Hyperspectral Foundation-Model Feature Upsampling for Unsupervised Unmixing: Training NAF on Frozen DOFA Features

[Tabouri Lukas, Obeye Essadani, Hussein El Gouch, Samy Bali]
Télécom Paris – IMA206 Project

June 18, 2026

Abstract

Hyperspectral foundation models such as DOFA produce semantically rich but spatially coarse feature maps, which is a fundamental obstacle for pixel-level tasks such as spectral unmixing. We study whether a recent, foundation-model-agnostic feature upsampler, Neighborhood Attention Filtering (NAF), can be adapted to the hyperspectral setting and trained to recover high-resolution features suitable for unmixing. Concretely, we keep both the DOFA encoder and a SpectralEarth spectral adapter frozen, and train *only* the NAF module, its dual-branch guidance encoder and cross-scale neighborhood attention, with a self-supervised feature-reconstruction objective: NAF learns to recover DOFA's own high-resolution feature map from a spatially sub-sampled copy of it, supervised by a single MSE term. This keeps the whole pipeline label-free and therefore directly compatible with unsupervised unmixing. We describe the architecture, the training objective, and an evaluation protocol on hyperspectral data, and report the behaviour of the upsampled features both quantitatively and qualitatively.

1 Use of generative AI

In accordance with the project guidelines, this section details how generative AI was used during the project and the writing of this report, with concrete examples.

During the project . Generative AI (Anthropic's Claude) was used as a programming and comprehension assistant:

- **Reading the source papers.** To clarify specific mechanisms of NAF [1] and DOFA [2]. For instance the exact roles of the queries, keys and values in the cross-scale neighborhood attention, we asked targeted questions and cross-checked the answers against the original papers.
- **Debugging.** For example : resolving channel-dimension mismatches between the spectral adapter (128 channels) and the NAF guidance encoder

All design decisions, the integration of the three code bases (valeoai/NAF, DOFA, SpectralEarth), the experiments, and the analysis of the numerical results were performed and validated by us.

2 Introduction

Scientific context. A hyperspectral image (HSI) records, at every pixel, a near-continuous reflectance spectrum sampled over hundreds of narrow contiguous bands. This dense spectral sampling enables fine material discrimination, but it comes with two well-known trade-offs: a high spectral dimensionality, and a comparatively coarse *spatial* resolution. As a consequence, the radiance measured at a pixel is, in general, a mixture of the signatures of several pure

materials (*endmembers*) present within the ground footprint of that pixel. *Spectral unmixing* is the inverse problem of recovering, for each pixel, both the endmember signatures and their fractional abundances. It is a dense, per-pixel task: the quantity of interest lives at the native spatial resolution of the image.

Foundation models and the resolution gap. Recent *hyperspectral foundation models* (HSFMs) learn, by large-scale self-supervised pre-training, transferable representations of HSIs. DOFA [2] is one such model: a wavelength-conditioned dynamic hypernetwork generates the patch-embedding weights, allowing a single Vision Transformer [3] backbone to ingest images with an arbitrary number of spectral bands. Like any ViT-based encoder, DOFA tokenises the image into patches and therefore outputs a feature map at a *reduced* spatial resolution: from an input $Y \in \mathbb{R}^{B \times H_{\mathrm{HR}} \times W_{\mathrm{HR}}}$ it produces features $F \in \mathbb{R}^{d \times H_{\mathrm{LR}} \times W_{\mathrm{LR}}}$ with $H_{\mathrm{LR}} \ll H_{\mathrm{HR}}$. These features are excellent semantic descriptors but, being spatially coarse, are not directly usable for a per-pixel task such as unmixing.

Feature upsampling. The natural remedy is to *upsample* the features back to (or towards) the image resolution. Rather than a fixed interpolation, we use a learnable, image-guided upsampler: NAF [1]. NAF reconstructs each high-resolution feature as a content- and position-aware weighted combination of the low-resolution features in a small spatial neighbourhood, using Cross-Scale Neighborhood Attention [4] and rotary position embeddings (RoPE) [5], with the high-resolution *image* as the only guidance signal.

