer perceptrons (MLPs), limiting its ability to capture spatial information and handling only global spectral relationships.

To address this limitation, SpACNN-LDVAE replaced the MLP encoder with a spatially aware CNN encoder while retaining the Dirichlet variational framework [5]. The

CNN encoder better preserves spatial coherence and extracts more expressive spectral–spatial representations. As a result, SpACNN-LDVAE improves both abundance estimation and endmember extraction compared to both LDVAE and CNNAEU.

Overall, Dirichlet-based VAEs represent an important evolution in hyperspectral unmixing by introducing a latent space that is physically aligned with mixture constraints while retaining the flexibility of deep generative modeling. These methods demonstrate that probabilistic abundance modeling can outperform deterministic encoders and can more faithfully capture the structure of hyperspectral mixtures. However, existing LDVAE architectures are still limited by their reliance on fixed endmember representations and their inability to fully account for spectral variability or long-range spectral–spatial dependencies. These remaining gaps motivate the need for more expressive models such as transformer based architectures with Dirichlet latent priors that can jointly capture spatial structure, global spectral relationships, and distributional endmember variability.

2.2.3 Transformer-Based Models for Hyperspectral Unmixing

Transformer architectures have recently emerged as a powerful alternative to CNN- and MLP-based unmixing models due to their ability to capture long-range dependencies and global spectral–spatial relationships through self-attention [6]. While CNNs effectively model local interactions, they struggle to incorporate distant spectral correlations or scene-level context, which are often essential for accurate abundance estimation and endmember extraction. Transformers address this limitation by processing hyperspectral pixels or patches as sequences of tokens and applying multi-head attention to learn contextual relationships across both spectral bands and spatial neighborhoods.

One of the earliest transformer-based approaches for hyperspectral unmixing is DeepTrans-HsU, which demonstrated that replacing convolutional or fully connected encoders with a Vision Transformer (ViT) backbone significantly improves abundance prediction [11]. The self-attention mechanism allows the model to learn long-range spectral interactions

that CNN-based models fail to capture. DeepTrans-HsU showed that even relatively simple transformer blocks can outperform deeper CNN architectures, establishing transformers as a promising direction for unmixing.

Building on this idea, UnDAT introduced a more sophisticated transformer encoder–decoder framework that incorporates spectral and spatial clustering modules [8]. By grouping pixels and bands into contextually coherent clusters, UnDAT leverages the attention mechanism more efficiently and improves robustness to spectral noise. However, the method is computationally heavy and relies on several auxiliary modules, making it less lightweight than other transformer-based models.

Another notable contribution is SSF-Net, which incorporates a spatial–spectral fusion module within a transformer-like architecture [31]. SSF-Net emphasizes the joint modeling of local spatial structure and global spectral relationships and demonstrates strong performance in abundance estimation.

Two recent studies further highlight the growing maturity of transformer-based unmixing. SCEELA (Spatial Context and Endmember Ensemble Learning with Attention) incorporates a multi-head attention mechanism into both endmember extraction and spatial contextualization [10]. It introduces three components: a Signature Predictor that uses an ensemble of endmember extraction algorithms as initialization; a Pixel Contextualizer that refines pixel embeddings; and an Abundance Predictor that outputs per-pixel proportions. By leveraging attention for both ensemble learning and spatial context modeling, SCEELA achieves strong performance across multiple real and synthetic datasets.

Similarly, SSLT-Net (Spatial–Spectral Linear Transformer) proposes a lightweight transformer unmixing framework that emphasizes linear attention mechanisms to reduce computational complexity [9]. SSLT-Net integrates both spatial and spectral cues while maintaining an efficient architecture suitable for larger datasets or real-time applications. While effective, SSLT-Net remains limited by the linearity of its attention design.

Despite the impressive progress of transformer-based unmixing models, existing ap-

proaches share two notable gaps. First, most methods assume fixed or externally initialized endmembers, making them less effective at modeling intrinsic spectral variability across a scene. Second, although transformers excel at spectral–spatial feature extraction, they do not inherently enforce the physical constraints of hyperspectral unmixing—such as the sum-to-one and non-negativity properties of abundances. These limitations motivate hybrid architectures that combine attention-based global modeling with probabilistic latent spaces tailored to the physics of spectral mixing, such as the Dirichlet-based transformer VAE framework introduced in this thesis.

Chapter 3

Problem Definition and Methodologies

In hyperspectral imaging, under the linear mixing model, each pixel $\mathbf{x} \in \mathbb{R}^C$ can be represented,

$$\mathbf{x} = \mathbf{E}\mathbf{z} + \mathbf{n}, \tag{3.1}$$

where C refers to the number of channels, $\mathbf{E} \in \mathbb{R}^{C \times K}$ is the endmember matrix that represents the spectra of K endmembers column-wise, $\mathbf{z} \in \mathbb{R}^{K \times 1}$ is the abundance vector representing the proportion of each endmember in pixel $\mathbf{x}$, and $\mathbf{n}$ is additive noise.

Here, the abundance vector $\mathbf{z}$ represents the mixing proportions of each endmember in the given pixel. Since $\mathbf{z}$ is a mixture, it must follow two constraints:

$$\forall_{k \in K} z_k \geq 0, \textbf{(non-negative)} \tag{3.2}$$

$$\sum_{k \in K} z_k = 1. \textbf{(sum-to-one)} \tag{3.3}$$

Now we can define unmixing as computing abundances $\mathbf{z}$ and endmembers matrix $\mathbf{E}$ given a hyperspectral pixel $\mathbf{x}$.

In the following sections we outline our three model progression, each with a trans-

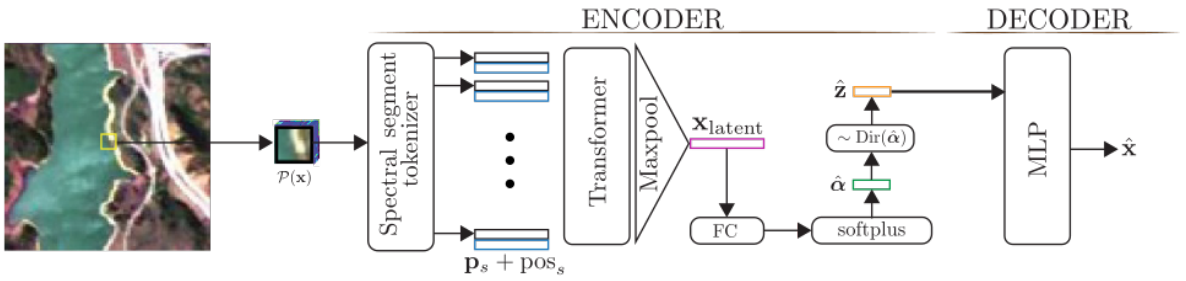


Figure 3.1: Overview. Transformer based encoder takes a hyperspectral image patch $\mathcal{P}(\mathbf{x})$ and construct $\mathbf{x}_{\text{latent}}$, $\hat{\boldsymbol{\alpha}}$ and abundances $\hat{\mathbf{z}}$. The MLP decoder takes $\hat{\mathbf{z}}$ and reconstructs pixel $\mathbf{x}$.

former encoder and Dirichlet enforced latent representation.

3.1 Baseline Transformer-VAE for Hyperspectral Unmixing

Our first approach was a straightforward advancement of LDVAE [22] and SpACNN-LDVAE [5] by replacing the encoder with a Vision Transformer architecture. The goal of this baseline model was to investigate whether self-attention mechanisms could improve spectral–spatial feature extraction compared to the MLP and CNN encoders used in prior work, while retaining the variational autoencoder framework. Our model overview is showcased in Figure 3.1

Similar to LDVAE, the model processes a local hyperspectral patch $\mathcal{P}(\mathbf{x})$ centered on pixel $\mathbf{x}$. The patch is first converted into a sequence of tokens $\mathbf{p}_s \in \mathbb{R}^d$ for $s \in \{1, \cdots, S\}$ using a simple linear embedding of each pixel spectrum. Learned positional encodings are added to preserve spatial relationships within the patch,

$$\tilde{\mathbf{p}}_s = \mathbf{p}_s + \text{pos}_s.$$

The sequence $\{\tilde{\mathbf{p}}_s\}_{s=1}^{S}$ is then passed through a transformer encoder consisting of L layers of multi-head self-attention followed by position-wise feed-forward networks.

Each layer enables the model to capture dependencies across both spatial locations and spectral channels within the patch. The encoder outputs $\{\mathbf{h}_s\}_{s=1}^{S}$ are aggregated using element-wise max pooling to form the latent representation for pixel $\mathbf{x}$,

$$\mathbf{x}_{\text{latent}} = \max_{s\in\{1,\cdots,S\}} \mathbf{h}_s.$$

In this baseline formulation, the latent vector parameterizes a Dirichlet distribution over abundances using a softplus transformation,

$$\boldsymbol{\alpha} = \text{softplus}(\mathbf{W}\mathbf{x}_{\text{latent}} + \mathbf{b}) + \epsilon\mathbf{1},$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters and $\epsilon > 0$ ensures strictly positive concentration values. Sampling from $\text{Dir}(\boldsymbol{\alpha})$ yields the abundance vector $\mathbf{z}$ that satisfies the non-negativity and sum-to-one constraints required for hyperspectral unmixing.

The decoder follows the structure of LDVAE and consists of a multilayer perceptron that maps the sampled abundance vector $\mathbf{z}$ to the reconstructed spectrum $\hat{\mathbf{x}}$,

$$\hat{\mathbf{x}} = f_\phi(\mathbf{z}),$$

where $f_\phi(\cdot)$ denotes the decoder network parameterized by ϕ. In this formulation the endmember matrix is implicitly represented within the decoder weights, as in other autoencoder-based unmixing approaches.

Training is performed by optimizing the loss function,

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{recon}},$$

the evidence lower bound (ELBO) loss is calculated as follows,

$$\mathcal{L}_{\text{ELBO}} = \mathbf{E}_{q_\theta(\mathbf{z}|\mathbf{x})}[\log p_\phi(\mathbf{x}|\mathbf{z})] - \text{KL}(q_\theta(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})),$$

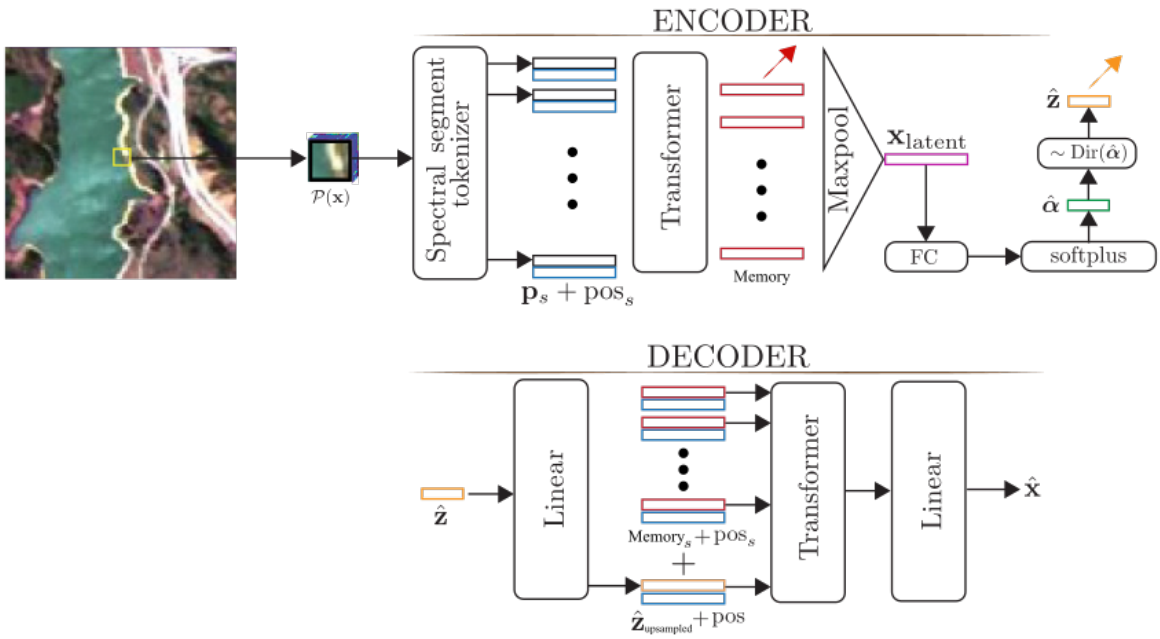


Figure 3.2: Overview. Transformer based encoder takes a hyperspectral image patch $\mathcal{P}(\mathbf{x})$ and construct $\mathbf{x}_{\text{latent}}$, $\hat{\boldsymbol{\alpha}}$ and abundances $\hat{\mathbf{z}}$. The transformer decoder takes **Memory** and $\hat{\mathbf{z}}$ and reconstructs pixel $\mathbf{x}$.

where $q_\theta(\mathbf{z}|\mathbf{x}) = \text{Dir}(\boldsymbol{\alpha})$ is the encoder distribution and $p(\mathbf{z}) = \text{Dir}(\boldsymbol{\alpha}^{\text{prior}})$ is the Dirichlet prior. The reconstruction term is implemented using mean squared error,

$$\mathcal{L}_{\text{recon}} = MSE(\hat{\mathbf{x}}, \mathbf{x}).$$

3.2 Enhanced Transformer-VAE

To improve on the baseline Transformer-VAE, we replace the MLP decoder with a transformer decoder. This change allows the reconstruction process to use spatial–spectral features extracted from the input patch, rather than reconstructing the pixel spectrum from the abundance vector alone. The model architecture is visualized in Figure 3.2.

In this model we keep the same data pre-prepossessing steps as in the previous, the first change lies at the end of the encoder where we instead have the encoder produce a

sequence of feature vectors

$$\{\mathbf{h}_s\}_{s=1}^{S},$$

which are stored as memory for the transformer decoder before the max-pooling operation. The pooled vector is still used to estimate the Dirichlet parameters and sample abundances $\mathbf{z}$ as in the baseline model.

Since the transformer decoder operates in the same embedding space as the encoder features, the abundance vector is first projected into the embedding dimension,

$$\mathbf{q} = \mathbf{W}_z\mathbf{z},$$

where $\mathbf{W}_z$ is a learnable projection matrix. The vector $\mathbf{q}$ acts as a query to the transformer decoder, while the encoder outputs $\{\mathbf{h}_s\}_{s=1}^{S}$ act as memory. The decoder reconstructs the pixel by attending to the encoder memory conditioned on the abundance vector,

$$\hat{\mathbf{x}} = g_\phi(\mathbf{z}, \{\mathbf{h}_s\}_{s=1}^{S}),$$

where $g_\phi(\cdot)$ represents the transformer decoder and output projection.

