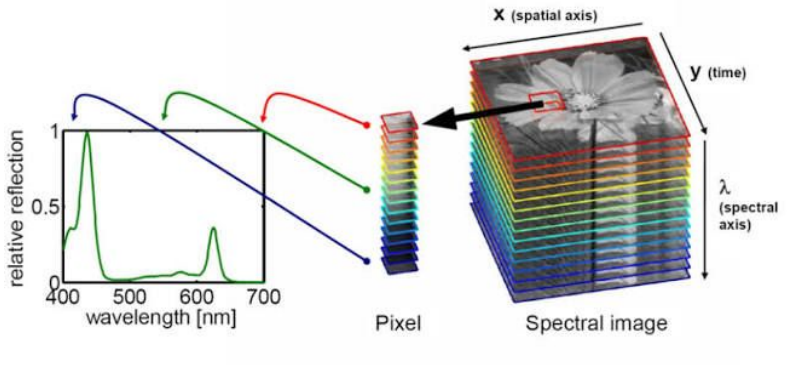
Hyperspectral Foundation-Model Feature Upsampling for Unsupervised Unmixing: Training NAF on Frozen DOFA Features

TABOURI Lukas, ESSAADANI Obeye, EL GOUCH Hussein, BALI Samy

Imagerie hyperspectrale & démélange spectral

- Un pixel hyperspectral = un spectre quasi-continu sur des centaines de bandes étroites.
- Compromis : forte dimension spectrale, mais résolution spatiale grossière.
- Conséquence : pixels mixtes mélange de plusieurs matériaux purs (endmembers).



Problème inverse modèle de mélange linéaire

$$Y = AS + N, \quad A \in \mathbb{R}^{B \times R}, \quad S \in \mathbb{R}^{R \times P}$$

$$S \geq 0 \ \text{(ANC)} \qquad \mathbf{1}^\top S = \mathbf{1}^\top \ \text{(ASC)}$$

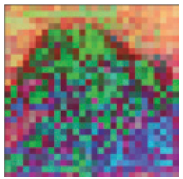
Colonnes de A = spectres ; S = abondances par pixel. Exactement une NMF.

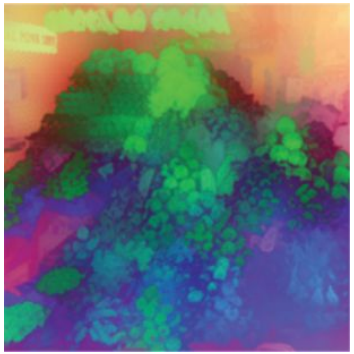
Foundation models & le fossé de résolution

DOFA : hypernetwork conditionné par la longueur d'onde qui génère les poids du patch-embedding → un seul ViT pour un nombre arbitraire de bandes.

Comme tout ViT, DOFA patchifie l'image → carte de features à résolution réduite :

$$F \in \mathbb{R}^{d \times H_{\text{LR}} \times W_{\text{LR}}}, \quad H_{\text{LR}} \ll H_{\text{HR}}$$





Sémantiquement riches mais **spatialement grossières** → inutilisables telles quelles pour une tâche par-pixel comme le démélange.

NAF

VFM : La résolution spatiale est réduite, on a une perte de précision. Comment y remédier ?

NAF : premier module d'upsampling totalement agnostique (indépendant du VFM utilisé) capable de surpasser les méthodes spécifiques. Il fonctionne en **"zero-shot"**, c'est-à-dire qu'une fois entraîné, il peut s'appliquer instantanément à n'importe quel modèle de vision sans aucun réentraînement

Pourquoi NAF ? **Filtres classiques :** résultats floues et indépendants du contexte

Upsamplers modernes : JAFAR, FeatUp sont spécifiques à chaque VFM, il faut donc réentraîner le module d’upsampling pour chaque VFM

OBJECTIF & CONTRIBUTION

Adapter NAF à l'hyperspectral et l'entraîner sur features DOFA gelées

Ce qui est gelé

DOFA + l'adaptateur spectral SpectralEarth restent figés.

Ce qui est entraîné

Seul le module NAF (encodeur de guidage + attention).

Objectif sans label

Self-distillation : un seul terme MSE en espace de features.

Pipeline entièrement label-free → directement compatible avec un démélange non-supervisé.

