MASK COMPONENTS, NOT BANDS: RETHINKING MASKED AUTOENCODING FOR HYPERSPECTRAL IMAGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperspectral imagery (HSI) captures hundreds of contiguous wavelengths per pixel, enabling material identification through spectral differences, but labeled data is scarce. Masked autoencoder (MAE) pretraining has emerged as the dominant self-supervised approach for learning HSI representations from unlabeled data. Existing HSI MAE methods adopt input-space masking from RGB imagery and video, a recipe designed around spatial redundancy. HSI distributes its redundancy differently. Neighboring pixels are often non-redundant, while adjacent wavelengths are so highly correlated that masked bands remain linearly recoverable from visible ones until most of the spectrum is removed. Input-space methods therefore rely on masking ratios of up to 90%, which make reconstruction difficult only by leaving the encoder a sparse view of each pixel's spectrum. We propose CoMask, which exploits this spectral correlation by projecting HSI pixels onto a small number of basis components that span all wavelengths. CoMask masks in this component basis, where zeroing a component removes its contribution from every wavelength. The encoder therefore sees every wavelength at reduced fidelity, with degradation controlled by how many components are masked. Reconstruction difficulty then depends on how much spectral signal is removed rather than on which wavelengths are hidden. That dependency aligns the task with the low-dimensional structure of HSI data. Under linear probing on Houston2018 at ViT-Tiny, CoMask exceeds every input-space masking configuration in our sweep, leading the strongest by 2.7 points of overall accuracy averaged over three pretraining seeds. On EnMAP-DFC the lead appears under finetuning rather than probing, at 2.0 points of macro accuracy over the strongest configuration in that sweep. Pretrained on the foundation-scale Spectral Earth corpus, CoMask outperforms an input-space baseline of the same architecture and recipe on all six downstream segmentation tasks, by 0.9 to 2.6 micro-IoU.

1 INTRODUCTION

Hyperspectral sensors record reflected light at hundreds of narrow, contiguous wavelengths, so every pixel carries a spectrum that identifies the materials present from their absorption and reflectance features (Goetz et al., 1985; Schaepman et al., 2009). Labels rather than data are the scarce resource, since pixel-level annotation needs expert interpretation or field survey (Thoreau et al., 2024; Li et al., 2019). Self-supervised pretraining on the unlabeled portion is the standard response, and masked autoencoding (Devlin et al., 2019; He et al., 2022) is the dominant recipe for hyperspectral imagery (Ibañez et al., 2022; Scheibenreif et al., 2023; Wang et al., 2025; Braham et al., 2025).

Nearly every hyperspectral masked autoencoder builds its masked unit out of raw band measurements, and we call any such strategy input-space masking whatever the unit, whether a single band, a token of contiguous bands (Ibañez et al., 2022; Zhu et al., 2023), a spatial-spectral block or a pixel tube (Scheibenreif et al., 2023). The exceptions mask coefficients of a transformed basis instead, and Section 2 takes them up. How much a method removes follows from how redundant its input is, since masking has to leave a task that interpolation from the visible part cannot solve (Vincent et al., 2010). Image and video MAE mask 75% and 90% of their tokens against redundancy that is spatial and temporal (He et al., 2022; Tong et al., 2022), and most hyperspectral methods inherit that reasoning