This modification allows the model to reconstruct spectra using both the abundance information and the spatial–spectral context of the input patch, which improves reconstruction accuracy compared to the baseline model that reconstructs using $\mathbf{z}$ alone.

The loss function remains similar to the baseline model, with the addition of a reconstruction weight,

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}},$$

where λ_{recon} controls the strength of the reconstruction term and is initialized to 0.5.

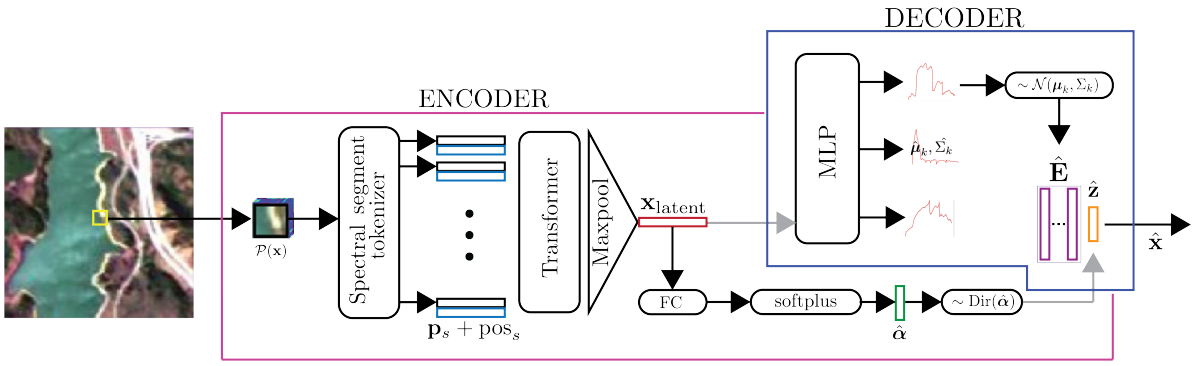


Figure 3.3: Overview. Transformer based encoder takes a hyperspectral image patch $\mathcal{P}(\mathbf{x})$ and construct $\mathbf{x}_{\text{latent}}$, $\hat{\boldsymbol{\alpha}}$ and abundances $\hat{\mathbf{z}}$. The decoder takes $\mathbf{x}_{\text{latent}}$ and $\hat{\mathbf{z}}$ and reconstructs pixel $\mathbf{x}$. The decoder also constructs endmembers $\hat{\mathbf{e}}_k$.

3.3 Latent Dirichlet Transformer VAE with Bundled Endmembers (LDVAE-T)

The proposed final *LDVAE-T* architecture consists of a transformer-based encoder that process a local hyperspectral image patch $\mathcal{P}(\mathbf{x})$ and predicts Dirichlet concentration parameters $\boldsymbol{\alpha}$ (abundance prior) for pixel $\mathbf{x}$. Additionally, it emits a latent code for $\mathbf{x}$ that the decoder uses to reconstruct the pixel's endmember spectra. Recall that endmember spectra are shared by all pixels. Under the assumption that each endmember follows a Gaussian distribution, the decoder maps the latent code to the Gaussian parameters mean (μ_k) and covariance (Σ_k) for each endmember k. Next, we sample endmembers and mix them explicitly. For each endmember k, draw $\mathbf{e}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ and draw abundances $\mathbf{z} \sim \text{Dir}(\boldsymbol{\alpha})$. The pixel is then reconstructed by the mixture $\hat{\mathbf{x}} = \phi(\sum_k z_k \mathbf{e}_k)$.

Figure 3.3 illustrates the model. The tokenizer in this model converts the local patch $\mathcal{P}(\mathbf{x}) \in \mathbb{R}^{P \times P \times C}$ into a sequence of tokens by splitting the patch both spatially and spectrally. Each pixel in the patch contains a spectrum with C bands. Rather than treating the entire spectrum as a single token, we divide the spectrum into segments of size C_{seg}. This results in

$$N_{\text{seg}} = \left\lfloor \frac{C}{C_{\text{seg}}} \right\rfloor$$

spectral segments per pixel.

Since the patch contains $P \times P$ pixels, the total number of tokens is

$$S = P^2 \cdot N_{\text{seg}}.$$

Each token therefore represents a small spectral segment from a specific spatial location. Each segment is projected into the embedding space using a linear layer to produce tokens

$$\mathbf{p}_s \in \mathbb{R}^d, \quad s \in \{1, \ldots, S\}.$$

Learned positional encodings are added to each token to preserve both spatial location and spectral segment position,

$$\tilde{\mathbf{p}}_s = \mathbf{p}_s + \text{pos}_s.$$

The sequence $\{\tilde{\mathbf{p}}_s\}_{s=1}^S$ is passed to a transformer encoder with $L = 4$ layers; each layer uses multi-head self-attention with $H = 16$ heads followed by a position-wise feed-forward network (with residual connections and layer normalization). Self-attention enables the model to capture long-range dependencies across both spectral bands and spatial locations within the patch, yielding a comprehensive representation of material mixtures. The encoder outputs $\{\mathbf{h}_s\}_{s=1}^S$ are max-pooled to construct the latent code for $\mathbf{x}$

$$\mathbf{x}_{\text{latent}} = \max_{s \in \{1, \cdots, S\}} \mathbf{h}_s \text{ (element-wise max)}.$$

The latent vector is passed through a softplus to produce the Dirichlet concentration parameters

$$\boldsymbol{\alpha} = \text{softplus}(\mathbf{W}\mathbf{x}_{\text{latent}} + \mathbf{b}) + \epsilon\mathbf{1}, \quad \boldsymbol{\alpha} \in \mathbb{R}^K_{>0}.$$

Here $\mathbf{W}$ and $\mathbf{b}$ are affine parameters for the soft-plus layer and small $\epsilon > 0$ ensures strictly positive entries. Sampling from $\text{Dir}(\boldsymbol{\alpha})$ yields an abundance vector $\mathbf{z}$ that satisfies non-

negativity and sum-to-one constraints.

The decoder comprises of two MLPs. First, MLP takes the latent vector $\mathbf{x}_{\text{latent}}$ and predicts endmembers means $\boldsymbol{\mu}_k$ and covariances Σ_k. The second MLP takes abundance vector $\mathbf{z}$ and sampled endmembers $\mathbf{e}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ and reconstruct pixel $\hat{\mathbf{x}}$. Similar to other approaches, we assume that the number of endmembers K is known *a priori*.

We employ the following losses during training. **Abundance loss** penalizes the divergence between predicted and ground truth abundances

$$\mathcal{L}_{\text{abundance}} = \text{MSE}(\hat{\mathbf{z}}, \mathbf{z})$$

where $\hat{\mathbf{z}}$ are the predicted abundances for pixel $\mathbf{x}$ and $\mathbf{z}$ are the ground-truth abundances.

Endmember bundles loss minimizes the KL divergence between the predicted endmember Gaussian and the ground-truth endmember bundle (estimated from *pure pixels*)

$$\mathcal{L}_{\text{endmembers}} = \sum_{k=1}^{K} \hat{\mathbf{z}}_k \text{KL}\left(\mathcal{N}(\hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k) \parallel \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)\right).$$

Here, the term $\hat{\mathbf{z}}_k$ is the predicted sampled abundance estimate for endmember k, obtained from the Dirichlet concentration parameters before drawing the abundance sample. This value weights the KL divergence so that endmembers predicted to be more present in the pixel contribute more strongly to the loss.

The distribution $\mathcal{N}(\hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)$ represents the predicted block-diagonal Gaussian distribution for endmember k, where $\hat{\boldsymbol{\mu}}_k$ is the predicted mean spectrum and $\hat{\Sigma}_k$ is the predicted covariance matrix. The distribution $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ represents the reference endmember bundle estimated from ground-truth pure pixels, where $\boldsymbol{\mu}_k$ is the reference mean spectrum and Σ_k is the reference covariance matrix for endmember k. This mixture-weighted regularization constrains the model most strongly for endmembers that are estimated to be present in the given pixel.

Lastly, the latent Dirichlet variational autoencoder optimizes

$$\mathcal{L}_{\text{ELBO}} = \mathbf{E}_{q_\theta(\mathbf{z}|\mathbf{x})}[\log p_\phi(\mathbf{x}|\mathbf{z})] - \text{KL}(q_\theta(\mathbf{z}|\mathbf{x}) \,\big\|\, p(\mathbf{z})).$$

with reconstruction term

$$\mathcal{L}_{\text{recon}} = MSE(\hat{\mathbf{x}}, \mathbf{x}),$$

and Dirichlet prior $p(\mathbf{z}) = \text{Dir}(\boldsymbol{\alpha}^{\text{prior}})$. When $q_\theta(\mathbf{z}|\mathbf{x}) = \text{Dir}(\hat{\boldsymbol{\alpha}})$

$$\begin{aligned}\text{KL}\left(\text{Dir}(\hat{\boldsymbol{\alpha}}) \,\big\|\, \text{Dir}(\boldsymbol{\alpha}^{\text{prior}})\right) = &\sum \log\Gamma(\alpha_k^{\text{prior}}) \\ &- \sum \log\Gamma(\hat{\alpha}_k) \\ &+ \sum(\hat{\alpha}_k - \alpha_k^{\text{prior}})\frac{d}{dx}\ln\Gamma(\hat{\alpha}_k).\end{aligned}$$

Here $\hat{\boldsymbol{\alpha}}$ is the concentration parameter of the estimated Dirichlet distribution and $\boldsymbol{\alpha}^{\text{prior}}$ is the concentration parameter of the Dirichlet prior. $\Gamma(\cdot)$ is the Gamma function. Thus the overall loss is

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \lambda_{\text{abundances}}\mathcal{L}_{\text{abundances}} + \lambda_{\text{endmembers}}\mathcal{L}_{\text{endmembers}}$$

with $\lambda_{\text{abundances}}$ and $\lambda_{\text{endmembers}}$ controlling the relative strengths of abundances and endmember terms. We set $\lambda_{\text{abundances}}$ to 1, and apply a heavy anneal to $\lambda_{\text{endmembers}}$ for experiments reported here. The annealing begins at 1×10^{-6} and progresses towards 1.0 over 80000 epochs (which is never reached). This is done to keep $\mathcal{L}_{\text{endmembers}}$ from significantly outweighing the other terms by a magnitude of 1×10^6 at the start of training.

The reconstruction loss is unsupervised, since it only requires the input hyperspectral pixel/patch. In contrast, the abundance loss is supervised because it requires ground-truth abundance vectors, z, for each training pixel. The endmember bundles loss is also supervised, or weakly supervised, because it requires reference endmember bundle

statistics estimated from ground-truth pure pixels. Thus, the full objective combines unsupervised reconstruction with supervised terms that use available abundance maps and pure-pixel-derived endmember information.

Chapter 4

Experimental Setup

This chapter describes the experimental framework used to evaluate the proposed LDVAE-T model. It introduces the hyperspectral datasets, the procedure used to construct non-overlapping training and testing splits, and the initialization of the endmember bundle distributions. This chapter also presents the model configurations, training settings, evaluation metrics, and baseline methods used to ensure that the reported comparisons are consistent and reproducible.

4.1 Datasets

We evaluate our model on three widely used hyperspectral unmixing benchmarks: Samson [19], Jasper Ridge [20], and HYDICE Urban [23]. All three datasets provide per-pixel ground-truth abundance maps and ground truth endmember spectra. To maximize available training data, we employ a special technique to select training and testing splits where patches do not overlap, while ensuring as many pixels as possible are used.

Recall, a hyperspectral image can be represented as a cube $X \in \mathbb{R}^{H \times W \times C}$, where H and W denote the spatial dimensions and C is the number of spectral bands. Training samples are extracted as square spatial patches of size $p \times p$. Each patch is indexed by the spatial coordinates of its center pixel (r, c). To ensure that patches do not overlap

between the training and testing sets, we first enumerate all valid center coordinates

$$\mathcal{S} = \{(r, c) \mid r \in [\lfloor p/2 \rfloor, H - \lfloor p/2 \rfloor),\ c \in [\lfloor p/2 \rfloor, W - \lfloor p/2 \rfloor)\}.$$

A patch centered at (r, c) therefore contains the spatial region

$$X[r - \lfloor p/2 \rfloor : r + \lfloor p/2 \rfloor,\ c - \lfloor p/2 \rfloor : c + \lfloor p/2 \rfloor, :].$$

To prevent overlap, we construct a subset of centers $\mathcal{S}_{\text{valid}}$ such that the spatial distance between any two selected centers is at least p pixels in both spatial dimensions. This guarantees that the spatial regions of any two patches are disjoint. In practice, this corresponds to sampling patch centers on a grid with stride p, producing the candidate index set

$$\mathcal{S}_{\text{valid}} = \{(r, c) \in \mathcal{S} \mid r \equiv r_0 \pmod{p},\ c \equiv c_0 \pmod{p}\},$$

where r_0 and c_0 ensure valid boundary placement.

The resulting index set $\mathcal{S}_{\text{valid}}$ contains all non-overlapping patches available in the image. These samples are then randomly partitioned into training and testing subsets using an 80/20 split:

$$|\mathcal{S}_{\text{train}}| = \lfloor 0.8\,|\mathcal{S}_{\text{valid}}| \rfloor, \qquad |\mathcal{S}_{\text{test}}| = |\mathcal{S}_{\text{valid}}| - |\mathcal{S}_{\text{train}}|.$$

The selected training indices $\mathcal{S}_{\text{train}}$ are used during model optimization, while the remaining indices $\mathcal{S}_{\text{test}}$ are reserved for evaluation. Because patches are constructed from $\mathcal{S}_{\text{valid}}$, no spatial region of the hyperspectral image appears in both training and testing sets, preventing spatial leakage while still allowing the model to utilize a large portion of the available pixels.

In Figures 4.2, 4.4, 4.6 the results of this selection process can be observed on each

of the three datasets.

4.1.1 Samson Dataset

The Samson dataset contains a 95×95 hyperspectral image with 156 spectral bands and three ground-truth endmembers: *Soil*, *Tree*, and *Water*. Figure 4.1 is an RGB approximation of the dataset and Figure 4.7 shows the normalized ground truth spectra for each endmember.

4.1.2 Jasper Ridge Dataset

Jasper Ridge comprises a 100×100 image with 198 spectral bands and four ground-truth endmembers: *Tree*, *Water*, *Dirt*, and *Road*. Figure 4.3 is an RGB approximation of the dataset and Figure 4.8 shows the normalized ground truth spectra for each endmember.

4.1.3 HYDICE Urban Dataset

HYDICE Urban has 307×307 pixels with 162 spectral bands and is available in three variants containing four, five, or six endmembers. We use the six-endmember variant with endmembers: *Asphalt Road*, *Grass*, *Tree*, *Roof*, *Metal*, and *Dirt*. Figure 4.5 is an RGB approximation of the dataset and Figure 4.9 shows the normalized ground truth spectra for each endmember.