Trois composants, deux gelés, un entraîné

DOFA extracteur de features (gelé)

Mappe l'HSI vers la carte basse résolution F, qui fournit les valeurs V_LR de l'attention.

Adaptateur spectral SpectralEarth (gelé)

Mappe B bandes (variables selon capteur) vers un tenseur fixe de 128 canaux → voie de guidage agnostique au capteur.

NAF upsampler image-guidé (entraîné)

Agnostique au foundation model : upsampling vu comme un filtrage adaptatif appris, guidé par l'image HR.

Attention de voisinage cross-échelle + RoPE

Pour chaque position HR p, on attend sur un petit voisinage N(p) de sa projection basse résolution :

$$F_{\text{up}}(p) = \sum_{q \in \mathcal{N}(p)} \alpha_{pq} V_{\text{LR}}(q)$$

$$\alpha_{pq} = \text{softmax}_{q \in \mathcal{N}(p)} \left(\frac{Q_{\text{HR}}(p)^{\top} K_{\text{LR}}(q)}{\sqrt{C}} \right)$$

Q, K portent RoPE → *α dépend du déplacement relatif p–q.*

Valeurs = features DOFA gelées : l'upsampler n'apprend que « où regarder », jamais à modéliser la distribution des features.

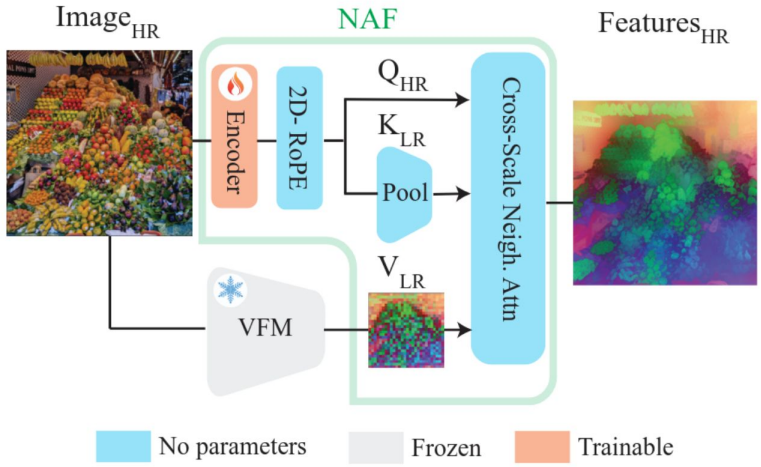
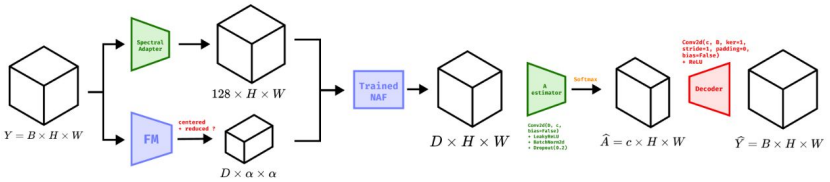


Figure 2: Naf components

VUE D'ENSEMBLE

La pipeline en quatre étapes



1. **Extraction gelée.** DOFA ⟶ carte de features = valeurs V_LR de l'attention.

2. **Encodage de guidage.** Adaptateur ⟶ 128 canaux ; encodeur dual-branch ⟶ carte de guidage HR.

3. **Construction Q / K.** RoPE sur le guidage ⟶ Q_HR ; average-pooling LR ⟶ K_LR.

4. **Attention cross-échelle.** Pour chaque position HR, attention sur un voisinage LR ⟶ F_up.

Seuls les paramètres θ de NAF (encodeur de guidage + attention) sont mis à jour ; DOFA et l'adaptateur spectral restent gelés.

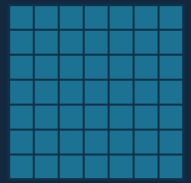
OBJECTIF D'ENTRAÎNEMENT

Self-distillation des features : protocole 7 → 14

Un seul terme MSE en espace de features

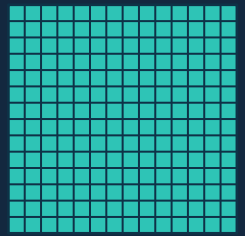
$$\mathcal{L}(\theta) = \| \mathrm{NAF}_\theta(\downarrow F_{\mathrm{HR}};\, Y) \;-\; F_{\mathrm{HR}} \|_2^2$$

DOFA (14×14) = professeur ; cible figée, aucun gradient vers DOFA.