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

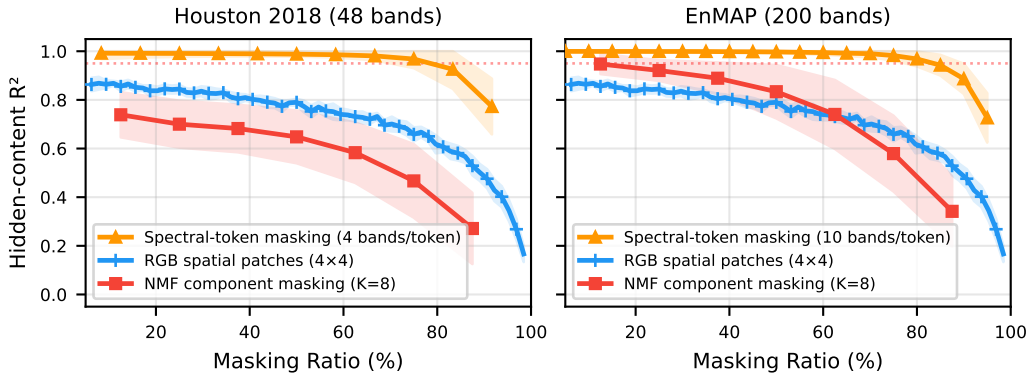


Figure 1: Hidden content stays linearly recoverable under input-space masking. A ridge regression predicts the hidden content from the visible content, and R^2 is scored on held-out samples over the hidden content only: the masked bands for spectral-token masking, the masked patches for RGB spatial masking on CIFAR-10, and the masked components' contribution for component masking at $K = 8$. Bands are one standard deviation across mask realizations.

and pretrain at 75 to 90% (Wang et al., 2025; Braham et al., 2025). Those ratios are not comparable across methods, since the token differs and some compose a spatial ratio with a spectral one. What the numbers agree on is only that these methods remove most of the input.

Hyperspectral redundancy does not sit where that reasoning places it. The spatial axis often carries little of it, since many applications work at resolutions where neighboring locations hold different materials. The spectral axis carries it instead. For most adjacent band pairs the correlation coefficient is close to one, and it falls only where atmospheric or material attenuation leaves the signal dominated by noise (Wang et al., 2007), while the intrinsic dimensionality of a scene's spectra lies far below its band count (Bioucas-Dias et al., 2012). Figure 1 measures what that redundancy leaves a masked autoencoder to do. Hidden bands stay linearly recoverable from visible ones at $R^2 \geq 0.968$ up to 75% masking on Houston2018 and 80% on EnMAP. Input-space masking therefore faces a trade between two levers. A low ratio keeps each pixel's spectrum nearly intact but leaves the reconstruction largely recoverable by interpolation, while a high ratio removes that recoverability and hands the encoder a sparse view of the spectrum it is meant to learn from. None of this is evidence that input-space masking fails, since these methods produce strong representations and underpin recent hyperspectral foundation models. It describes the design space they work in.

CoMask breaks that trade by masking outside the wavelength domain. Each pixel spectrum is approximated by a small number of components that span every wavelength, fitted once per dataset on unlabeled pixels and then held fixed, and CoMask zeros a subset of the coefficients, so the unit the mask removes is a component rather than a set of band measurements. Recombining what survives hands the encoder a spectrum with a value at every wavelength, degraded in fidelity rather than in coverage, so difficulty follows from how much spectral signal was removed rather than from which wavelengths are hidden. At 75% masking, component masking leaves the hidden content at R^2 0.47 on Houston2018 and 0.58 on EnMAP, against 0.97 and 0.98 for spectral tokens. That lower recoverability is a property of the masking and not of the representations it yields, and the experiments below test whether it translates.

Contributions.

- CoMask, a masking strategy that operates in a fitted spectral basis and reconstructs in wavelength space, running before tokenization with no change to the encoder or decoder.
- A controlled attribution of the gain, at matched input degradation against an input-space control and against random, fixed-frequency and wavelet bases. No prior method that masks in a transformed basis compares one basis against another, so this comparison carries the originality of the work.
- An evaluation on three benchmarks through a single downstream harness with stated comparison rules, reporting where CoMask does not win, together with label-efficiency and

basis-provenance results. We release the code and pretrained checkpoints for all three benchmarks as supplementary material.