Objective and contribution. The goal of this project is to adapt NAF to the hyperspectral setting and to train it, on top of frozen DOFA features, so that the resulting high-resolution features support unmixing. Our specific choices are:

- **What is frozen, what is trained.** DOFA and the SpectralEarth spectral adapter [6] are kept frozen; *only* the NAF module is trained. This isolates the upsampler as the sole object of study and keeps training light.
- **A label-free objective.** NAF is trained by self-distillation: it must reproduce DOFA's own high-resolution features from a sub-sampled version of them, under a single mean-squared-error term in feature space. The pipeline therefore needs no unmixing ground truth and remains compatible with *unsupervised* unmixing.

3 Background and related work

3.1 Spectral unmixing and the linear mixing model

Let $Y \in \mathbb{R}^{B \times P}$ collect the spectra of P pixels over B bands. Under the *linear mixing model* (LMM), each observed spectrum is a non-negative combination of R endmember signatures:

$$Y = A\,S + N, \qquad A \in \mathbb{R}^{B \times R}, \;\; S \in \mathbb{R}^{R \times P}, \tag{1}$$

where the columns of A are the endmember spectra, S holds the per-pixel abundances, and N is additive noise. Physically meaningful solutions satisfy the *abundance non-negativity constraint* (ANC), $S \geq 0$, and the *abundance sum-to-one constraint* (ASC), $\mathbf{1}^\top S = \mathbf{1}^\top$. Equation (1) together with $A, S \geq 0$ is exactly a non-negative matrix factorisation, which links classical unmixing to the NMF tools seen in class; modern variants replace the bilinear factorisation by an auto-encoder whose decoder is linear (its weights play the role of A) and whose latent code is constrained to the simplex (playing the role of S). A recurring difficulty is that spectral information alone is often insufficient to disambiguate mixtures; incorporating *spatial* context, which is precisely what spatially resolved deep features provide, improves abundance estimation.

3.2 DOFA as a hyperspectral feature extractor

DOFA [2] is a multimodal Earth-observation foundation model. Its key idea is a wavelength-conditioned dynamic hypernetwork that, given the central wavelengths of the input bands, *generates* the weights of the patch-embedding layer. A single shared ViT [3] backbone can therefore process inputs with heterogeneous spectral configurations, including band counts unseen during pre-training. DOFA is pre-trained with a masked-image-modelling objective combined with a distillation strategy. In our pipeline DOFA is used purely as a frozen feature extractor: it maps an input HSI to the low-resolution feature map F that becomes the *values* of the attention (Section 4).

3.3 The spatial-resolution gap and feature upsampling

Feature upsampling addresses the gap between coarse ViT features and the dense outputs required by pixel-level tasks. Parameter-free interpolation (nearest, bilinear) is fast but content-agnostic and over-smooths boundaries; joint/guided filters use the high-resolution image to sharpen edges; learnable upsamplers achieve higher accuracy but are often tied to a specific backbone and must be retrained for each one. NAF [1] was introduced to break this trade-off: it is foundation-model-agnostic and image-guided, treating upsampling as a learned, adaptive filtering operation. We rely on NAF as the trainable core of our system.

3.4 Building blocks

Vision Transformer. The ViT [3] splits an image into non-overlapping patches, linearly embeds them into tokens, and applies global self-attention. Patchification is the origin of the spatial downsampling that motivates this work.

Rotary position embedding (RoPE). RoPE [5] encodes *relative* positions by rotating query and key vectors as a function of their coordinates, so that the attention score depends on the displacement between two locations. Its two-dimensional extension is well suited to image grids and to the cross-scale matching between high- and low-resolution coordinates.

Neighborhood attention. Neighborhood Attention [4] restricts each query to a local window instead of attending globally, which makes attention linear in the number of tokens and aligns naturally with local interpolation. NAF builds on a *cross-scale* variant of this mechanism.