4.2 Endmember Bundle Initialization

These datasets provide a single "true" spectrum for each endmember. In practice, endmember spectra are not unique: the same material can exhibit measurable spectral variability under different illumination, viewing geometry, sensor characteristics, and environmental conditions. For the first two models, we use this ground truth provided in each dataset, but our third model with bundled endmembers requires a more detailed



Figure 4.1: An RGB approximation of the Samson dataset

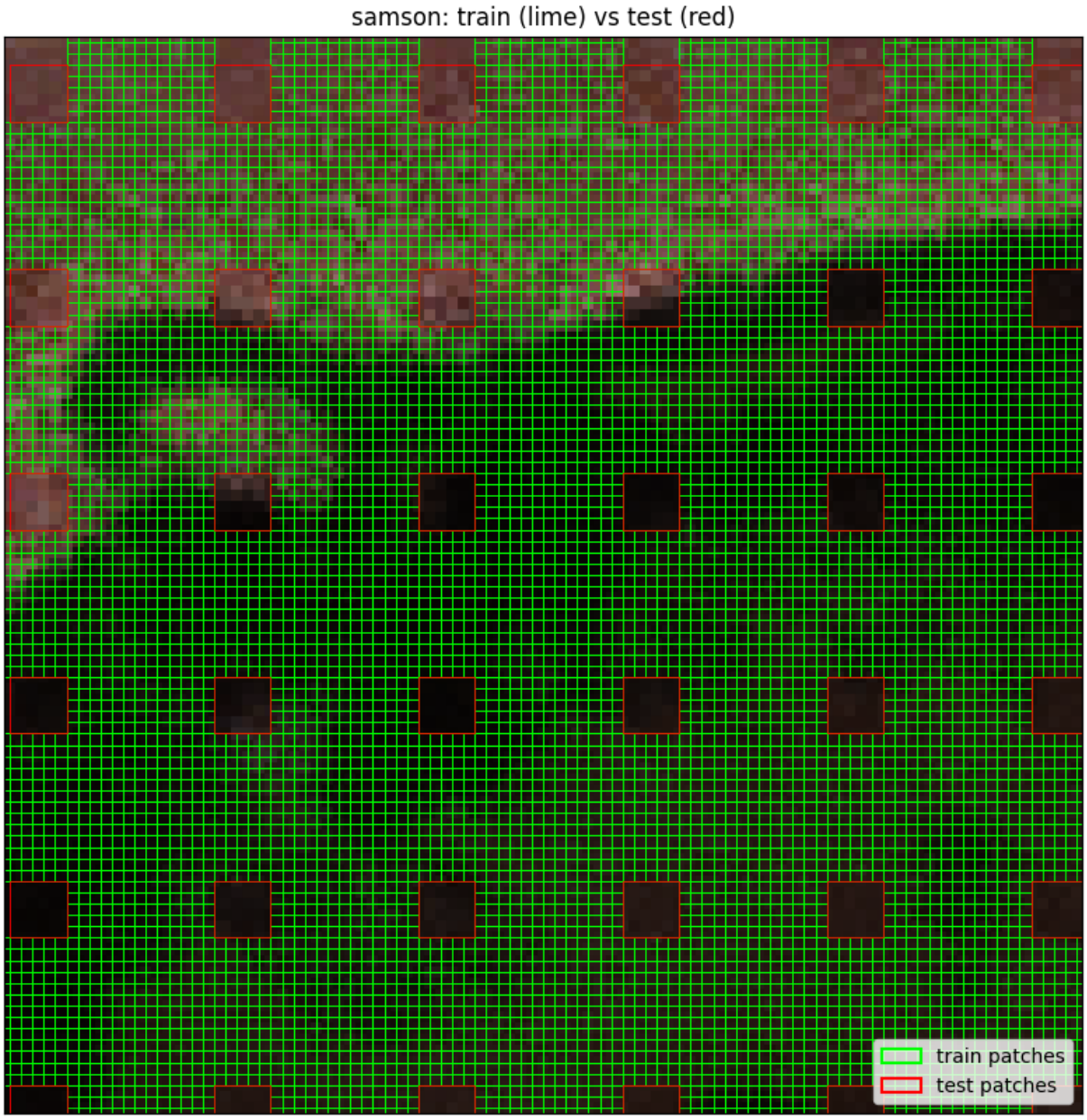


Figure 4.2: Visualization of the train/test split used for the experiments on Samson

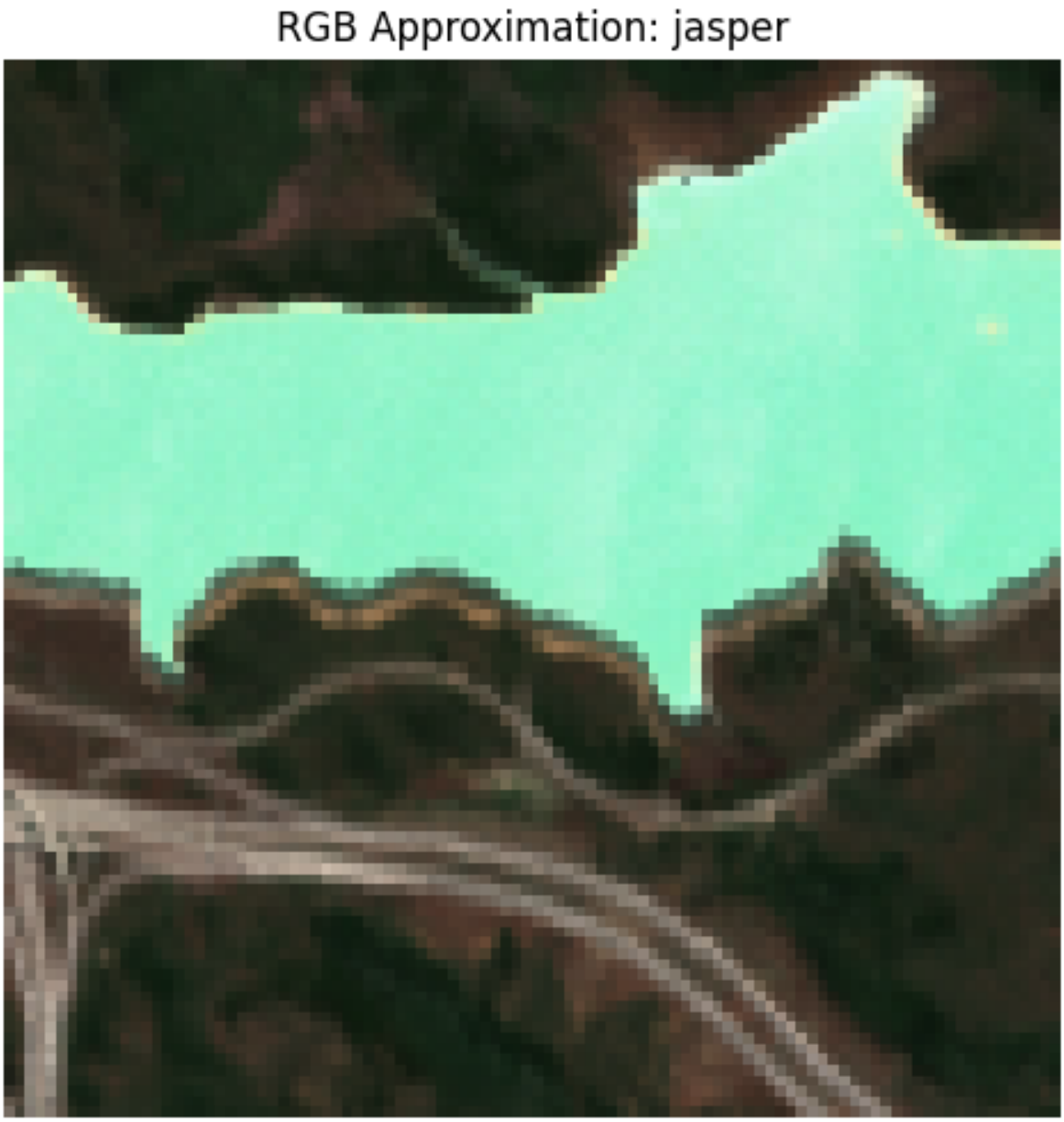


Figure 4.3: An RGB approximation of the Jasper Ridge dataset

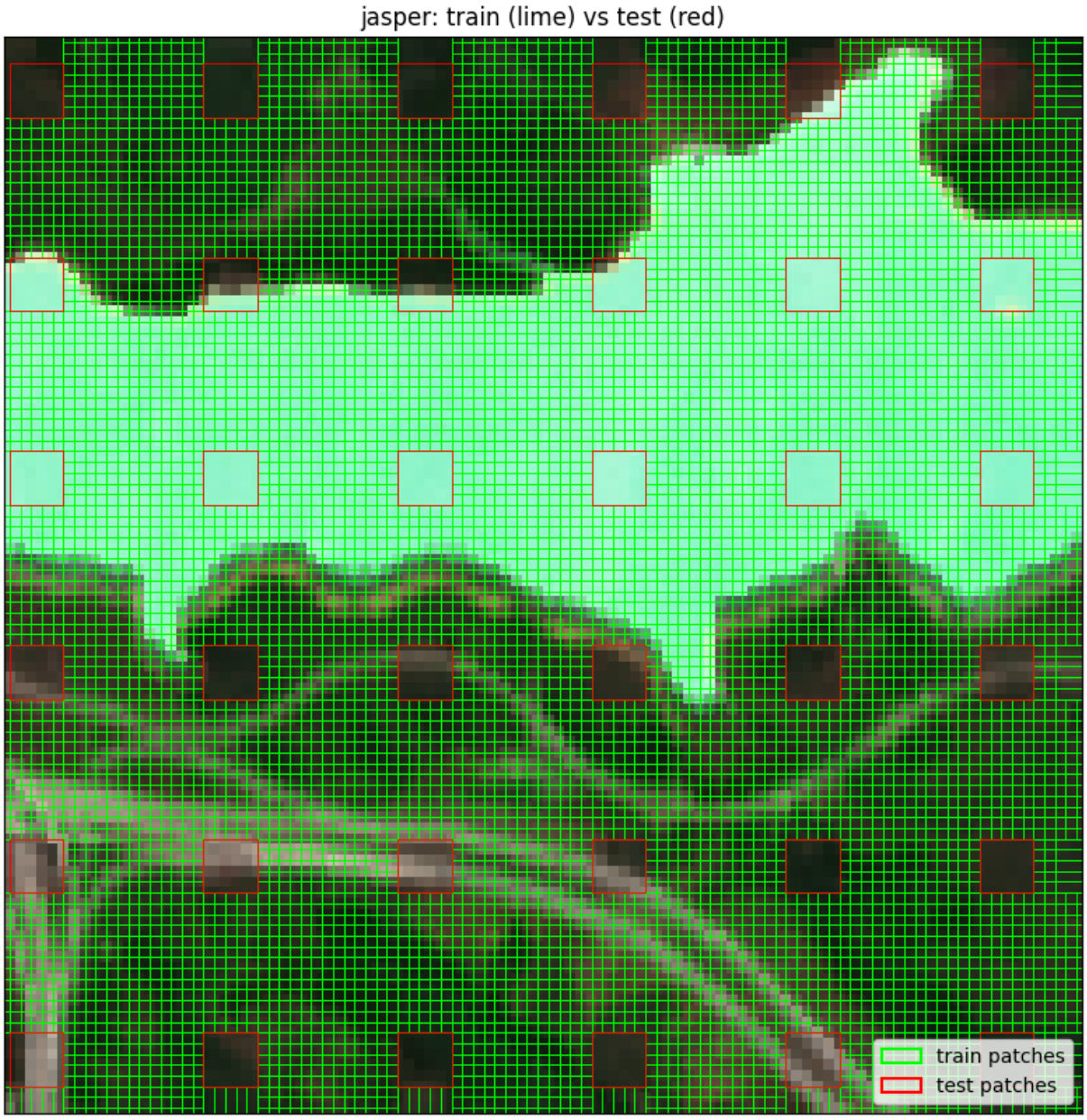


Figure 4.4: Visualization of the train/test split used for the experiments on Jasper Ridge