2 Related Work

The methods that mask outside the input domain form a small but established area of work. Four properties separate them: which axes the basis spans, whether the basis is fixed or fitted, which coefficients the mask removes together, and whether the model reconstructs the coefficients or the original input. PMAE fits a PCA basis over whole images and MFM masks a fixed two-dimensional Fourier basis, so both transform image content rather than the spectrum (Bizeul et al., 2025; Xie et al., 2023), and WaveMAE does the same with a fixed Haar basis on the spatial axes of multispectral imagery (Bernuzzi et al., 2025). All three were designed against redundancy that lies along the spatial axes, which is where it lies in RGB imagery and not where it lies in hyperspectral data. Applied unchanged to a hyperspectral cube, a basis over image content masks the axis that carries the least redundancy and leaves the correlation between adjacent wavelengths intact.

The methods that transform the spectrum itself sit closer, and differ in the origin of the basis and in what the mask removes. SSLSM and HSIMAE fit PCA but mask whole components or whole band groups, and both reconstruct in the transformed space (Liu et al., 2023; Wang et al., 2024). KARMA instead places a linear mixing model in the decoder, leaving the masked units as raw patches (Matin et al., 2025). SFMIM is the nearest prior work, since it also transforms each pixel's spectrum and reconstructs the original spectrum. Its basis is fixed, though, and its mask removes a frequency band rather than a selected set of coefficients (Mohamed et al., 2025). CoMask is the only combination of a fitted spectral basis, a selected set of masked coefficients, and reconstruction in wavelength space. That combination is what turns the basis into something the pretext task can be tested against, because the mask acts on coefficients that a fitting procedure produced while the loss stays on the measured spectrum, so one basis can be exchanged for another without altering the encoder, the target or the terms of the comparison.

Decomposition itself is long-standing as hyperspectral preprocessing, from the maximum noise fraction transform (Green et al., 1988) to principal components taken ahead of classification (Rodarmel & Shan, 2002; Licciardi et al., 2012; Kang et al., 2017), with the limits of that practice documented alongside it (Prasad & Bruce, 2008). What is new here is therefore not the use of a basis but the question of what the basis has to provide. Every method above adopts one basis and none tests it against another, so whether the choice matters is unestablished. Section 6 runs that comparison, replacing CoMask's fitted basis with a random basis, a fixed frequency basis and a wavelet basis at matched input degradation. To our knowledge it is the first evidence in the hyperspectral setting of which properties of a basis the masking task depends on, a question every method above leaves open.

3 Method

Decomposition. CoMask begins by re-expressing each pixel spectrum $\boldsymbol{x} \in \mathbb{R}^C$ as a weighted sum of a small number of components that span every wavelength,

$$\boldsymbol{x} \approx \sum_{k=1}^{K} a_k \boldsymbol{b}_k, \tag{1}$$

where each component $\boldsymbol{b}_k \in \mathbb{R}^C$ is a full-length spectrum and a_k is its coefficient at that pixel. The representation is useful only when $K \ll C$, and hyperspectral data permits that, because the intrinsic dimensionality of a scene lies far below its band count (Bioucas-Dias et al., 2012; Cawse-Nicholson et al., 2022). Applied at every location of a cube $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, Equation 1 yields a coefficient cube $\mathbf{A} \in \mathbb{R}^{H \times W \times K}$ and a component matrix $\boldsymbol{B} \in \mathbb{R}^{K \times C}$ whose rows are the components, so that $\mathbf{X} \approx \mathbf{A}\boldsymbol{B}$. Equation 1 is the linear mixing model of hyperspectral unmixing (Keshava & Mustard, 2002; Bioucas-Dias et al., 2012). CoMask borrows that model as a basis in which to mask, and makes no claim that its components are physical materials. The decomposition is fitted once per dataset on unlabeled training pixels, with test-region pixels excluded throughout, and is held fixed for the whole of pretraining. Where a corpus is too large for a single fit, the basis is fitted on a sample of pixels with a minibatch solver, and coefficients for the remaining pixels are obtained by projecting onto that fixed basis.