Spectral adapter. The SpectralEarth spectral adapter [6] is a lightweight module that preprocesses the hundreds of HSI bands into a fixed-width representation, so that standard backbones can handle hyperspectral inputs. We use it to map an arbitrary number of input bands to the fixed channel count expected by the NAF guidance encoder.

4 Method

4.1 Pipeline overview

Given a high-resolution HSI $Y \in \mathbb{R}^{B \times H_{\mathrm{HR}} \times W_{\mathrm{HR}}}$, the system produces high-resolution features $F_{\mathrm{up}} \in \mathbb{R}^{d \times H_{\mathrm{HR}} \times W_{\mathrm{HR}}}$ (or at a finer intermediate resolution with scale factor s) in four stages:

1. **Frozen feature extraction.** DOFA maps Y to a feature map that provides the attention *values* V_{LR}. At training time these values are a spatially sub-sampled copy of DOFA's own high-resolution features, which also serve as the regression target (Section 4.4); at inference the learned operator is applied to DOFA's native features to upsample them.

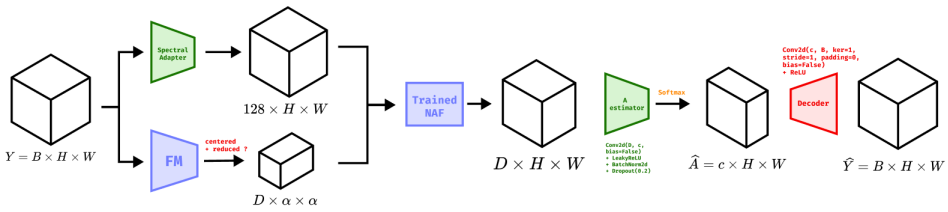

Figure 1: Project Pipeline

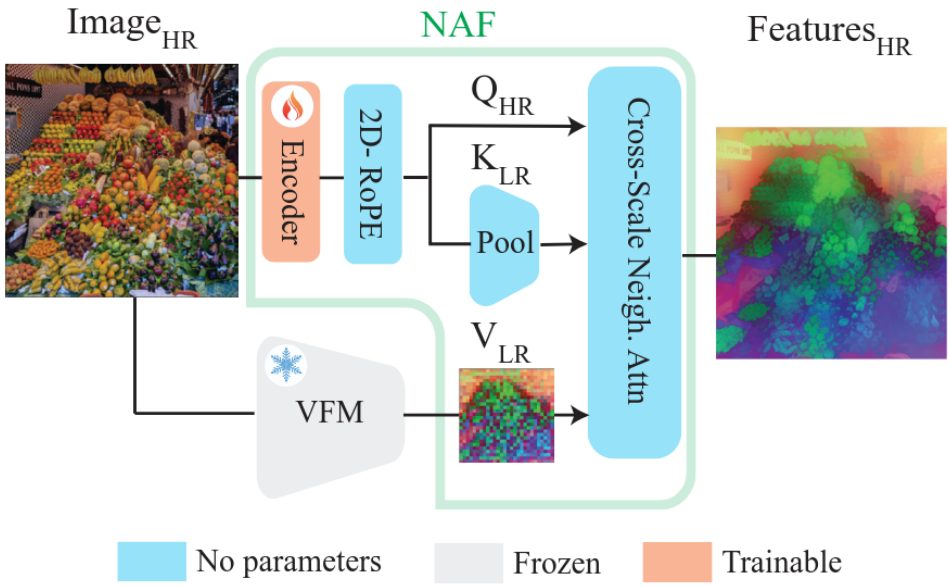

Figure 2: Naf components

2. **Guidance encoding.** The spectral adapter maps Y to 128 channels; the NAF dual-branch guidance encoder turns this into a high-resolution guidance map $\mathrm{Enc}_\theta(Y) \in \mathbb{R}^{C \times H_{\mathrm{HR}} \times W_{\mathrm{HR}}}$.
3. **Query/key construction.** RoPE is applied to the guidance map to obtain the high-resolution *queries* Q_{HR}; average pooling of the guidance map to the low resolution yields the *keys* K_{LR}.
4. **Cross-scale neighborhood attention.** For each high-resolution location, attention over a compact neighbourhood of its low-resolution counterpart predicts F_{up}.