Figure 4.5: An RGB approximation of the HYDICE Urban dataset

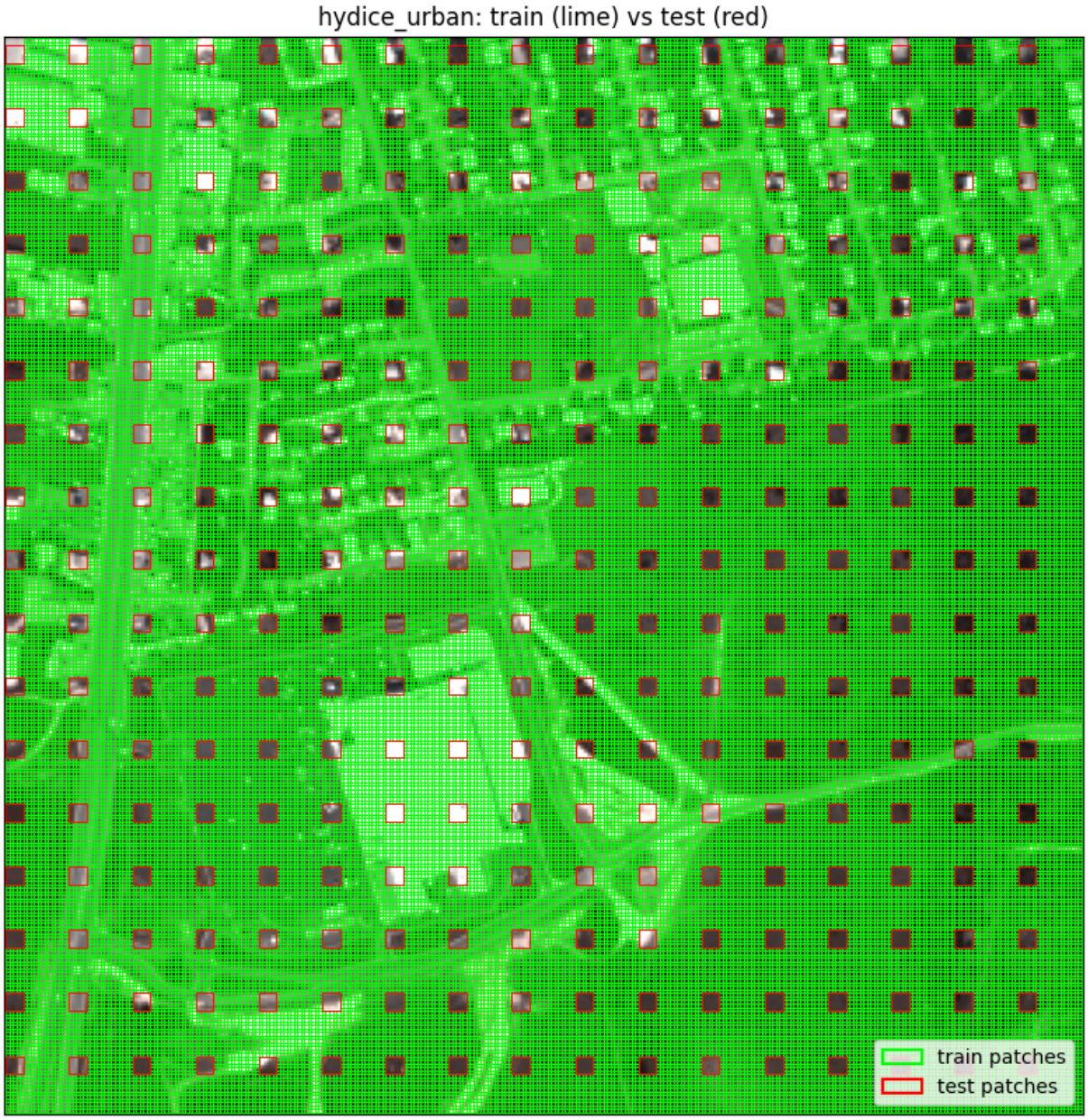
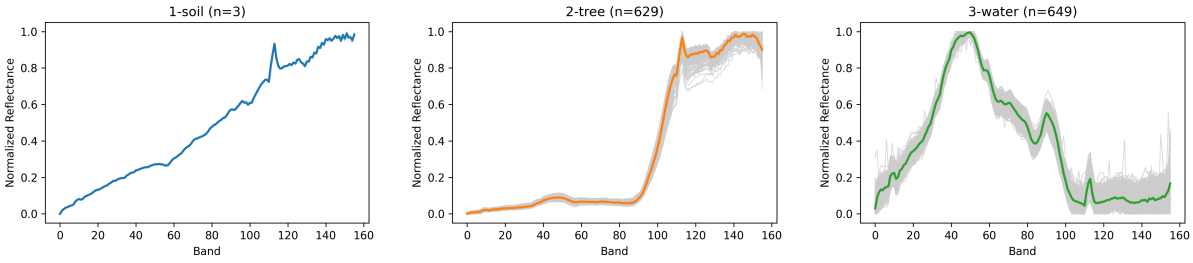
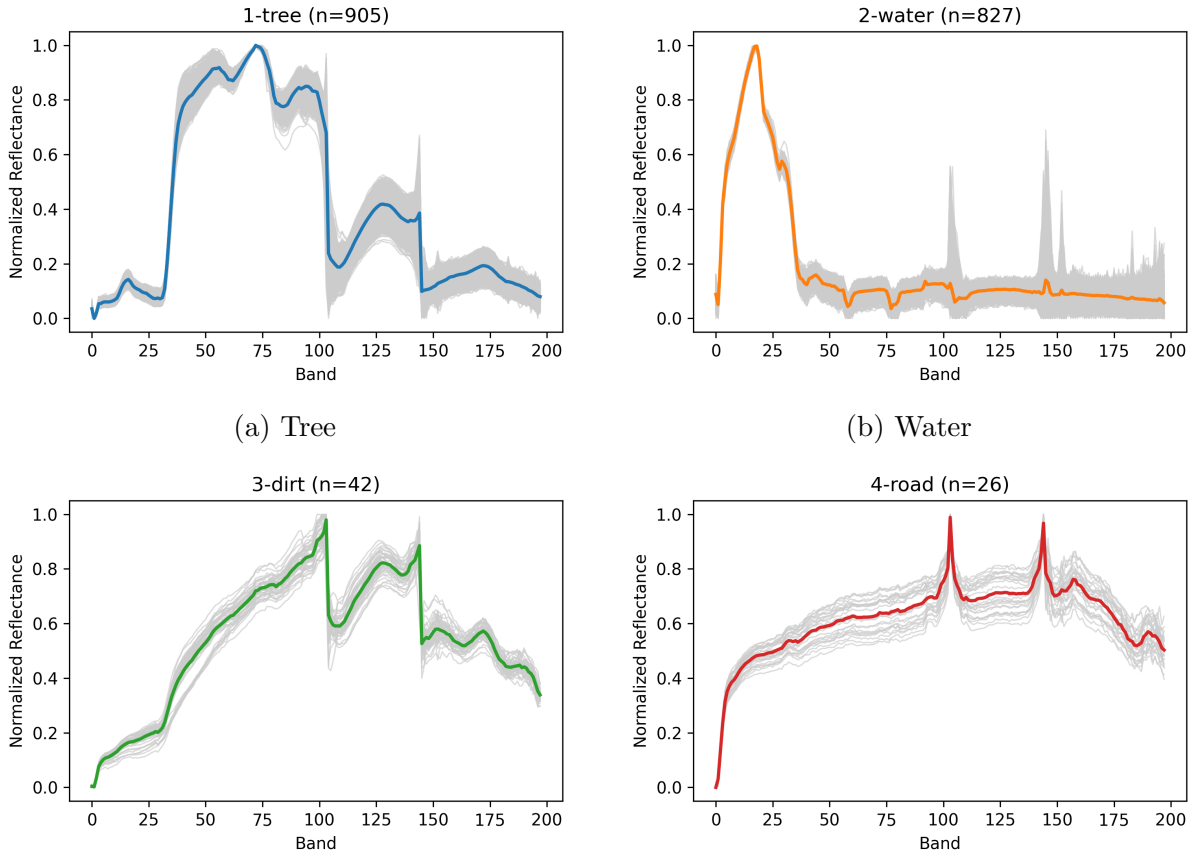


Figure 4.6: Visualization of the train/test split used for the experiments on HYDICE Urban



(a) Soil (b) Tree (c) Water

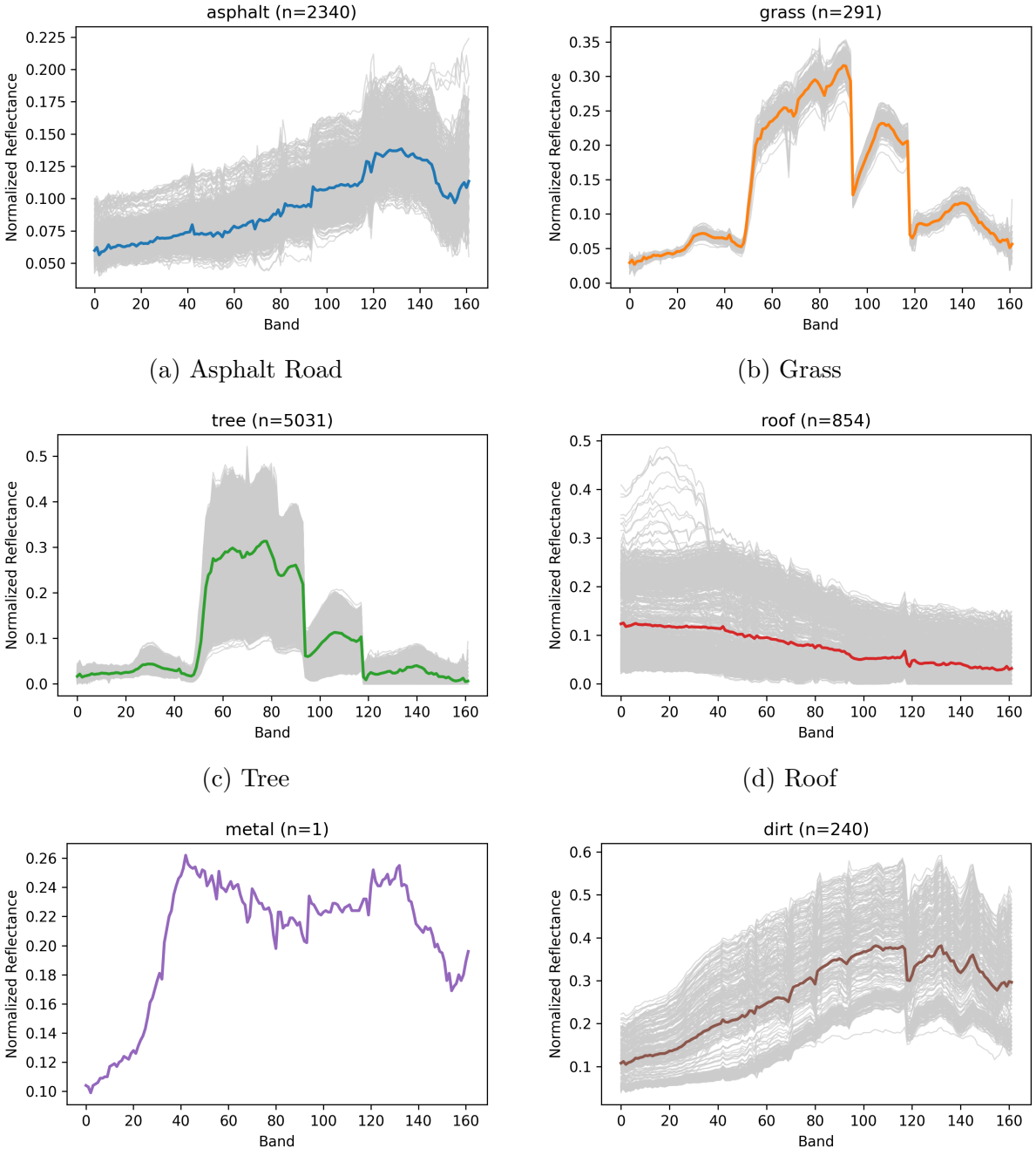
Figure 4.7: Normalized spectra of the pure pixels and the corresponding distribution mean for each endmember in the Samson dataset. In each subplot, the individual pure-pixel spectra are shown in grey and the distribution mean is shown in colour. The value n denotes the number of identified pure pixels.



(a) Tree (b) Water

(c) Dirt (d) Road

Figure 4.8: Normalized spectra of the pure pixels and the corresponding distribution mean for each endmember in the Jasper Ridge dataset. In each subplot, the individual pure-pixel spectra are shown in grey and the distribution mean is shown in colour. The value n denotes the number of identified pure pixels.



(a) Asphalt Road

(b) Grass

(c) Tree

(d) Roof

(e) Metal

(f) Dirt

Figure 4.9: Normalized spectra of the pure pixels and the corresponding distribution mean for each endmember in the HYDICE Urban dataset. In each subplot, the individual pure-pixel spectra are shown in grey and the distribution mean is shown in colour. The value n denotes the number of identified pure pixels.

ground truth. To achieve this, we manually calculate a new ground truth, first we identify pure pixels (where the ground truth abundance is 1.0 for any endmember) and then use their spectra to initialize the spectral distribution of each endmember. We model each endmember's spectral variability with a (multivariate) Gaussian distribution. Each endmember bundle is modeled as a multivariate Gaussian distribution,

$$\mathbf{e}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\mathbf{e}_k \in \mathbb{R}^C$ represents the spectral signature of endmember k, C is the number of spectral bands, $\boldsymbol{\mu}_k \in \mathbb{R}^C$ is the mean spectrum of endmember k, and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{C\times C}$ is the covariance matrix that models spectral variability across bands.

The parameters of this distribution are estimated from the set of pure pixels associated with each endmember. Let $\mathcal{S}_k = \{\mathbf{x}_i\}_{i=1}^{N_k}$ denote the set of pure pixel spectra for endmember k, where N_k is the number of pure pixels. The Gaussian parameters are then estimated as

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{x}_i,$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k - 1}\sum_{i=1}^{N_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T.$$

This allows each endmember to be represented by a distribution over spectra rather than a single fixed spectral signature, capturing the natural variability of materials in hyperspectral scenes. We refer to these as endmember bundles, that is, each endmember is represented by a distribution over spectra rather than a single spectral signature.

A visualization of each distribution and their respective means (μ_k) can be seen in Figures 4.7, 4.8, 4.9. The difference between the means of our calculated distribution and the ground truth spectra is reported in Tables 4.1, 4.2, 4.3. In figures 4.10, 4.11, 4.12, we show where identified pure pixels are located for each endmember across each dataset.

Table 4.1: SAD between mean of all GT pure pixels in the dataset and provided GT spectra for HYDICE Urban.

Endmember	SAD
asphalt road	0.073103
grass	0.042624
tree	0.130733
roof	0.174057
metal	0.029286
dirt	0.095820
Average	0.090937

Table 4.2: SAD between mean of all GT pure pixels in the dataset and provided GT spectra for Jasper Ridge.

Endmember	SAD
tree	0.070803
water	0.091218
soil	0.048795
road	0.042779
Average	0.063399

Table 4.3: SAD between mean of all GT pure pixels in the dataset and provided GT spectra for samson

Endmember	SAD
soil	0.098539
tree	0.043549
water	0.189304
Average	0.110464

4.3 Evaluation Metrics

The performance of our model is evaluated on two standard metrics, *Reconstruction accuracy* where our decoders reconstructed pixel is compared to the original input pixel and *Abundance Estimation* where our latent prediction is compared to the ground truth abundance of the input pixel.

We capture endmember reconstruction accuracy using Spectral Angle Distance (SAD), which quantifies the discrepancy between predicted and ground-truth endmember spectra by measuring the angle between their spectral vectors in a high-dimensional space, making it invariant to per-pixel intensity (scaling) changes:

$$\mathrm{SAD}(\hat{\mathbf{e}}, \mathbf{e}) = \cos^{-1}\left(\frac{\hat{\mathbf{e}}_k^\top \mathbf{e}_k}{\|\hat{\mathbf{e}}_k\| \, \|\mathbf{e}_k\|}\right),$$

where $\hat{\mathbf{e}}_k$ is the predicted endmember distribution's mean and $\mathbf{e}_k$ is the mean spectrum of the corresponding GT endmember bundle. Lower SAD values indicate closer alignment between the predicted and reference spectra. We use SAD to assess endmember-reconstruction accuracy for each material class in the dataset.