F_LR 7×7

NAF (s = 2)



F_up 14×14

Pourquoi sous-échantillonner ?

Aucune feature HR n'existe au-delà de DOFA → on simule l'upsampling en partant d'une copie sous-échantillonnée.

Opérateur image-guidé & résolution-agnostique → à l'inférence on dépasse 14×14.

Zéro label → démélange aval non-supervisé.

Données, backbones gelés & métriques

Données

- Entraînement : EnMAP (SpectralEarth).
- Évaluation : images avec endmembers & abondances de référence.

Backbones gelés

- DOFA-Large (ViT-L) ⟶ d = 1024.
- Adaptateur spec_ViTb_mae ⟶ 128 canaux.
- Seuls les poids de NAF sont entraînables.

Métriques d'évaluation

MSE de reconstruction entre F_up et la cible DOFA (7⟶14) **SAD ↓** angle spectral endmembers estimés / référence.

Abundance RMSE/NMSE ↓ cartes d'abondance ; **PCA de F vs F_up** diagnostic qualitatif (détail récupéré, pas lissé).
Baseline : upsampling bilinéaire.

NAF vs bilinéaire — mêmes features, même tête

−27%

SAD endmembers

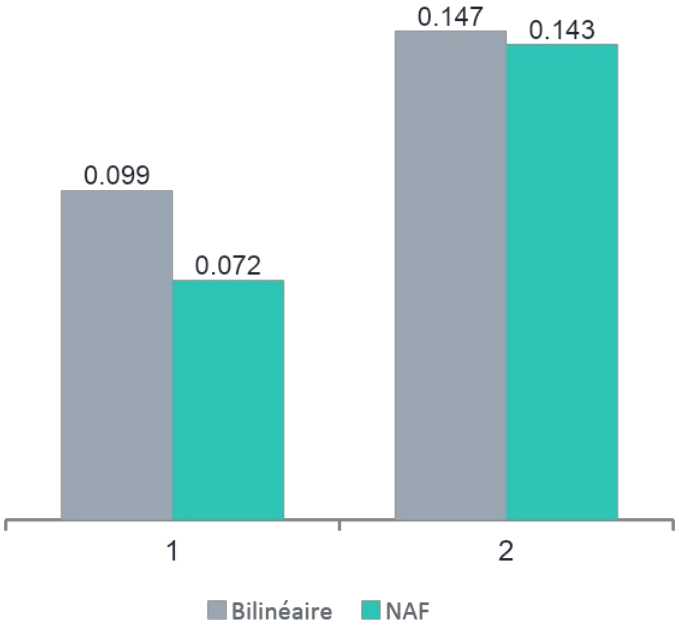
0.099 → 0.072

≈ 3%

NMSE abondances

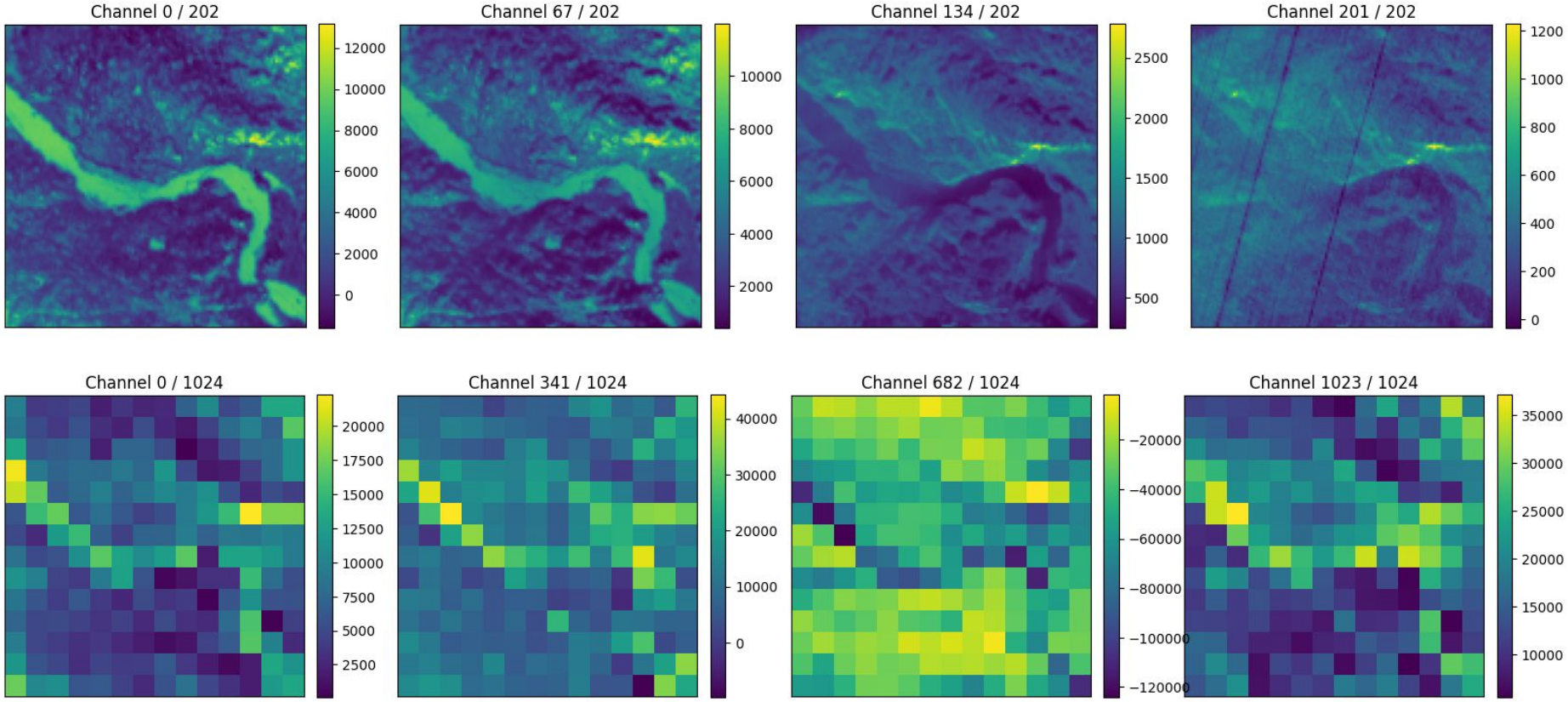
0.147 → 0.143

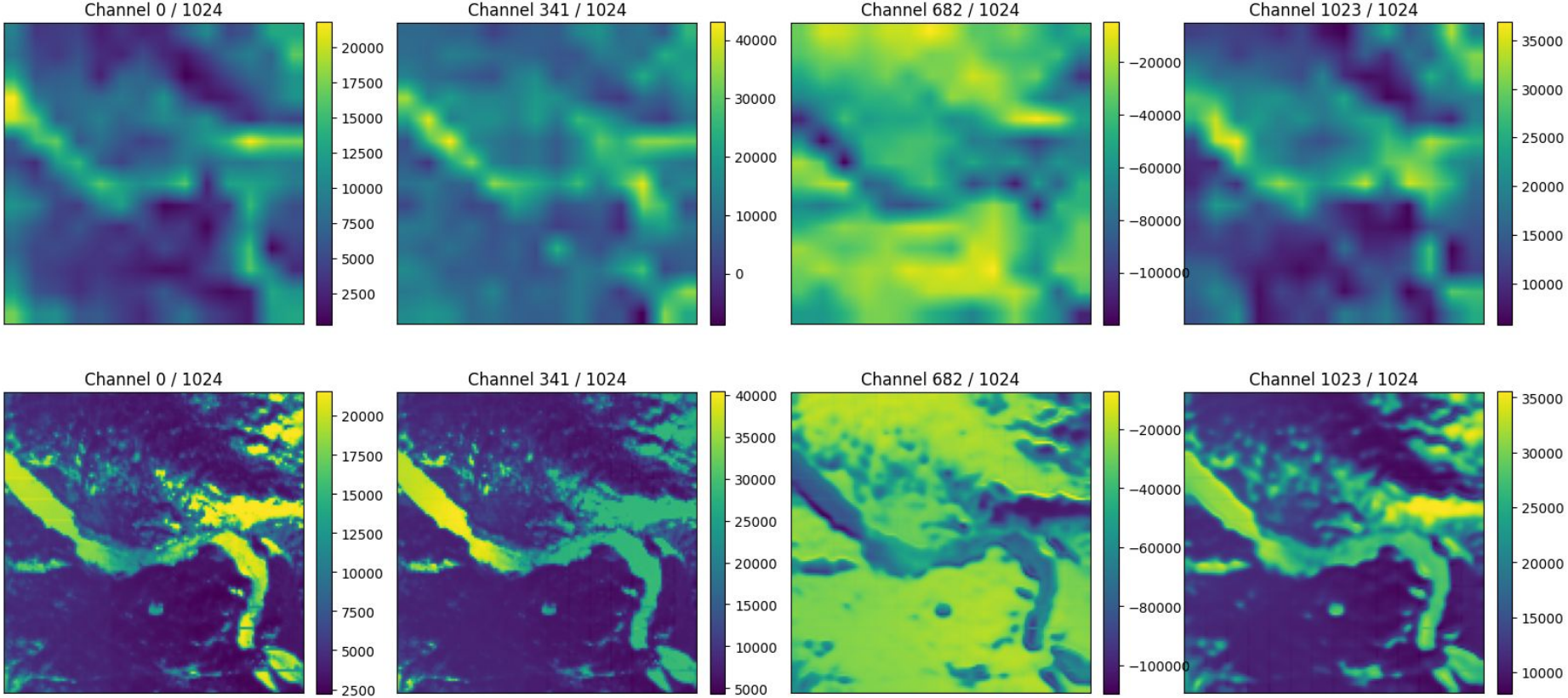
Gain net et cohérent sur les spectres ; quasi-égalité sur les abondances, dominée par un seul endmember (EM1).