Only the parameters θ of the NAF module (guidance encoder + attention) are trained; the spectral adapter and DOFA are frozen.

4.2 Spectral adaptation and the dual-branch guidance encoder

The NAF guidance encoder is a dual-branch convolutional network: a *pixel-encoding* branch that captures local, fine-grained detail, and a *contextual-encoding* branch that aggregates a wider receptive field; each branch is a stack of L convolutional layers, and the two outputs are concatenated along the channel dimension. Because this encoder expects a *fixed* number of input channels, the raw HSI (B bands, B varying across sensors and scenes) cannot be fed directly. We therefore insert the frozen SpectralEarth spectral adapter upstream, which maps the B bands to a fixed 128-channel tensor; the guidance encoder then operates on these 128 channels. This keeps the guidance pathway sensor-agnostic while leaving only the upsampler to be trained.

4.3 Cross-scale neighborhood attention with RoPE

Let p be a location on the high-resolution grid and $\mathcal{N}(p)$ the set of low-resolution locations in a compact neighbourhood of its projection onto the low-resolution grid. The upsampled feature is computed as an attention-weighted combination of the corresponding values:

$$F_{\mathrm{up}}(p) \;=\; \sum_{q\in\mathcal{N}(p)} \alpha_{pq}\, V_{\mathrm{LR}}(q), \qquad \alpha_{pq} \;=\; \mathrm{softmax}_{q\in\mathcal{N}(p)}\!\left(\frac{Q_{\mathrm{HR}}(p)^{\top} K_{\mathrm{LR}}(q)}{\sqrt{C}}\right), \tag{2}$$

where Q_{HR} and K_{LR} are obtained from the guidance map as described above and carry RoPE so that α_{pq} depends on the relative displacement between p and q. Equation (2) is exactly a learned, content- and position-adaptive interpolation kernel: each high-resolution query attends only to a small neighbourhood around its low-resolution location, which aligns receptive fields across the two scales while keeping attention local and efficient [1, 4]. Crucially, the *values* are the frozen DOFA features, so the upsampler never needs to model the foundation model's feature distribution—it only learns where to look.

4.4 Training objective: feature self-distillation

NAF is trained with a single mean-squared-error term in *feature space*, using DOFA itself to provide the high-resolution target. The idea is to turn upsampling into a self-supervised regression: we never need any external (image or unmixing) ground truth, only DOFA's own features.

Let $F_{\mathrm{HR}} = \mathrm{DOFA}(Y) \in \mathbb{R}^{d\times 14\times 14}$ be the feature map produced directly by the frozen encoder, and let $\downarrow$ denote a fixed spatial sub-sampling operator that halves the resolution, so that $F_{\mathrm{LR}} = \downarrow F_{\mathrm{HR}} \in \mathbb{R}^{d\times 7\times 7}$. We feed F_{LR} to NAF as the low-resolution values V_{LR}, and ask it to reconstruct the 14×14 map. The loss is the pixel-wise MSE between this reconstruction and the original DOFA features:

$$\mathcal{L}(\theta) \;=\; \Big\| \underbrace{\mathrm{NAF}_{\theta}\big(\downarrow F_{\mathrm{HR}};\, Y\big)}_{F_{\mathrm{up}}\in\mathbb{R}^{d\times 14\times 14}} \;-\; \underbrace{F_{\mathrm{HR}}}_{\text{target}} \Big\|_2^2, \tag{3}$$

where the second argument Y recalls that the queries and keys are derived from the image guidance (Section 4.3). The target F_{HR} is treated as a constant (no gradient flows back into DOFA), and gradients update only θ.