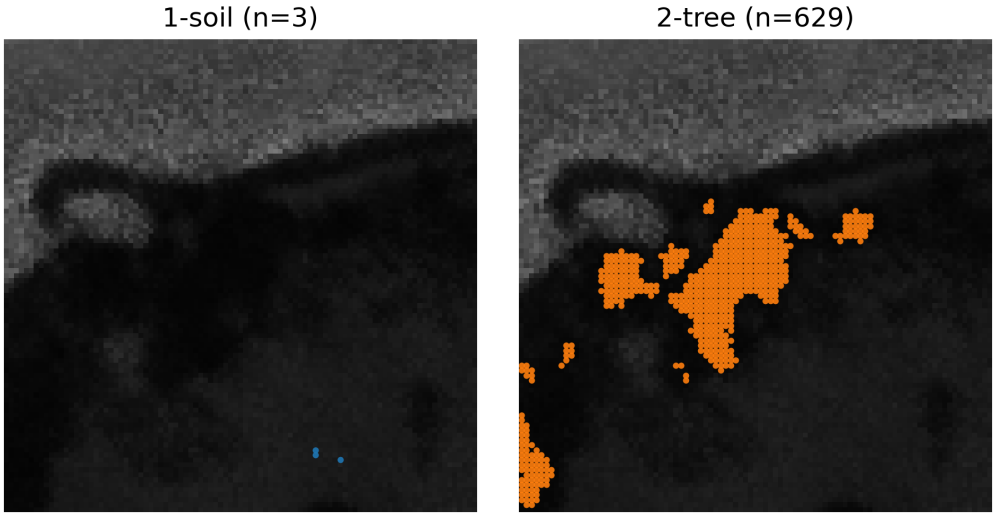
SAD is adopted for endmember evaluation because it compares the shape of spectral signatures while being insensitive to changes in magnitude. Since illumination variations can scale the observed spectra without changing the underlying material identity, this makes SAD well suited for assessing extracted endmembers.

We evaluate abundance-estimation accuracy using the Root Mean Squared Error (RMSE), which measures the average deviation between predicted and ground-truth abundances across pixels:

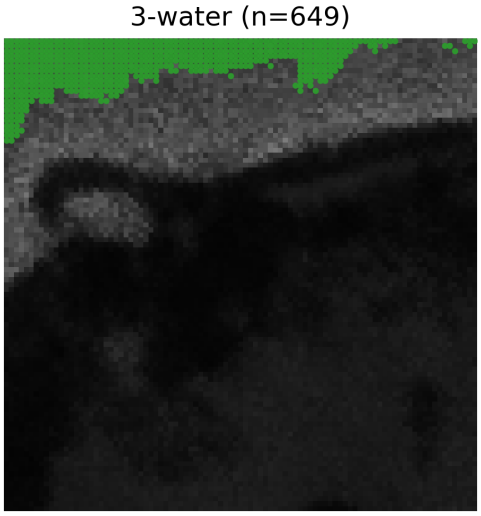
$$\mathrm{RMSE}(\hat{\mathbf{z}}, \mathbf{z}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \|\hat{\mathbf{z}}_n - \mathbf{z}_n\|^2},$$

where $\hat{\mathbf{z}}$ and $\mathbf{z}$ refers to the predicted and the ground truth abundances. N is the number

of pixels. A lower RMSE value indicates better abundance estimation. RMSE is used for abundance evaluation because abundance maps represent per-pixel material proportions, where the absolute numerical error is the target.

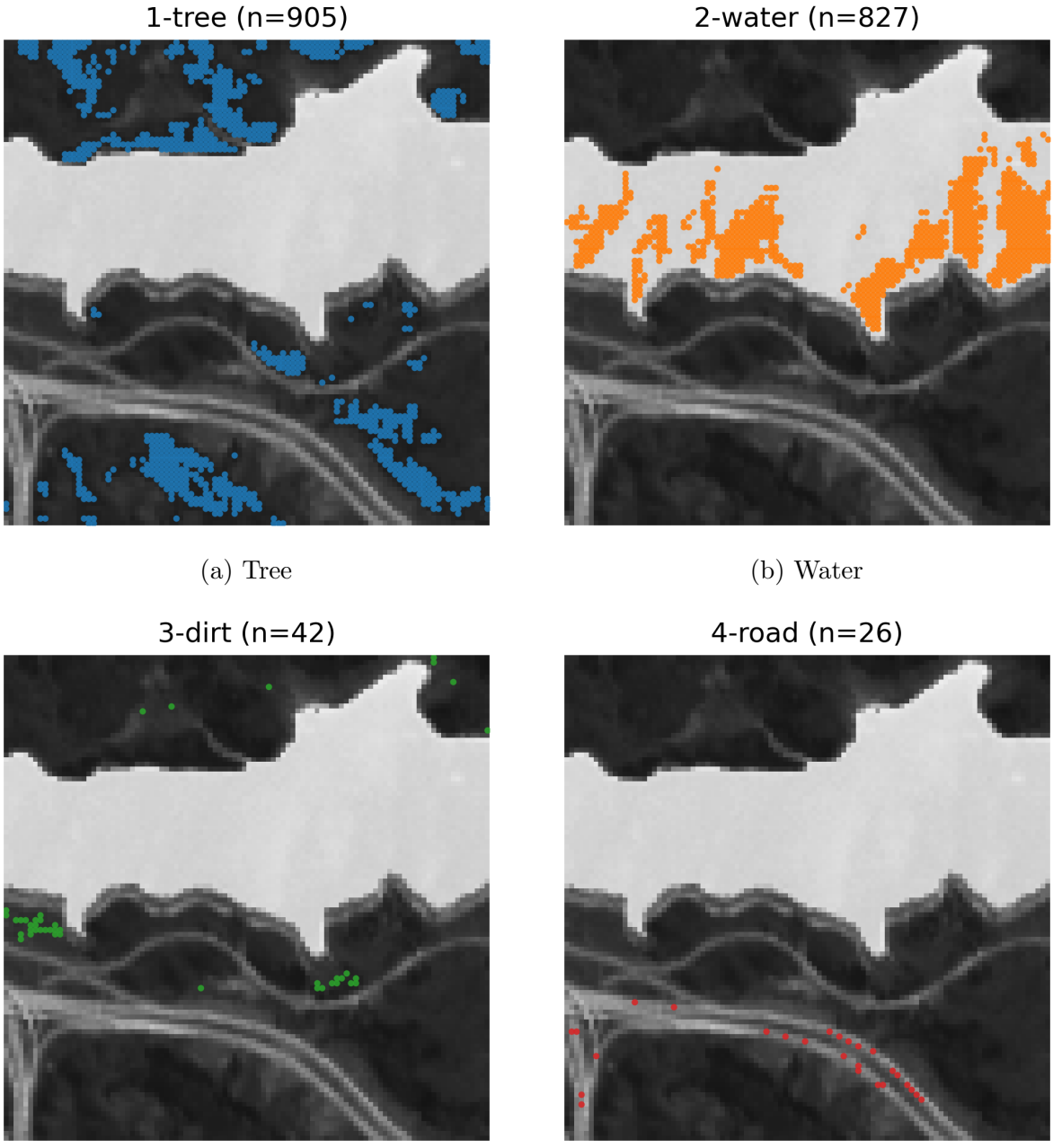


(a) Soil (b) Tree



(c) Water

Figure 4.10: Locations of the identified pure pixels for each endmember in the Samson dataset. The corresponding pure pixels are highlighted over a grayscale approximation of the hyperspectral scene.



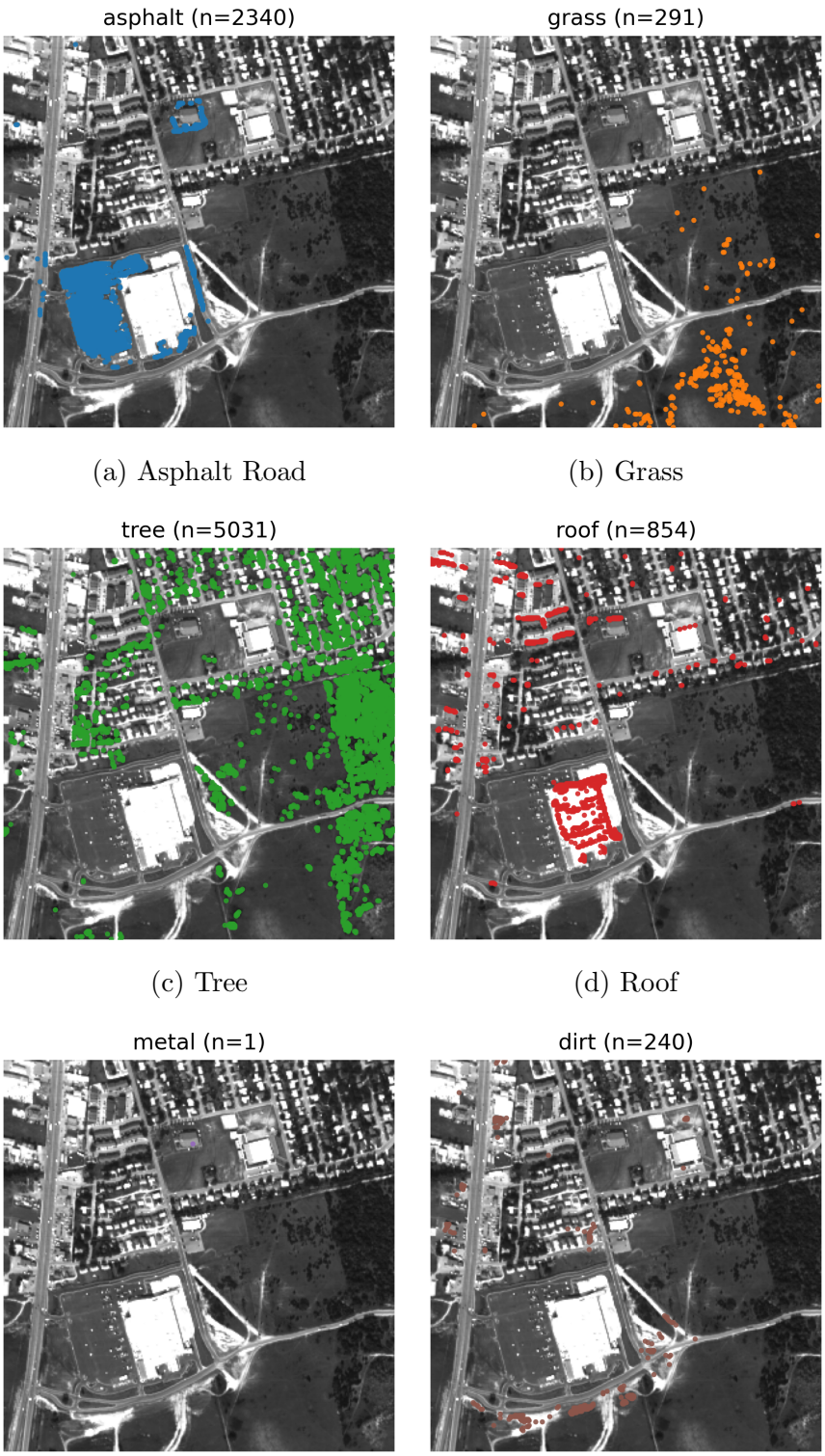
(a) Tree

(b) Water

(c) Dirt

(d) Road

Figure 4.11: Locations of the identified pure pixels for each endmember in the Jasper Ridge dataset. The corresponding pure pixels are highlighted over a grayscale approximation of the hyperspectral scene.



(a) Asphalt Road (b) Grass

(c) Tree (d) Roof

(e) Metal (f) Dirt

Figure 4.12: Locations of the identified pure pixels for each endmember in the HYDICE Urban dataset. The corresponding pure pixels are highlighted over a grayscale approximation of the hyperspectral scene.

Chapter 5

Experimental Results and Analysis

This chapter presents and analyzes the experimental results of the proposed LDVAE-T model. Performance is evaluated in terms of both endmember extraction and abundance estimation across the Samson, Jasper Ridge, and HYDICE Urban datasets. The proposed model variants are compared with one another and with existing hyperspectral unmixing methods using Spectral Angle Distance and Root Mean Square Error, supported by visual comparisons of the extracted spectra and predicted abundance maps.

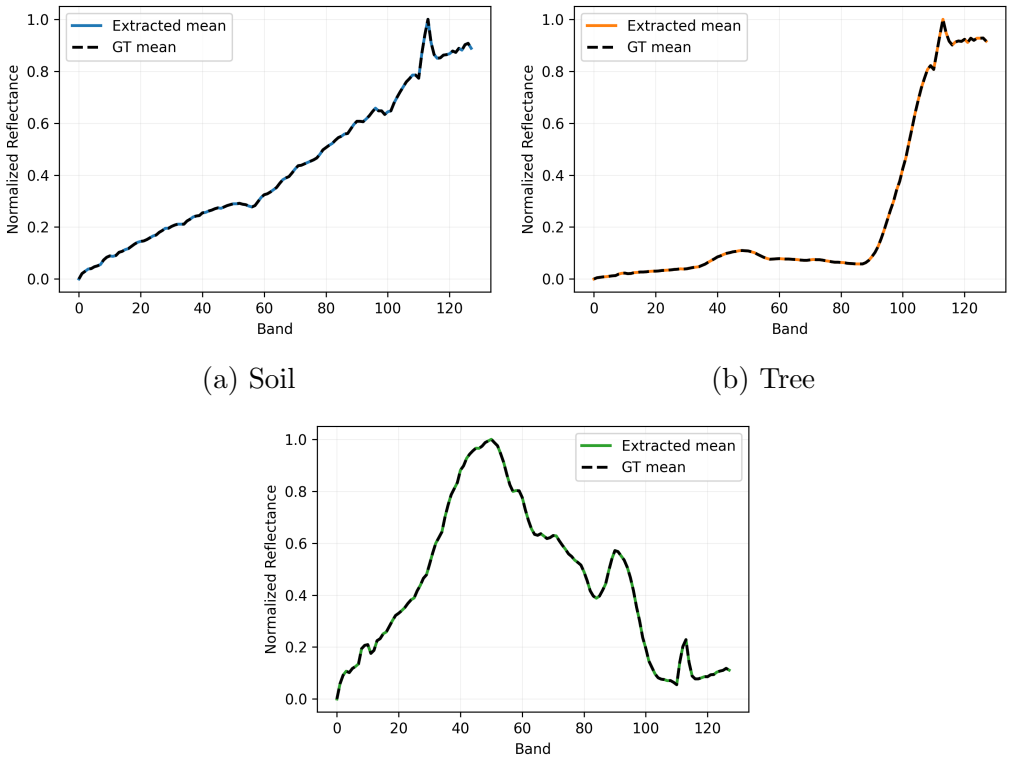
5.1 Endmember Extraction Results

Tables 5.1, 5.2, and 5.3 report Spectral Angle Distance (SAD) for the LDVAE-T model on the Samson, Jasper Ridge, and HYDICE Urban datasets compared to other works. Table 5.8 shows our three methods in comparison to each other.

For our Models, we observe that the change to a transformer decoder negatively impacted the endmember extraction of the model, while the change to bundled endmembers significantly improves upon it.

In comparison to other models, we observe consistently significantly lower SAD across datasets and endmembers, indicating that LDVAE-T extracts signatures that align closely with our reference GT endmembers under our evaluation pipeline. We attribute these

gains to the use of *bundled endmembers*; instead of enforcing a single fixed spectrum per material, the model estimates a distribution of plausible spectra. As a result of this, we are unable to achieve one-to-one compatibility with SAD reported in other papers; however, our approach better reflects real scenes where material signatures vary with illumination, sensor noise, and intra-class variability, and it reduces the need for the model to ignore meaningful spectral structure. These improvements are also consistent with the lower abundance RMSE reported in Tables 5.4, 5.5, and 5.6, suggesting that improved endmember characterization and improved mixing proportion estimates reinforce each other. Figures 5.1, 5.2, and 5.3 further visualize extracted versus reference spectra, illustrating how bundled endmembers enable closer agreement with the observed scene statistics while maintaining physically plausible reconstructions.