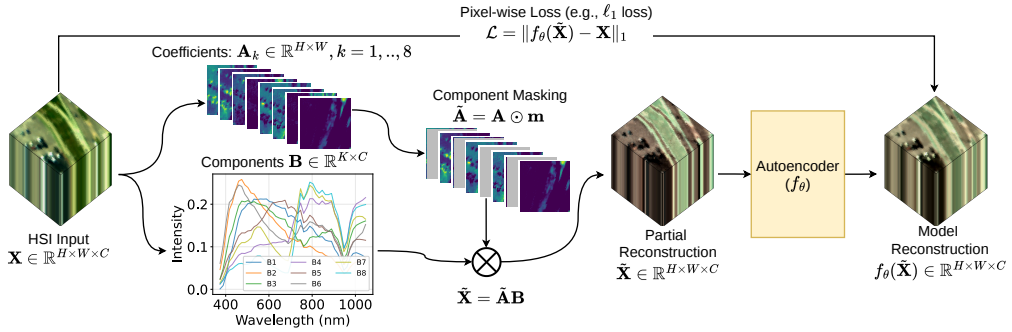


Figure 2: CoMask pretraining. A hyperspectral cube $\mathbf{X}$ is decomposed once per dataset into coefficient maps $\mathbf{A}_k$ and a component matrix $\boldsymbol{B}$; a binary mask $\boldsymbol{m}$ zeros a subset of the coefficients, and the retained components are recombined into the partial reconstruction $\tilde{\mathbf{X}}$ that the encoder and decoder $f_{\boldsymbol{\theta}}$ reconstruct the original cube from. The basis shown is the $K = 8$ NMF basis fitted on Houston2018. The figure shows $\rho_s = 1$; when $\rho_s < 1$ a fraction of locations keep their original spectra.

Choice of decomposition. Nothing in CoMask depends on how the components are obtained, so any procedure that returns components spanning all C wavelengths can serve as the masking basis. We study three that differ in what they constrain: principal component analysis, which returns orthogonal components ordered by variance and coefficients of either sign (Jolliffe, 2002); non-negative matrix factorization, which constrains components and coefficients to be non-negative so that each component contributes additively (Lee & Seung, 1999); and SPICE, an unmixing method that keeps the coefficients non-negative and adds a sum-to-one constraint, solved by alternating quadratic programming (Zare & Gader, 2007). Each supplies the coefficients through its own transform, so the constraint a decomposition imposes is the constraint its coefficients satisfy, and Appendix I gives the rule used for each dataset. NMF is our default, and Section 4 states which decompositions were fitted for each dataset. Two further families are excluded by cost and by assumption rather than by result. Kernel PCA forms and decomposes a matrix with one row and column per pixel, which is impractical beyond roughly 10^5 pixels against 1.4M in our smallest scene (Schölkopf et al., 1998). Independent component analysis assumes independent sources, which the sum-to-one coefficients of the mixing model violate by construction (Hyvärinen & Oja, 2000).

Component masking. Information is removed from the input by zeroing coefficients rather than by deleting wavelength measurements. Let $\boldsymbol{m} \in \{0, 1\}^K$ be a binary mask in which $m_k = 1$ marks component k as retained; one mask is drawn per training sample and shared by every masked location within it. Masking multiplies a pixel's coefficients elementwise by $\boldsymbol{m}$ and recombines what survives,

$$\tilde{\boldsymbol{a}} = \boldsymbol{a} \odot \boldsymbol{m}, \qquad \tilde{\boldsymbol{x}} = \sum_{k\,:\,m_k=1} a_k \boldsymbol{b}_k, \tag{2}$$

so the partial reconstruction $\tilde{\boldsymbol{x}}$ is Equation 1 with the masked components dropped, and at cube level $\tilde{\mathbf{X}} = \tilde{\mathbf{A}}\boldsymbol{B}$. Each masked component therefore removes a coherent pattern $a_k \boldsymbol{b}_k$ that spans the whole spectrum, rather than a set of wavelengths. What the encoder receives still carries a value at every wavelength position, so the input is degraded in fidelity rather than in coverage.