This is a self-distillation scheme: DOFA's native output at 14×14 acts as the teacher, and NAF must recover it from a 7×7 sub-sampled copy. Because no high-resolution *ground-truth* feature exists beyond DOFA's own resolution, simulating the upsampling task by first downsampling is the only way to obtain a supervised target; the operator NAF learns this way is image-guided and resolution-agnostic, so at inference it can be applied to DOFA's native features to reach resolutions *finer* than 14×14. The objective uses no labels, which is precisely what keeps the downstream unmixing *unsupervised*.

4.5 From upsampled features to unmixing

Once trained, the upsampler yields a per-pixel feature $F_{\mathrm{up}}(p) \in \mathbb{R}^d$ at (close to) native resolution. We obtain abundances by using a lightweight unmixing neural network.

4.6 Training configuration: the 7→14 protocol

Training instances are built directly from DOFA features at a working resolution of 14×14. For each image Y we (i) run DOFA once to obtain the 14×14 feature map F_{HR}; (ii) spatially sub-sample it to 7×7 to obtain F_{LR}; (iii) pass F_{LR} to NAF as low-resolution values, together with the image-derived guidance, so that NAF reconstructs a 14×14 map; and (iv) back-propagate the MSE of Eq. (3) between this reconstruction and F_{HR}. This corresponds to an upsampling factor $s = 2$ ($7\times 7 \to 14\times 14$). Only θ is updated.

5 Experimental protocol

5.1 Data

Training is performed on EnMAP hyperspectral imagery from SpectralEarth [6] (patch size, number of bands B, spatial resolution, normalisation statistics). Evaluation of the upsampled features for unmixing is carried out on Apex dataset, or held-out EnMAP scene, with reference endmembers/abundances obtained from this dataset.

5.2 Implementation details

Frozen backbones. The feature extractor is the **DOFA-Large** model (ViT-Large backbone) [2]; its output therefore has feature dimension $d = 1024$, which is the channel dimension of the values V_{LR} and of the reconstructed map F_{up}. The guidance pathway uses the spectral adapter taken from the SpectralEarth `spec_ViTb_mae` checkpoint [6] (a Spectral ViT-Base pre-trained on EnMAP with masked auto-encoding); it is kept frozen and maps the input HSI bands to the 128 channels consumed by the NAF dual-branch guidance encoder. Both backbones are frozen, so the only trainable parameters are those of NAF.

5.3 Evaluation metrics

We report:

- **Feature reconstruction MSE** between F_{up} and the target DOFA features (the quantity of Eq. (3)), as the primary measure of upsampling quality; we evaluate it on held-out images and, for an upsampler, on the 7→14 self-supervised task;
- **Spectral Angle Distance** (SAD) between estimated and reference endmembers (lower is better);
- **Abundance RMSE** between estimated and reference abundance maps;
- **Qualitative feature diagnostics**, e.g. PCA of F vs. F_{up}, to show that detail is recovered rather than merely smoothed.

6 Results and discussion

We evaluate on an urban hyperspectral scene with $R = 4$ endmembers, comparing our NAF upsampler against bilinear upsampling of the same frozen DOFA-Large features. All abundance and endmember plots are collected in the appendix.

6.1 Quantitative results

Table 1 reports the mean metrics; Table 2 gives the per-endmember breakdown.

Table 1: Unmixing metrics, averaged over the four endmembers. Lower is better; best in bold. The two methods share the same DOFA-Large features and unmixing head, and differ only in the upsampler.