(a) Soil

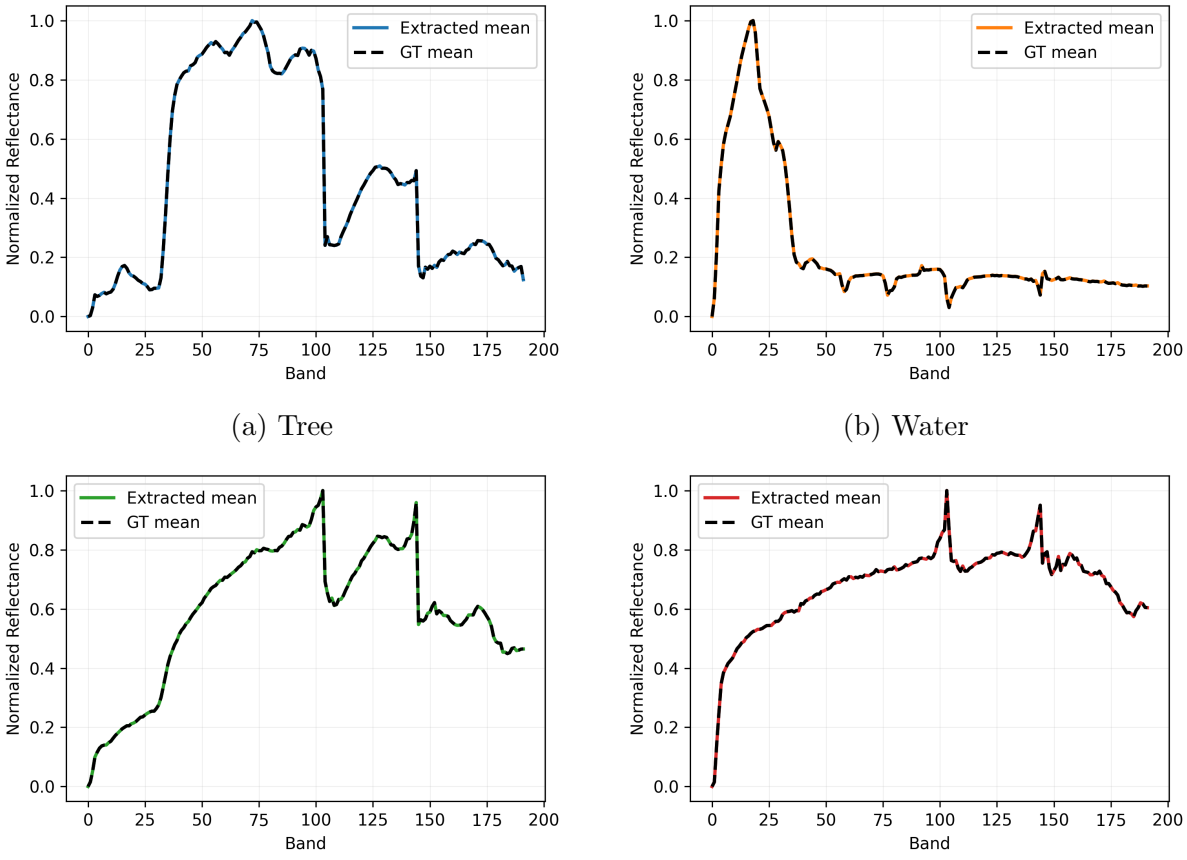
(b) Tree

(c) Water

Figure 5.1: Comparison between the extracted endmember spectra and the corresponding ground-truth mean spectra for the Samson dataset.

Table 5.1: SAD Scores for Endmember Extraction on Samson Dataset

Endmember	NMF-QMV [36] 2022	CNNAEU [24] 2021	LDVAE [22] 2024	SpACNN-LDVAE [5] 2024	DeepTrans-HsU [11] 2022	UnDAT [8] 2023	SSF-Net [31] 2024	SCEELA [10] 2025	SSLT-Net [9] 2025	LDVAE-T 2026
soil	0.02326	0.07565	0.0959	0.2097	0.0128	0.01191	0.0092	0.0117	0.0237	**0.000114** ↑98.76%
tree	0.06086	0.05440	1.2788	0.5347	0.0674	0.03775	0.0314	0.0309	0.0797	**0.000202** ↑99.36%
water	1.45643	0.03642	0.4022	0.8233	0.0729	0.00813	0.0373	0.0326	0.1052	**0.000213** ↑97.38%
average	0.51352	0.05549	0.5923	0.5525	0.0510	0.01926	0.0260	0.0250	0.0695	**0.000177** ↑99.08%



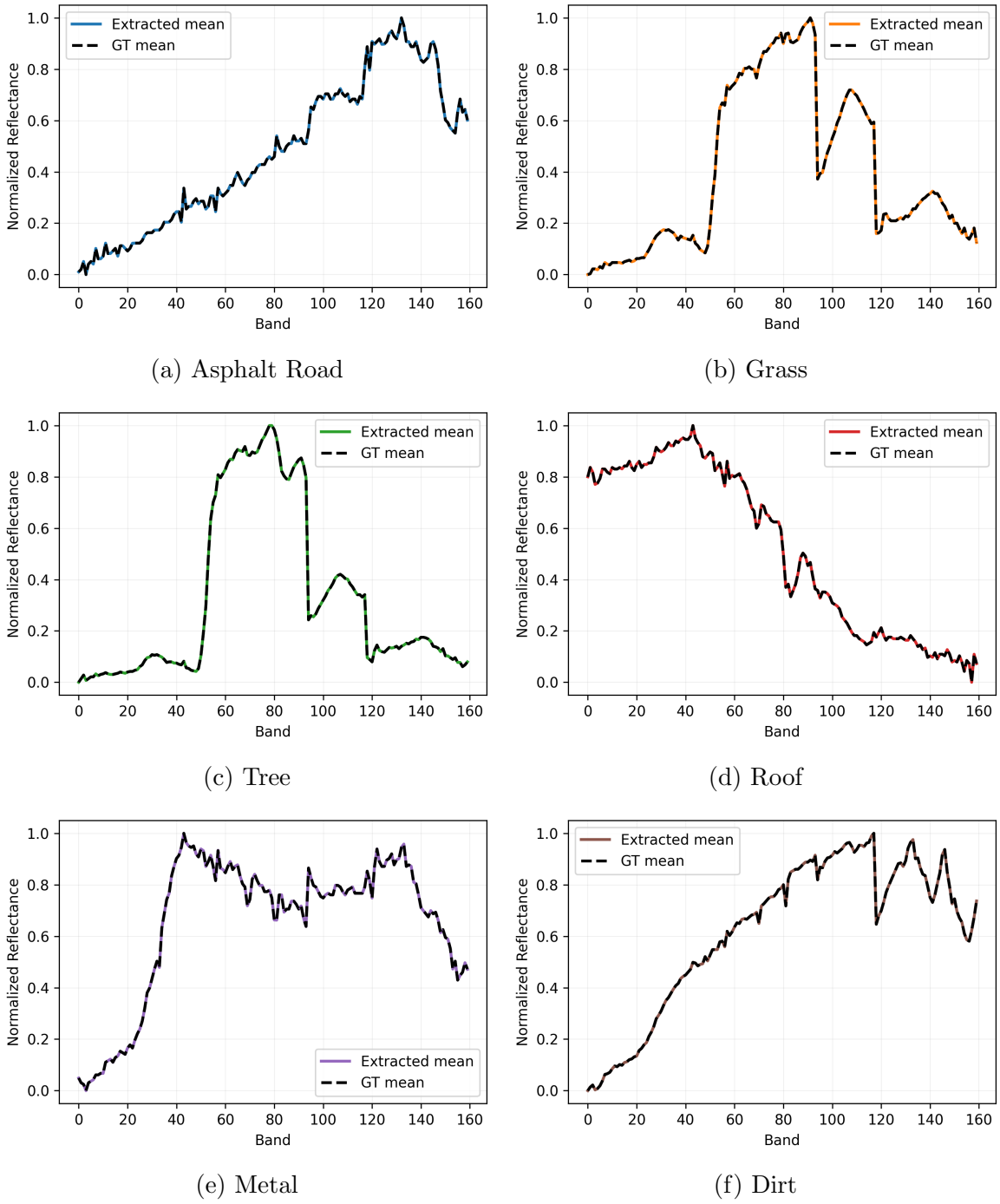
(a) Tree (b) Water

(c) Dirt (d) Road

Figure 5.2: Comparison between the extracted endmember spectra and the corresponding ground-truth mean spectra for the Jasper Ridge dataset.

Table 5.2: SAD Scores for Endmember Extraction on Jasper Ridge Dataset

Endmember	NMF-QMV [36] 2022	CNNAEU [24] 2021	LDVAE [22] 2024	SpACNN-LDVAE [5] 2024	DeepTrans-HsU [11] 2022	UnDAT [8] 2023	SSF-Net [31] 2024	SCEELA [10] 2025	SSLT-Net [9] 2025	LDVAE-T 2026
tree	0.05016	0.3104	-	-	-	0.04519	0.0781	0.1222	0.1587	**0.000255** ↑99.44%
water	0.28387	0.6082	-	-	-	0.02811	0.0293	0.1355	0.0920	**0.001269** ↑95.49%
soil	0.17326	0.3381	-	-	-	0.09074	0.0484	0.0427	0.0988	**0.000270** ↑99.44%
road	1.46974	0.0519	-	-	-	0.03306	0.0238	0.0421	0.0675	**0.000288** ↑98.79%
average	0.49426	0.3271	-	-	-	0.04927	0.0449	0.0870	0.1043	**0.000520** ↑98.84%



(a) Asphalt Road

(b) Grass

(c) Tree

(d) Roof

(e) Metal

(f) Dirt

Figure 5.3: Comparison between the extracted endmember spectra and the corresponding ground-truth mean spectra for the HYDICE Urban dataset.

Table 5.3: SAD Scores for Endmember Extraction on HYDICE Urban Dataset. *SSF-Net used a five-endmember version of HYDICE Urban for their experiments.

Endmember	SGSNMF [32] 2017	SSWNMF [35] 2022	LDVAE [22] 2024	SpACNN-LDVAE [5] 2024	SSF-Net* [31] 2024	LDVAE-T 2025
Asphalt road	0.0841	0.0782	0.4262	0.2786	0.0629	**0.000216** ↑99.66%
Grass	0.1516	0.1490	0.3323	0.1936	0.0411	**0.000145** ↑99.65%
Tree	0.1199	0.1173	0.3177	0.4411	0.0850	**0.000155** ↑99.82%
Roof	0.0731	0.0713	0.4393	0.4502	0.0487	**0.000208** ↑99.57%
Metal	0.1250	0.1241	0.7004	0.3241	-	**0.000338** ↑99.73%
Dirt	0.0859	0.0802	0.2806	0.2026	0.1065	**0.000088** ↑99.89%
Average	0.1060	0.1034	0.4161	0.3151	0.0688	**0.000192** ↑99.72%

5.2 Abundance Estimation Results

Tables 5.4, 5.5, and 5.6 report Root Mean Squared Error (RMSE) for the LDVAE-T model on the Samson, Jasper Ridge, and HYDICE Urban datasets compared to outside methods, while Table 5.8 shows our three methods in comparison to each other.

Concerning our methods, both the bundled endmembers and the transformer decoder models show a slight drop across the board for RMSE compared to the base model. As far as the bundled endmembers, we believe this is attributed to the added variability of spectral signatures causing slightly less concrete abundance estimation.

In comparison to other works, our model consistently achieves lower RMSE across all endmembers and datasets, outperforming state-of-the-art baselines. This performance is primaraly attributable to the Dirichlet prior imposed in the latent space, which naturally enforces the sum-to-one and non-negativity constraints expected of abundance vectors, yielding stable and physically meaningful estimates. By contrast, methods based on unconstrained regression can violate these constraints or require *ad hoc* normalization. Figures 5.4, 5.5, and 5.6 show heatmaps of the predicted abundances vs. ground truth abundances for the Samson, Jasper Ridge, and HYDICE Urban datasets.

Table 5.4: RMSE Scores for Abundance Estimation on Samson Dataset

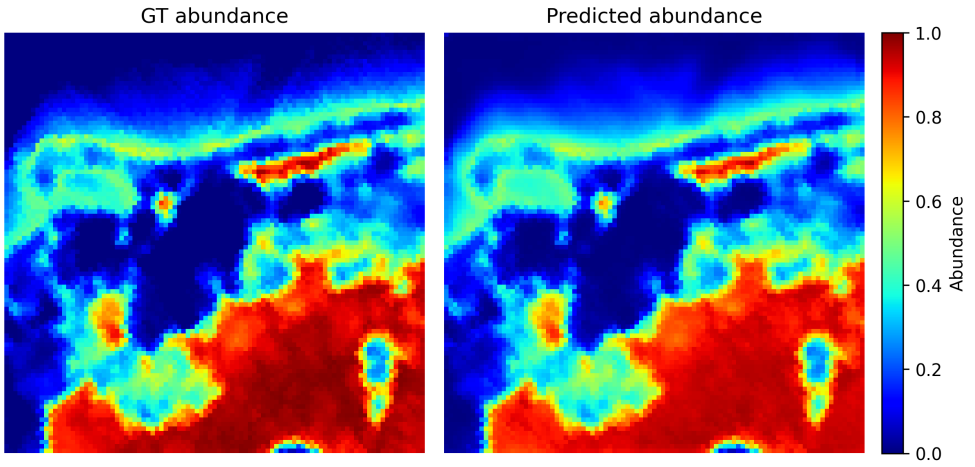
Endmember	NMF-QMV [36] 2022	CNNAEU [24] 2021	LDVAE [22] 2024	SpACNN-LDVAE [5] 2024	DeepTrans-HsU [11] 2022	UnDAT [8] 2023	SSF-Net [31] 2024	SCEELA [10] 2025	SSLT-Net [9] 2025	LDVAE-T 2026
soil	0.23298	0.3157	0.2609	0.2097	0.0712	0.04306	0.0511	0.0442	-	**0.03032** ↑29.59%
tree	0.24432	0.2911	0.3431	0.5347	0.0683	0.02854	0.0502	0.0266	-	**0.02401** ↑15.87%
water	0.37621	0.1552	0.3165	0.2098	0.0930	0.03128	0.0272	0.0258	-	**0.02249** ↑17.32%
average	0.28450	0.2540	0.3078	0.2412	0.0783	0.03429	0.0428	0.0333	0.0687	**0.02561** ↑25.31%