Masking configuration. Two ratios and a selection rule fix how much is removed and where. The component ratio $\rho_c \in (0, 1)$ sets the fraction of components masked at each masked pixel, giving $\lfloor \rho_c K \rfloor$ masked components, and the spatial ratio $\rho_s \in (0, 1]$ sets the fraction of locations that receive masking at all. Random selection draws the masked components uniformly without replacement, while dominant selection ranks components by their mean coefficient magnitude across the sample and masks the largest, so that the strongest spectral contributions are the ones removed. Our implementation also supports budget-based strategies that fix a target degradation rather than a component count (Appendix I). When $\rho_s < 1$, a set $\mathcal{M}$ of $\lfloor \rho_s HW \rfloor$ locations is drawn uniformly per sample, and the encoder input at location (i, j) is $\tilde{\boldsymbol{x}}_{ij} = \tilde{\boldsymbol{a}}_{ij}\boldsymbol{B}$ for $(i, j) \in \mathcal{M}$ and the original spectrum $\boldsymbol{x}_{ij}$ otherwise. Retaining some unmasked input is standard in masked autoencoding rather than particular to CoMask, with image MAE reconstructing from a visible quarter of the patches (He et al., 2022) and pixel-tube masking on hyperspectral cubes leaving whole spatial locations untouched (Scheibenreif et al., 2023).

Reconstruction target. The encoder and decoder $f_{\boldsymbol{\theta}}$ are trained to recover the original cube from the partial reconstruction, under

$$\mathcal{L} = \left\| f_{\boldsymbol{\theta}}(\tilde{\mathbf{X}}) - \mathbf{X} \right\|_1, \tag{3}$$

averaged over the pixels in a sample and over the samples in a batch. The target is the original spectrum $\boldsymbol{x}$ rather than its approximation $\boldsymbol{aB}$, and the two differ because the decomposition leaves a residual. A model that learned only to undo the masking would produce $\boldsymbol{aB}$, leaving exactly the structure the decomposition fails to capture unrecovered. Targeting $\boldsymbol{x}$ forces the model to account for that residual, and confines the decomposition to its role as the basis in which masking happens.

Preprocessing. Pixels pass through the same three steps before the decomposition is fitted and before every training run. Bands corrupted by water absorption or sensor noise are removed, with the affected band indices listed per dataset in Appendix J. Negative reflectance values, which arise from atmospheric correction, are clipped to zero, because the non-negative decompositions require it. Each spectrum is then rescaled to unit L_2 norm, which separates spectral shape from absolute magnitude, a quantity that varies with illumination and topography rather than with material. Every method compared in this paper is trained and evaluated on these same normalized spectra, baselines included, so the normalization cannot account for any difference between them. All later references to $\boldsymbol{x}$ are to the normalized spectrum.

4 Experimental Setup

Datasets. Three benchmarks cover the conditions the method is meant to work under, and Appendix J summarizes their band counts, tile geometry and fitted decompositions. Houston2018 (Xu et al., 2019) is a single 1 m aerial scene with the official IEEE GRSS DFC 2018 spatial split, small enough for controlled comparison and for the configuration grid. EnMAP-DFC moves to 30 m satellite imagery and severe class imbalance, with 66% of its labeled pixels in a single class. Pretraining there uses the unlabeled EnMAP collection assembled by Scheibenreif et al. (2023) from the EnMAP mission (Guanter et al., 2015; Storch et al., 2023), and evaluation uses the paired labeled subset, whose land-cover labels come from the DFC 2020 challenge (Schmitt et al., 2019). Spectral Earth (Braham et al., 2025) is a globally distributed corpus from the same sensor, and its 8.8 billion pretraining pixels place CoMask at foundation-model pretraining scale. Downstream evaluation there uses the six semantic segmentation tasks released with the corpus, which span crop type, land cover and forest cover across three continents.