Upsampler	Endmember SAD ↓	Abundance NMSE ↓
Bilinear (baseline)	0.099	0.147
NAF (ours)	**0.072**	**0.143**

Table 2: Per-endmember SAD and abundance NMSE for the two upsamplers

Metric	Upsampler	EM1	EM2	EM3	EM4
2*SAD ↓	Bilinear	0.14	0.14	0.12	0.05
	NAF	0.10	0.14	0.10	0.04
2*NMSE ↓	Bilinear	0.99	0.05	0.16	0.12
	NAF	0.98	0.06	0.15	0.10

Endmembers: a clear and consistent gain. NAF lowers the mean SAD from 0.099 to 0.072, a relative reduction of about 27%. The breakdown shows where this comes from: the two endmembers the baseline recovered worst among the non-trivial cases improve the most, EM1 ($0.14 \rightarrow 0.10$) and EM3 ($0.12 \rightarrow 0.10$), while the already excellent EM4 improves marginally ($0.05 \rightarrow 0.04$) and EM2 is unchanged (0.14). The interpretation is consistent with the role of an image-guided upsampler: by restoring spatial high-frequency content that bilinear interpolation smooths away, NAF helps the head isolate purer pixels, sharpening exactly the endmember spectra that were the most biased and saturating on those already well estimated.

Abundances: a near-tie dominated by one endmember. The mean abundance NMSE barely moves ($0.147 \rightarrow 0.143$, $\approx 3\%$ relative), and the per-endmember picture is mixed: EM4 improves ($0.12 \rightarrow 0.10$), EM3 marginally ($0.16 \rightarrow 0.15$), EM2 marginally degrades ($0.05 \rightarrow 0.06$) and EM1 is essentially unchanged ($0.99 \rightarrow 0.98$). The aggregate is entirely governed by EM1, whose NMSE near 1 means the predicted map explains virtually none of the variance of the reference. This is consistent with EM1 being a very low-energy endmember: its abundances are tiny (the colour bars saturate at 0.005–0.010, Figure 6,5), so the NMSE, normalised by the reference energy, is ill-conditioned and saturates regardless of the upsampler. Crucially, EM1's *spectrum* is recovered with moderate and *improving* SAD ($0.14 \rightarrow 0.10$): the failure is therefore in the spatial localisation performed by the unmixing head, not in the resolution of the features, which is exactly why both upsamplers fail equally on it. The bottleneck for the overall NMSE is the unmixing stage, not NAF.

6.2 Qualitative results

Abundance maps (Figure 5, 6). The NAF maps are visibly sharper than the bilinear ones: the road network and the building blocks have crisp, well-localised boundaries that follow the ground truth closely, whereas the bilinear maps are diffuse and blurred. (The two figures use different colour-bar ranges, 0.010 vs. 0.005, so the comparison is on spatial *structure*, not absolute magnitude.) That this clear visual improvement coincides with an almost unchanged global NMSE is itself informative: an ℓ_2 map metric rewards an energy match far more than edge fidelity, and is here capped by EM1.

Endmember spectra (Figure 3, 4). EM4 is recovered almost exactly (SAD 0.04); its single sharp short-wavelength peak is characteristic of water, consistent with the river-shaped EM4 abundance map. EM2 is the hardest (SAD 0.14, unchanged): the predicted curve underestimates the amplitude of its main peak around the mid-band region and the secondary structure that follows. For EM1 and EM3, NAF visibly tightens the predicted curve onto the reference, matching the SAD gains in Table 2.

6.3 Limitations and ablations

The main limitation surfaced here is downstream of the upsampler: the unmixing head fails to localise the low-energy endmember EM1, capping the achievable abundance NMSE for both

methods. A spatially regularised or sparsity-aware head is the natural next step.

7 Conclusion and perspectives

We presented a pipeline that adapts the NAF feature upsampler to hyperspectral imagery and trains it, and only it, on top of a frozen DOFA encoder and a frozen SpectralEarth spectral adapter, using a self-supervised feature self-distillation objective (a single MSE between NAF's reconstruction and DOFA's own high-resolution features). By turning coarse foundation-model features into spatially resolved ones without any label, the approach is naturally suited to unsupervised unmixing.

Promising extensions include coupling the upsampler with a spatially regularised unmixing head and studying robustness to the number of input bands across sensors.