Table 5.5: RMSE Scores for Abundance Estimation on Jasper Ridge Dataset

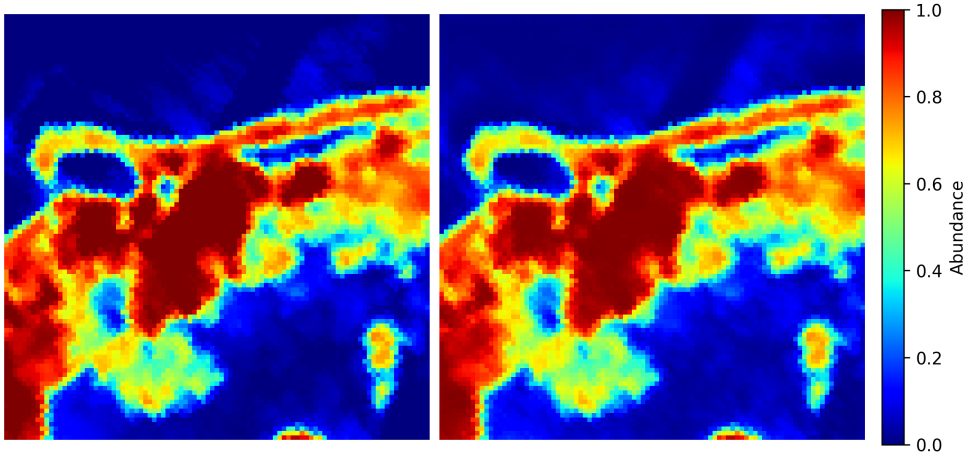
Endmember	NMF-QMV [36] 2022	CNNAEU [24] 2021	LDVAE [22] 2024	SpACNN-LDVAE [5] 2024	DeepTrans-HsU [11] 2022	UnDAT [8] 2023	SSF-Net [31] 2024	SCEELA [10] 2025	SSLT-Net [9] 2025	LDVAE-T 2026
tree	0.12528	0.3169	-	-	-	0.06031	0.0721	0.0822	-	**0.02609** ↑56.74%
water	0.20375	0.2118	-	-	-	0.04211	0.0761	.0807	-	**0.02892** ↑31.32%
soil	0.14647	0.2978	-	-	-	0.07007	0.0930	0.0776	-	**0.03855** ↑44.98%
road	0.18446	0.2043	-	-	-	0.07792	0.0782	0.1048	-	**0.03819** ↑50.99%
average	0.16499	0.2577	-	-	-	0.06260	0.0798	0.0854	0.1143	**0.03294** ↑47.38%

Table 5.6: RMSE Scores for Abundance Estimation on HYDICE Urban Dataset. *SSF-Net used a five-endmember version of HYDICE Urban for their experiments.

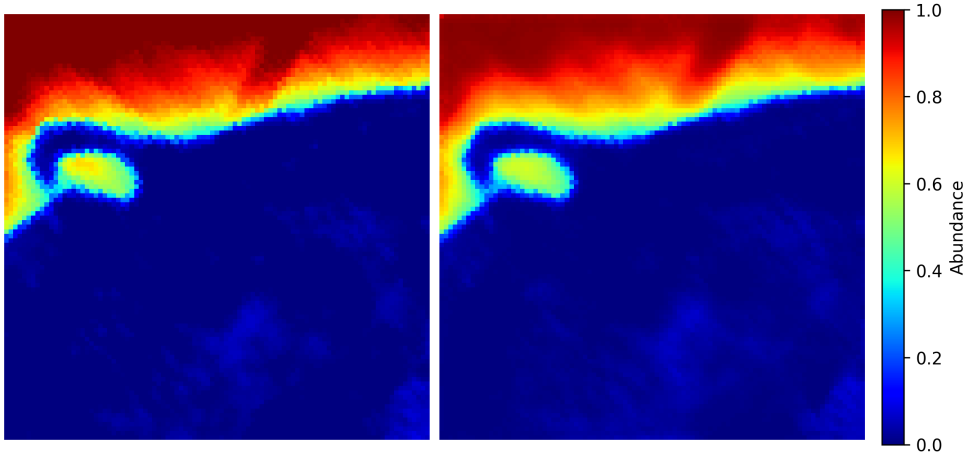
Endmember	SGSNMF [32] 2017	SSWNMF [35] 2022	LDVAE [22] 2024	SpACNN-LDVAE [5] 2024	SSF-Net* [31] 2024	LDVAE-T 2025
Asphalt road	-	-	0.2889	0.1566	0.1578	**0.03538** ↑77.41%
Grass	-	-	0.1832	0.1977	0.1416	**0.03083** ↑78.23%
Tree	-	-	0.1737	0.1632	0.1179	**0.02576** ↑78.15%
Roof	-	-	0.1250	0.1283	0.0909	**0.02476** ↑72.76%
Metal	-	-	0.2599	0.0992	-	**0.03830** ↑61.39%
Dirt	-	-	0.1334	0.1894	0.1192	**0.03128** ↑73.76%
Average	-	-	0.1840	0.1558	0.1256	**0.03105** ↑75.28%



(a) Soil

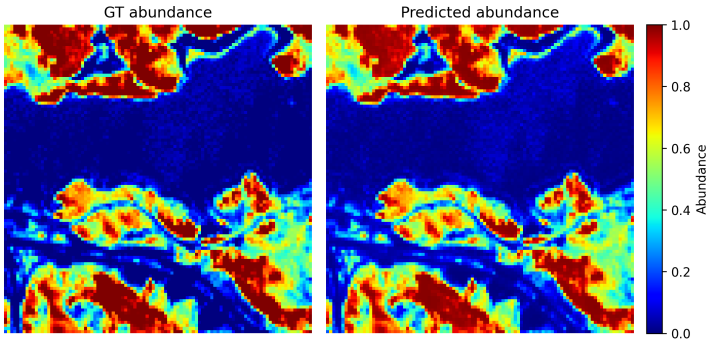


(b) Tree

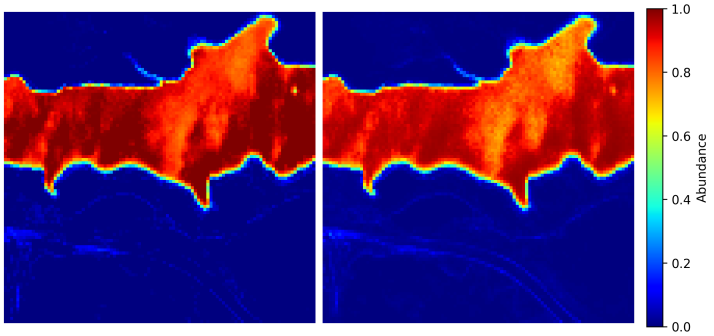


(c) Water

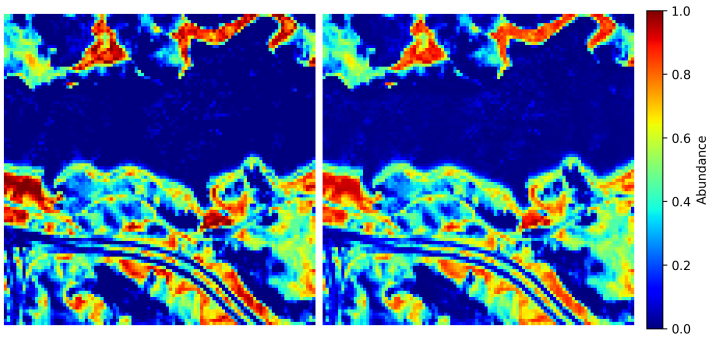
Figure 5.4: Ground-truth and predicted abundance maps for each endmember in the Samson dataset.



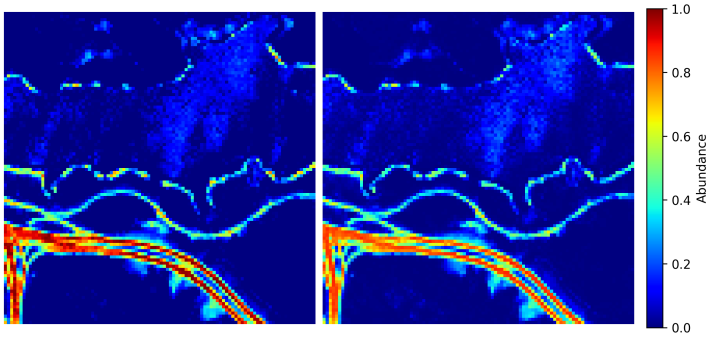
(a) Tree



(b) Water

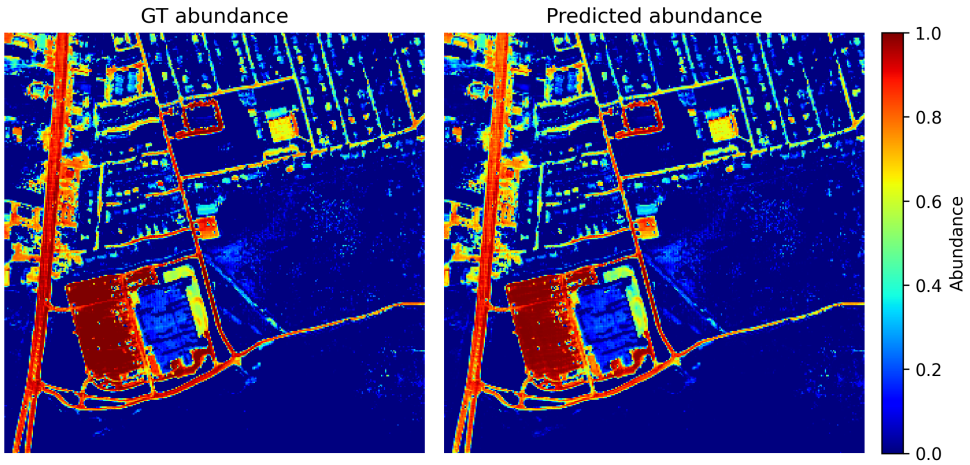


(c) Dirt

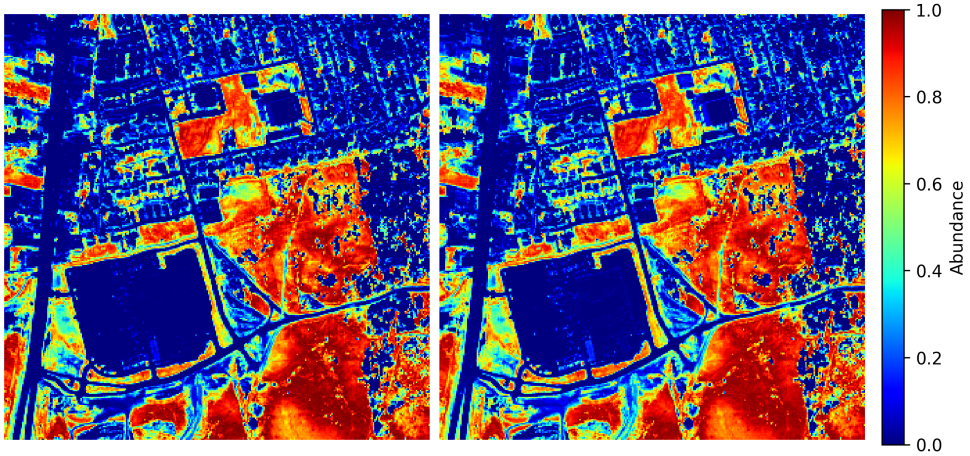


(d) Road

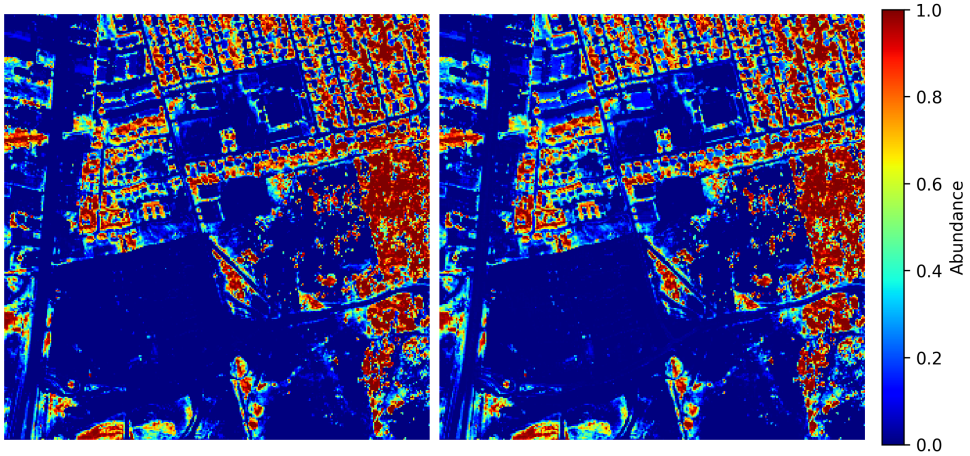
Figure 5.5: Ground-truth and predicted abundance maps for each endmember in the Jasper Ridge dataset.