Pipeline and architectures. CoMask is evaluated inside a published HSI masked-autoencoder pipeline rather than one of our own. We take MaskedSST (Scheibenreif et al., 2023) and change only the masking. The blockwise spectral tokenizer, the encoder and decoder, the optimizer and the training schedule are all as published, and CoMask replaces the input-space mask with the component mask of Section 3. Because that mask is applied before tokenization, the encoder receives a cube of the usual shape and tokenizes it unchanged. We run three encoder scales, ViT-Tiny at 1.7M, ViT-Small at 21.5M and ViT-Base at 85.4M parameters, to test how the masking interacts with capacity.

Baselines. Every table compares CoMask against input-space masking under the same encoder and recipe. The input-space baselines are MaskedSST-Block, which masks contiguous spatial-spectral blocks, and MaskedSST-Tube, which masks pixel tubes across all wavelengths and so removes spectral interpolation as a reconstruction route. The two bracket how input-space masking removes information from a cube, by contiguous region and by whole-spectrum location. Recent HSI foundation models are not used as baselines, because they vary architecture, capacity and pretraining data at once, so a comparison would confound the masking strategy with all three. Spectral Earth supplies the comparison at foundation-model pretraining scale instead. Supervised training from scratch and a frozen, never-pretrained encoder are also reported, the latter as a reference floor rather than as a method.

Comparison rules. Every CoMask and baseline number in this paper comes from the same downstream harness, with the same head, the same five evaluation seeds and the same metric definitions, so rows differ only in what was pretrained and how. Each input-space baseline is reported at its strongest cell over the masking-ratio sweep, eleven ratios on Houston2018 and three per mode on EnMAP. CoMask uses one configuration for every Houston2018 and EnMAP row, dominant selection with $\rho_c = 0.75$, $\rho_s = 0.70$ and $K = 8$, held fixed across encoder scales and decompositions. Appendix D

gives the grid that configuration was chosen on. On Spectral Earth the comparison is against an in-house MaskedSST-Tube run at the same ViT-Tiny encoder, pretraining data and training budget with a matched segmentation head, with CoMask using NMF at $K = 10$, random selection, $\rho_c = 0.6$ and $\rho_s = 0.8$. The published Spec-ViT-S rows there are printed as reported and do not pass through our harness.

Metrics. Houston2018 reports linear probing and EnMAP-DFC reports finetuning, with macro accuracy primary on EnMAP because of its class imbalance. That split is forced by the data rather than chosen. On EnMAP no pretrained encoder exceeds the frozen random-init floor by more than 1.5 points of overall accuracy at any encoder scale, and on macro accuracy the floor sits above the input-space baseline at ViT-Small and ViT-Base, so probing there does not separate methods. A random-init floor row is reported throughout.

Cost of the decomposition. Fitting the decomposition is a one-off cost per dataset and is small against pretraining. At Spectral Earth scale it takes about 1% of pretraining GPU-hours, since the fit runs mostly on CPUs while pretraining uses four B200s. On Houston2018 it is under 3% of pretraining wall time at every encoder scale. On EnMAP it is 7 to 20% of wall time, highest at ViT-Tiny, where the share is large because pretraining there is shortest rather than because fitting is more expensive. Per-iteration cost during pretraining is unchanged, since the coefficients are precomputed and loaded as fixed buffers.

5 RESULTS

On Houston2018 CoMask leads input-space masking under linear probing. CoMask-NMF reaches 58.19 ± 0.62 overall accuracy at ViT-Tiny and beats all 22 input-space cells, block and tube masking at each of eleven ratios. The closest input-space cell is tube masking at ratio 0.3, which CoMask-NMF leads by 1.08 even though that cell was chosen on test accuracy. The lead grows over three pretraining seeds, to 2.73 across the four cells pretrained at all three, with CoMask-NMF at 57.91 ± 1.13 against 55.18 ± 2.61.