References

[1] Loick Chambon, Paul Couairon, Eloi Zablocki, Alexandre Boulch, Nicolas Thome, and Matthieu Cord. NAF: Zero-shot feature upsampling via neighborhood attention filtering. *arXiv preprint arXiv:2511.18452*, 2025.

[2] Zhitong Xiong, Yi Wang, et al. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

[3] 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. In *International Conference on Learning Representations (ICLR)*, 2021.

[4] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 2024. arXiv:2104.09864.

[6] Nassim Ait Ali Braham, Conrad M. Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. SpectralEarth: Training hyperspectral foundation models at scale. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025. arXiv:2408.08447.

Appendix: figures

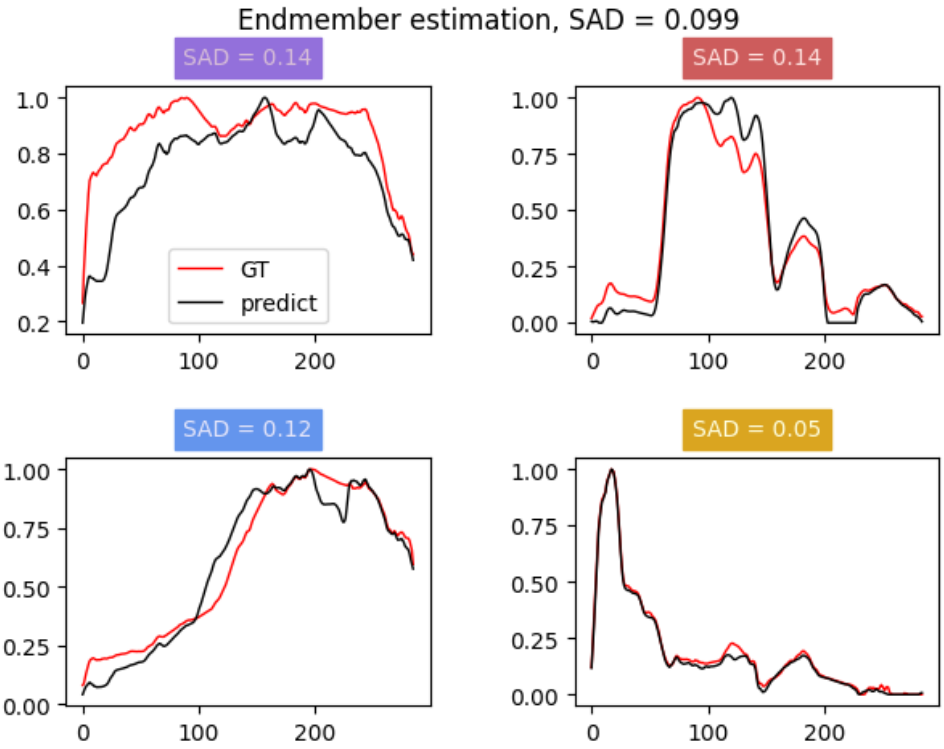

Figure 3: endmember DOFA+bilinear

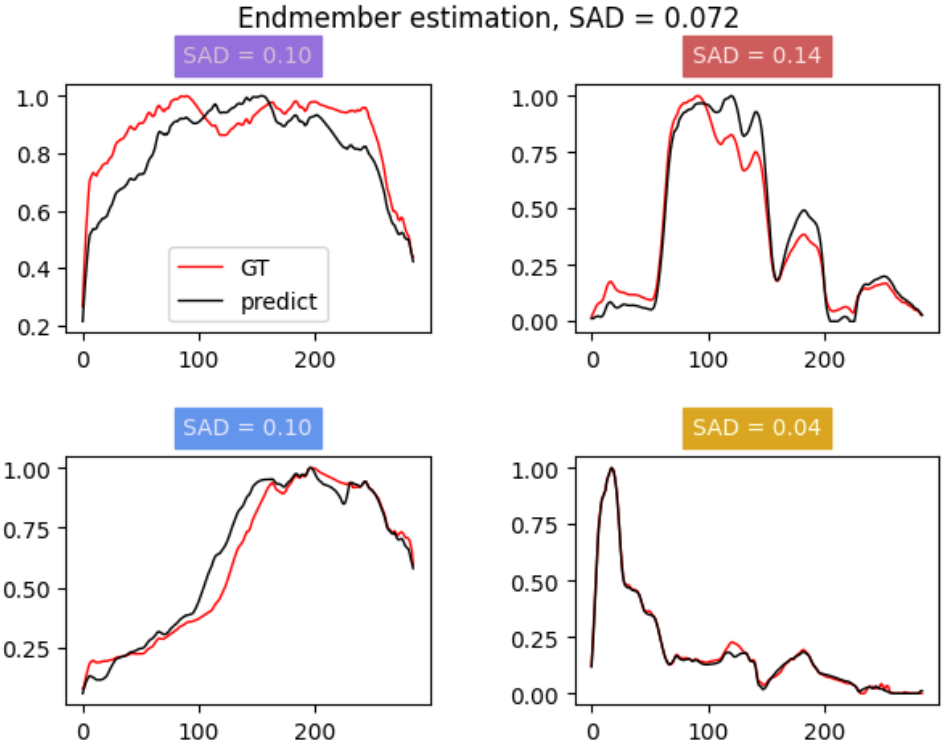

Figure 4: endmember NAF

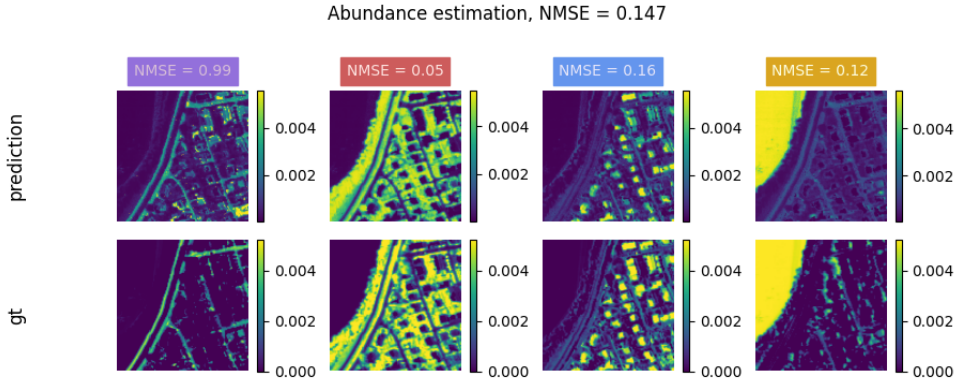

Figure 5: abundance DOFA+bilinear

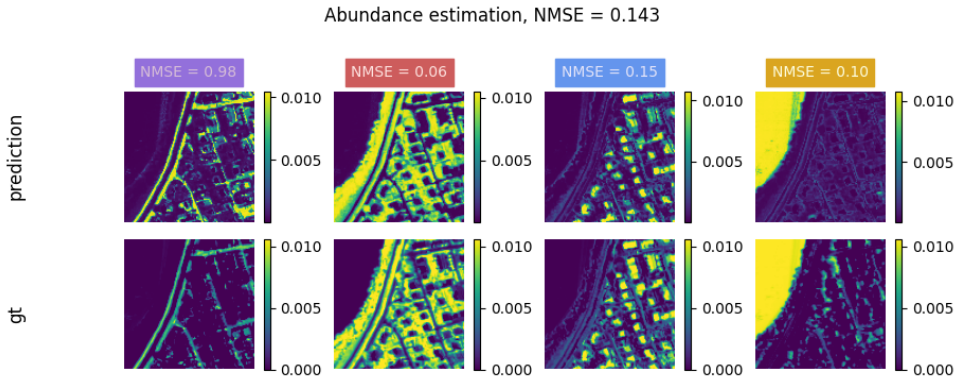

Figure 6: abundance NAF