(a) Asphalt Road

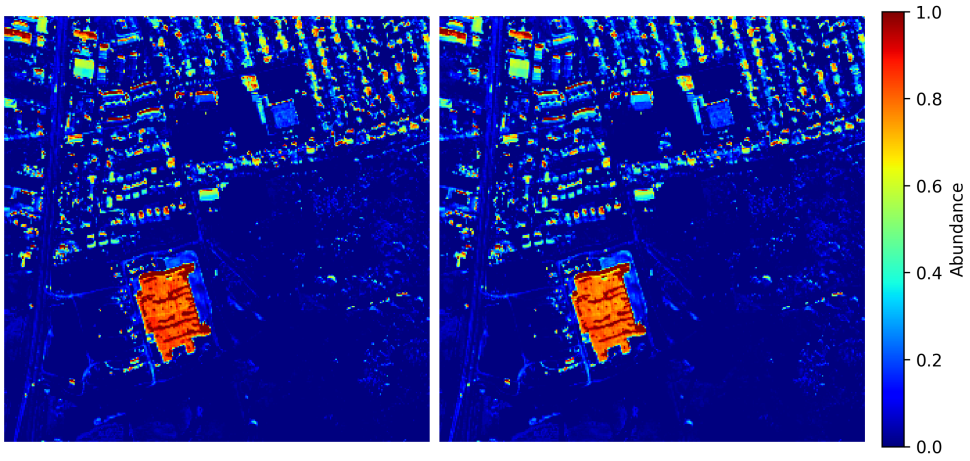


(b) Grass

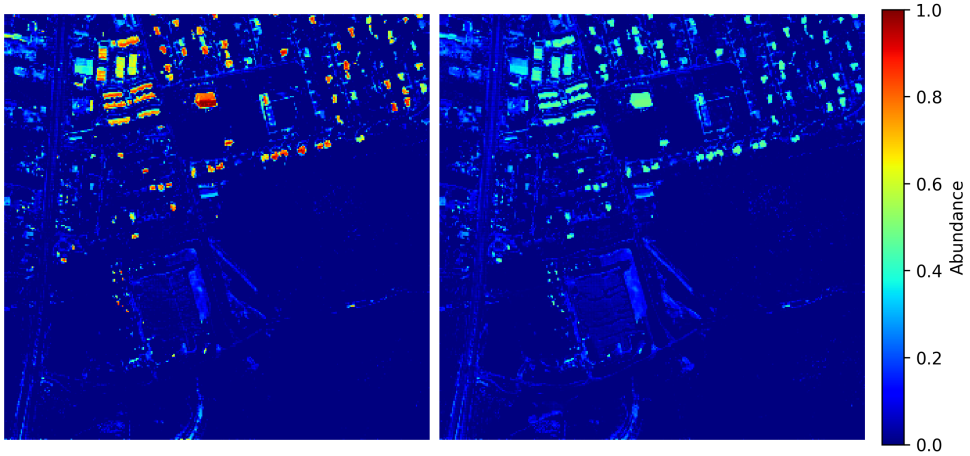


(c) Tree

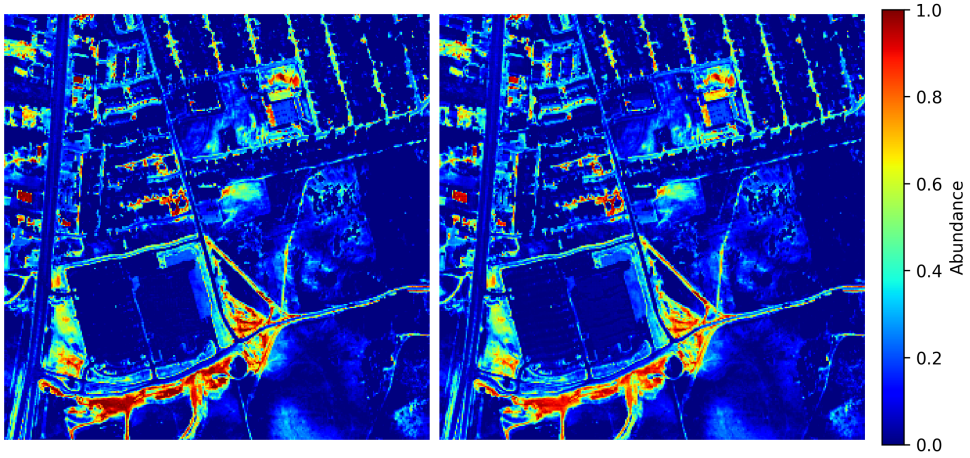
Figure 5.6: Ground-truth and predicted abundance maps for each endmember in the HYDICE Urban dataset.



(d) Roof



(e) Metal



(f) Dirt

Figure 5.6: Ground-truth and predicted abundance maps for each endmember in the HYDICE Urban dataset (continued).

5.3 Ablation Studies

We ran an ablation study with the aim of determining optimal hyperparameters for the LDVAE-T model. We experimented with different patch sizes, number of attention heads and segment size. Results for this test can be observed in Table 5.7. All tests in this table were run for 500 epochs (500 less then reported final results, which explains why numbers vary). We decided on a middle-ground 5/16/32 for all the above experiments as a result of this ablation study.

Table 5.7: Ablation study results for patch size 3/5/7, number of attention heads 4/8/16, and segment size 20/24/32. Note, for best results, number of heads should divide segment size. All run for 500 epochs

Dataset	Patch	Heads	Segment	SAD	RMSE
hydice urban	3	4	20	**0.000157**	**0.038208**
	3	8	24	0.000187	0.043541
	3	16	32	0.000206	0.041650
	5	4	20	0.000215	0.045332
	5	8	24	0.000329	0.049416
	5	16	32	0.000254	0.046787
	7	4	20	0.000215	0.062915
	7	8	24	0.000326	0.054691
	7	16	32	0.000377	0.064046
jasper	3	4	20	0.001767	0.030693
	3	8	24	0.001738	0.025851
	3	16	32	0.001495	**0.022713**
	5	4	20	0.002272	0.031072
	5	8	24	0.002281	0.029395
	5	16	32	0.001922	0.029866
	7	4	20	0.001085	0.044746
	7	8	24	0.001640	0.023704
	7	16	32	**0.000892**	0.023634
samson	3	4	20	0.000252	0.028158
	3	8	24	**0.000180**	0.021613
	3	16	32	0.000691	0.020097
	5	4	20	0.000443	0.026597
	5	8	24	0.000461	0.021319
	5	16	32	0.000472	0.019943
	7	4	20	0.000623	0.037254
	7	8	24	0.000752	0.031850
	7	16	32	0.000426	**0.019499**

Table 5.8: Abundance estimation (RMSE) and endmember extraction (SAD) average across our three models.

Dataset	Baseline Transformer		Transformer Decoder		Bundled Endmembers	
	RMSE (↓)	SAD (↓)	RMSE (↓)	SAD (↓)	RMSE (↓)	SAD (↓)
Samson	0.015920	0.070140	0.01710	0.115240	0.02561	0.000177
Jasper Ridge	0.015667	0.109783	0.02020	0.185030	0.03294	0.000520
HYDICE Urban	0.015050	0.240190	0.01680	0.360019	0.03105	0.000192

Chapter 6

Conclusion and Future Work

We introduce the *Latent Dirichlet Transformer Variational Autoencoder* (LDVAE-T), a framework for hyperspectral unmixing that couples transformer-based modeling with a Dirichlet latent structure. By imposing a Dirichlet prior in the latent space, the model naturally enforces the sum-to-one and non-negativity constraints required for physically plausible abundance estimates. Beyond this probabilistic scaffold, LDVAE-T contributes a decoder based on *bundled endmembers*: instead of treating each material as a single, fixed prototype, the decoder predicts for each patch a distributional prototype comprising a mean spectrum and a structured covariance defined over spectral segments. Reconstructions are formed by mixing these bundles with Dirichlet-distributed abundances, enabling the network to capture intrinsic intra-class variability while preserving interpretability. Experiments on three standard benchmarks:*Samson*, *Jasper Ridge*, and *HYDICE Urban*, show that LDVAE-T delivers state-of-the-art performance in both endmember extraction and abundance estimation, as measured by spectral angle distance and root mean squared error, respectively.

The empirical gains of LDVAE-T stem from the combination of three ingredients: 1) a *physically meaningful latent space* via the Dirichlet prior, which allows for a VAE framework with a distribution that works within the constraints of a mixture; 2) *distributional*

endmembers that explicitly model spectral variability present in real-world scenes; and 3) *transformer-based encoding* that leverages long-range spectral–spatial dependencies often missed by MLP or CNN encoders. We find that Segment-wise covariance parameterization strikes a balance between flexibility (capturing correlated variation within bands) and tractability (avoiding full dense covariances over all wavelengths). Collectively, these elements improve both abundance estimation and the stability of extracted endmember signatures.

6.1 Limitations and Future Research Directions

One limitation of the current model is that it is not designed to be directly applied to a new hyperspectral scene without additional training or adaptation. Since the learned endmember representations and abundance estimation process are optimized for the spectral characteristics of a specific dataset, changes in sensor conditions, illumination, atmospheric effects, material composition, or captured wavelengths may require retraining. However, this limitation is acceptable in many industry settings, where models are often trained or calibrated for a specific sensor, site, or operational environment before deployment.

Future work may extend LDVAE-T in several directions. One direction is to incorporate the model into an iterative LDVAE framework, following Mantripragada and Qureshi [22], to progressively refine endmember and abundance estimates. Another direction is to investigate richer representations for endmember distributions that can better model spectral variability. Further evaluation on additional hyperspectral benchmarks is also needed to assess robustness across different datasets and imaging conditions. Finally, future work could explore few-shot and zero-shot applications.

Bibliography

[1] Anuja Bhargava, Ashish Sachdeva, Kulbhushan Sharma, Mohammed H. Alsharif, Peerapong Uthansakul, and Monthippa Uthansakul. Hyperspectral imaging and its applications: A review. *Heliyon*, 10(12):e33208, 2024.

[2] Anuja Bhargava, Ashish Sachdeva, Kulbhushan Sharma, Mohammed H. Alsharif, Peerapong Uthansakul, and Monthippa Uthansakul. Hyperspectral imaging and its applications: A review. *Heliyon*, 10(12):e33208, 2024.

[3] Ricardo Augusto Borsoi, Tales Imbiriba, and José Carlos Moreira Bermudez. Deep generative endmember modeling: An application to unsupervised spectral unmixing. *IEEE Transactions on Computational Imaging*, 6:374–384, 2020.

[4] Thomas Burr. Pattern recognition and machine learning. *Journal of the American Statistical Association*, 103(482):886–887, 2008.

[5] Soham Chitnis, Kiran Mantripragada, and Faisal Z. Qureshi. Spacnn-ldvae: Spatial attention convolutional latent dirichlet variational autoencoder for hyperspectral pixel unmixing. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 7714–7719, 2024.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words:

Transformers for image recognition at scale, 2021. Copyright - © 2021. This work is published under http://arxiv.org/licenses/nonexclusive-distrib/1.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2024-10-16.

[7] Lucas Drumetz, Jocelyn Chanussot, and Christian Jutten. Spectral unmixing: A derivation of the extended linear mixing model from the hapke model. *IEEE Geoscience and Remote Sensing Letters*, 17(11):1866–1870, 2020.

[8] Yuexin Duan, Xia Xu, Tao Li, Bin Pan, and Zhenwei Shi. Undat: Double-aware transformer for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

[9] Deng et al. Sslt-net: A spatial–spectral linear transformer unmixing network for hyperspectral image. *IEEE GRS Letters*, 22:1–5, 2025.

[10] Ratnayake et al. Hyperspectral unmixing with spatial context and endmember ensemble learning with attention mechanism. *ISPRS Open J. of Photogrammetry and Remote Sensing*, 15:100086, 2025.

[11] Preetam Ghosh, Swalpa Kumar Roy, Bikram Koirala, Behnood Rasti, and Paul Scheunders. Hyperspectral unmixing using transformer network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[12] Scott J. Goetz, Patrick Jantz, and Claire A. Jantz. Connectivity of core habitat in the northeastern united states: Parks and protected areas in a landscape context. *Remote Sensing of Environment*, 113(7):1421–1429, 2009. Monitoring Protected Areas.

[13] Bruce Hapke. Bidirectional reflectance spectroscopy: 1. theory. *Journal of Geophysical Research: Solid Earth*, 86(B4):3039–3054, 1981.

[14] Wei He, Hongyan Zhang, and Liangpei Zhang. Total variation regularized reweighted sparse nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3909–3921, 2017.

[15] John Janiczek, Parth Thaker, Gautam Dasarathy, Christopher S. Edwards, Philip Christensen, and Suren Jayasuriya. Differentiable programming for hyperspectral unmixing using a physics-based dispersion model, 2020. Copyright - © 2020. This work is published under http://arxiv.org/licenses/nonexclusive-distrib/1.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2020-07-15.

[16] Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognition*, 107:107514, 2020.

[17] N. Keshava and J.F. Mustard. Spectral unmixing. *IEEE Signal Processing Magazine*, 19(1):44–57, 2002.

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, 2022. Copyright - © 2022. This work is published under http://arxiv.org/licenses/nonexclusive-distrib/1.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2022-12-14.

[19] Hyperspectral Imaging Laboratory. Samson hyperspectral dataset, 2011. Publicly available dataset commonly used for hyperspectral unmixing studies.

[20] NASA Jet Propulsion Laboratory. AVIRIS jasper ridge hyperspectral dataset. https://aviris.jpl.nasa.gov/, 1995. Accessed: 2025-04-30.

[21] Xiaoqiang Lu, Hao Wu, Yuan Yuan, Pingkun Yan, and Xuelong Li. Manifold regularized sparse nmf for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2815–2826, 2013.

[22] Kiran Mantripragada and Faisal Z. Qureshi. Hyperspectral pixel unmixing with latent dirichlet variational autoencoder. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

[23] U.S. Army Corps of Engineers. Hydice urban hyperspectral dataset, 1995. Captured by the Hyperspectral Digital Imagery Collection Experiment (HYDICE) sensor. Available via academic requests.

[24] Burkni Palsson, Magnus O. Ulfarsson, and Johannes R. Sveinsson. Convolutional autoencoder for spectral–spatial hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):535–549, 2021.

[25] Billy G. Ram, Peter Oduor, C. Igathinathane, Kirk Howatt, and Xin Sun. A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. *Computers and Electronics in Agriculture*, 222:109037, 2024.

[26] Michal Shimoni, Rob Haelterman, and Christiaan Perneel. Hyperspectral imaging for military and security applications: Combining myriad processing and sensing techniques, June 2019. Publisher Copyright: © 2019 IEEE.

[27] Alpana Shukla and Rajsi Kot. An overview of hyperspectral remote sensing and its applications in various disciplines. *IRA-International Journal of Applied Sciences (ISSN 2455-4499)*, 5:85, 12 2016.

[28] Ben Somers, Laurent Tits, and Pol Coppin. Quantifying nonlinear spectral mixing in vegetated areas: Computer simulation model validation and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):1956–1965, 2014.

[29] Mary B. Stuart, Andrew J. S. McGonigle, and Jon R. Willmott. Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems. *Sensors*, 19(14), 2019.

[30] Lingzhi Sun and Paul G. Lucey. Unmixing mineral abundance and mg# with radiative transfer theory: Modeling and applications. *Journal of Geophysical Research: Planets*, 126(2):e2020JE006691, 2021. e2020JE006691 2020JE006691.

[31] Bin Wang, Huizheng Yao, Dongmei Song, Jie Zhang, and Han Gao. Ssf-net: A spatial–spectral features integrated autoencoder network for hyperspectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:1781–1794, 2024.

[32] Xinyu Wang, Yanfei Zhong, Liangpei Zhang, and Yanyan Xu. Spatial group sparsity regularized nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6287–6304, 2017.

[33] Zhiru Yang, Mingming Xu, Shanwei Liu, Hui Sheng, and Jianhua Wan. Ust-net: A u-shaped transformer network using shifted windows for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[34] Alina Zare and K.C. Ho. Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing. *IEEE Signal Processing Magazine*, 31(1):95–104, 2014.

[35] Shaoquan Zhang, Guorong Zhang, Fan Li, Chengzhi Deng, Shengqian Wang, Antonio Plaza, and Jun Li. Spectral-spatial hyperspectral unmixing using nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[36] Min Zhao, Tiande Gao, Jie Chen, and Wei Chen. Hyperspectral unmixing via non-negative matrix factorization with handcrafted and learned priors. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.