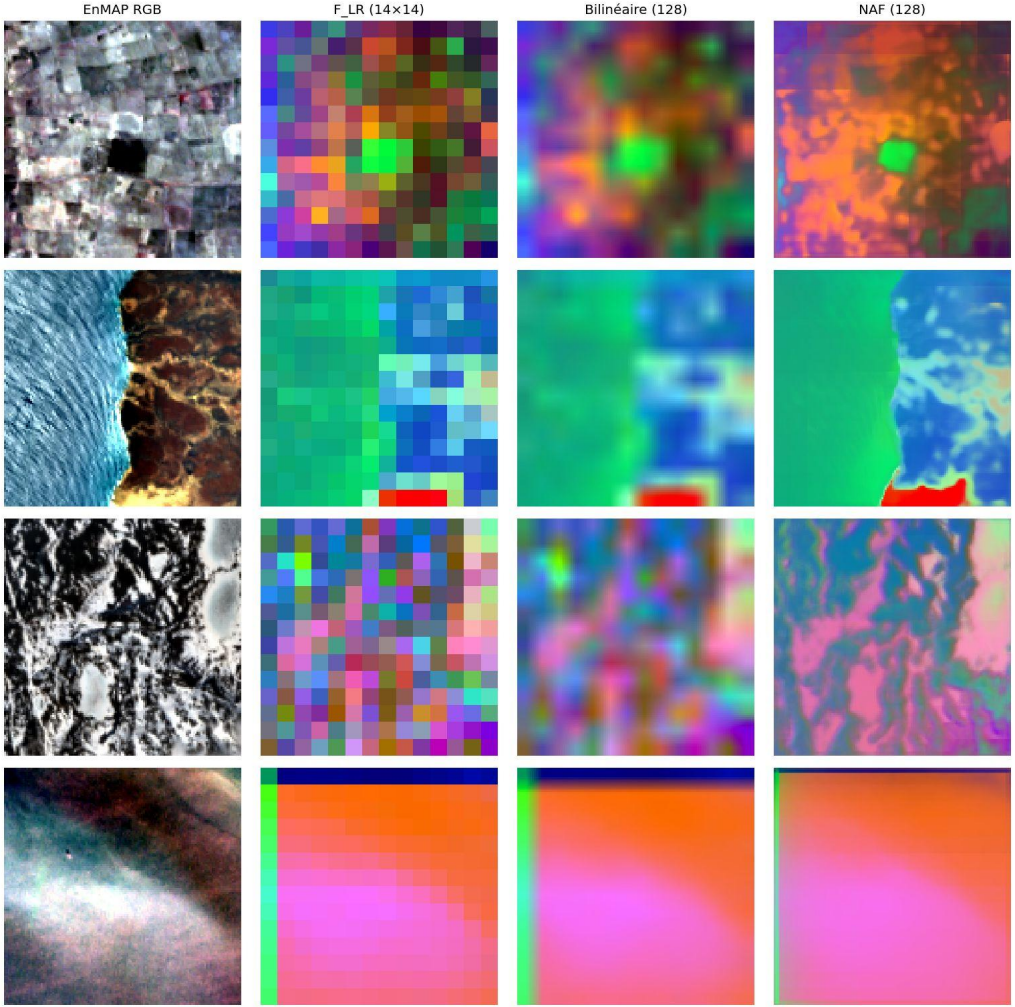


Détail par endmember

	EM1	EM2	EM3	EM4
SAD · Bilin	0.14	0.14	0.12	0.05
SAD · NAF	**0.10**	0.14	**0.10**	**0.04**
NMSE · Bilin	0.99	0.05	0.16	0.12
NMSE · NAF	0.98	0.06	**0.15**	**0.10**


Channel 0 / 202
12000
10000
8000
6000
4000
2000
0
Channel 67 / 202
10000
8000
6000
4000
2000
Channel 134 / 202
2500
2000
1500
1000
500
Channel 201 / 202
1200
1000
800
600
400
200
0
Channel 0 / 1024
20000
17500
15000
12500
10000
7500
5000
2500
Channel 341 / 1024
40000
30000
20000
10000
0
Channel 682 / 1024
−20000
−40000
−60000
−80000
−100000
−120000
Channel 1023 / 1024
35000
30000
25000
20000
15000
10000


Channel 0 / 1024
Channel 341 / 1024
Channel 682 / 1024
Channel 1023 / 1024
Channel 0 / 1024
Channel 341 / 1024
Channel 682 / 1024
Channel 1023 / 1024


EnMAP RGB
F_LR (14×14)
Bilinéaire (128)
NAF (128)

Cartes plus nettes & spectres resserrés