The strongest CoMask row beats the input-space baseline at every encoder scale, on both probing columns. Which decomposition supplies that row changes with capacity. NMF is best at ViT-Tiny, where CoMask-PCA reaches only 52.19 and falls below both input-space baselines. SPICE is best at ViT-Small and ViT-Base, at 59.88 ± 0.75 and 57.10 ± 0.26 against 51.30 and 55.32 for the input-space baseline. At ViT-Small that baseline sits at the random-init probing floor of 51.53. The choice of decomposition matters most at ViT-Base. There CoMask-SPICE beats the baseline on three of the four columns and ties the fourth, while CoMask-NMF trails the baseline on all four. Every CoMask row still clears the random-init probing floor, by 11.7, 8.4 and 4.7 points at the best row of each scale and by 1.4 for CoMask-NMF at ViT-Base.

Full-label finetuning narrows the differences that probing shows. At ViT-Tiny the three best finetuning rows on overall accuracy fall within one standard deviation of each other, at 61.66, 61.50 and 61.19. The input-space baselines lead on macro accuracy at that scale, where CoMask-NMF trails the stronger of the two by 1.75. The narrowing is uneven across scales, since CoMask-SPICE leads both finetuning columns at ViT-Small and leads on macro accuracy at ViT-Base.

On EnMAP-DFC CoMask leads input-space masking under finetuning at every encoder scale, by 4.25, 2.95 and 1.70 points of macro accuracy over the cells in Table 2. At ViT-Tiny those cells are each method's strongest of the six-cell sweep, scored on the mean of the two macro columns. Scored on finetuning macro accuracy alone, the strongest cell is tube masking at ratio 0.5, which reaches 55.48 ± 1.51. CoMask-NMF leads that cell by 1.95. The lead over supervised training from scratch is thinner than the lead over input-space masking. CoMask-NMF ties from-scratch at ViT-Tiny, 85.08 against 85.08, and at ViT-Small, 84.70 against 84.72. It wins only at ViT-Base, by 1.49 points of overall accuracy and 5.23 points of macro accuracy.

The advantage over from-scratch training is largest where labels are scarce. At 1% of Houston2018 labels CoMask-NMF finetunes to 51.10 against 30.38 for from-scratch training, a margin of 20.7 points of overall accuracy. That margin narrows as labels are added, reaching 5.0 points at full supervision. CoMask-NMF also leads both input-space baselines at every Houston fraction on overall accuracy, although the margins at 50% and 100%, 0.05 and 0.37, fall inside the standard deviations. EnMAP-DFC repeats the pattern on macro accuracy, where CoMask-NMF leads from-scratch by 4.9

Table 1: Houston2018 across encoder scales. CoMask uses the fixed configuration (dominant selection, $\rho_c = 0.75$, $\rho_s = 0.70$, $K = 8$). Input-space rows use block masking at ratio 0.6 and tube masking at 0.3 at ViT-Tiny, and tube masking at 0.7 at the two larger scales. The random-init row is a frozen, never-pretrained encoder, reported as a reference. Bold marks the best value per column and architecture.

Architecture	Method	LP-OA	LP-Macro	FT-OA	FT-Macro
ViT-Tiny	From-scratch	n/a	n/a	56.15 ± 1.92	53.21 ± 2.20
	Random-init floor	46.45 ± 2.33	44.59 ± 2.25	n/a	n/a
	MaskedSST-Block	55.26 ± 1.09	52.81 ± 1.07	61.50 ± 0.93	58.96 ± 0.47
	MaskedSST-Tube	57.11 ± 0.77	54.41 ± 0.91	**61.66 ± 1.14**	**59.33 ± 1.20**
	CoMask-PCA	52.19 ± 0.32	50.00 ± 0.28	60.38 ± 0.74	57.13 ± 1.02
	CoMask-NMF	**58.19 ± 0.62**	**55.35 ± 0.53**	61.19 ± 0.45	57.58 ± 1.60
	CoMask-SPICE	53.58 ± 0.58	51.32 ± 0.45	58.96 ± 1.42	56.67 ± 0.99
ViT-Small	From-scratch	n/a	n/a	57.45 ± 1.37	55.32 ± 1.43
	Random-init floor	51.53 ± 1.64	48.75 ± 1.47	n/a	n/a
	MaskedSST-Tube	51.30 ± 0.67	49.88 ± 0.57	58.28 ± 1.59	55.91 ± 1.39
	CoMask-NMF	55.44 ± 0.68	52.78 ± 0.52	58.52 ± 1.33	55.92 ± 1.16
	CoMask-SPICE	**59.88 ± 0.75**	**56.01 ± 0.75**	**60.62 ± 1.23**	**57.88 ± 0.88**
ViT-Base	From-scratch	n/a	n/a	56.75 ± 1.05	54.92 ± 1.01
	Random-init floor	52.42 ± 2.37	49.95 ± 1.98	n/a	n/a
	MaskedSST-Tube	55.32 ± 0.45	52.63 ± 0.58	**59.70 ± 0.24**	57.01 ± 0.38
	CoMask-NMF	53.84 ± 0.79	51.76 ± 0.88	56.56 ± 2.16	54.05 ± 1.99
	CoMask-SPICE	**57.10 ± 0.26**	**54.91 ± 0.29**	59.61 ± 0.62	**57.60 ± 0.41**