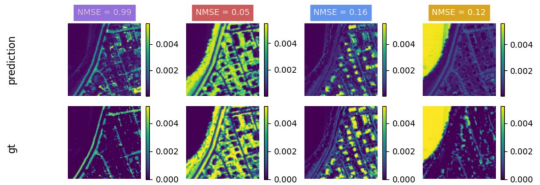


Figure 5: abundance DOFA+bilinear

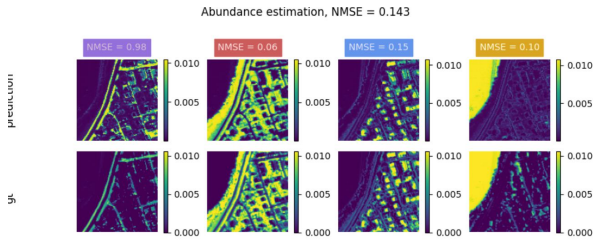


Figure 6: abundance NAF

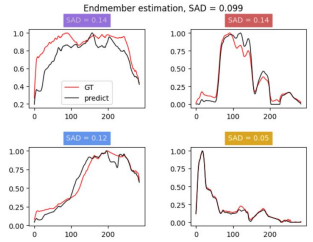


Figure 3: endmember DOFA+bilinear

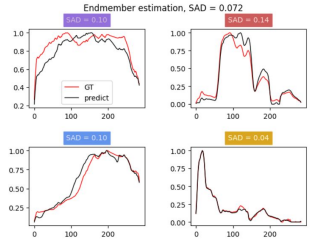


Figure 4: endmember NAF

Cartes d'abondance

Réseau routier & blocs de bâtiments nets et bien localisés ; le bilinéaire reste flou.

Spectres

EM4 ≈ eau (pic court, SAD 0.04). EM2 le plus dur (0.14). EM1/EM3 resserrés sur la référence.

BILAN & PERSPECTIVES

Un upsampling sans label, fait pour le démélange non-supervisé

Conclusion

NAF entraîné sur DOFA gelé, par self-distillation MSE.

Des features grossières deviennent spatialement résolues, sans aucun label.

Perspectives

Entrainer NAF avec plus d'images hyperspectrales (ce qui demande plus de ressources)

Utiliser d'autres methodes d'entrainement de NAF (par exemple crop & stitch)