Table 2: EnMAP-DFC across encoder scales. CoMask uses the fixed configuration (dominant selection, $\rho_c = 0.75$, $\rho_s = 0.70$, $K = 8$). Input-space rows use block masking at ratio 0.5 and tube masking at 0.7. The random-init row is a frozen, never-pretrained encoder, reported as a reference. Bold marks the best finetuning value per column and architecture. The probing columns carry no bold, since no method there separates from the random-init floor.

Architecture	Method	LP-OA	LP-Macro	FT-OA	FT-Macro
ViT-Tiny	From-scratch	n/a	n/a	**85.08 ± 0.62**	56.88 ± 1.96
	Random-init floor	81.45 ± 0.72	45.30 ± 2.46	n/a	n/a
	MaskedSST-Block	82.64 ± 0.63	50.50 ± 1.66	84.00 ± 0.52	53.18 ± 2.19
	MaskedSST-Tube	82.34 ± 0.46	48.64 ± 1.08	83.34 ± 0.89	51.80 ± 2.18
	CoMask-PCA	82.51 ± 0.42	50.61 ± 1.49	84.99 ± 0.93	**57.91 ± 1.45**
	CoMask-NMF	82.28 ± 0.57	50.30 ± 0.90	**85.08 ± 0.73**	57.43 ± 1.37
	CoMask-SPICE	82.89 ± 0.46	49.60 ± 1.66	83.75 ± 1.48	55.09 ± 4.25
ViT-Small	From-scratch	n/a	n/a	**84.72 ± 0.50**	54.88 ± 0.67
	Random-init floor	82.75 ± 0.43	51.81 ± 0.94	n/a	n/a
	MaskedSST-Tube	82.52 ± 0.91	49.41 ± 1.77	83.63 ± 0.95	53.10 ± 1.96
	CoMask-NMF	82.40 ± 0.76	52.11 ± 2.05	84.70 ± 0.75	**56.05 ± 2.01**
ViT-Base	From-scratch	n/a	n/a	82.71 ± 0.54	48.86 ± 1.17
	Random-init floor	82.38 ± 0.78	51.41 ± 3.47	n/a	n/a
	MaskedSST-Tube	82.34 ± 0.59	45.94 ± 1.05	**84.30 ± 0.32**	52.39 ± 2.05
	CoMask-NMF	82.70 ± 1.03	52.67 ± 1.49	84.20 ± 0.58	**54.09 ± 1.33**

points at 10% of labels. It leads tube masking at every EnMAP fraction too, by 1.7 to 5.6 points of macro accuracy, though the margin at 10% falls inside the standard deviations. Appendix A reports both sweeps in full. Its input-space rows use one pretrained checkpoint per architecture, selected on validation accuracy at full labels, rather than the strongest cell of the masking-ratio sweep that Tables 1 and 2 report.

The advantage holds at foundation-model pretraining scale as well. Pretrained on Spectral Earth and evaluated with a frozen encoder, CoMask-NMF exceeds the in-house MaskedSST-Tube run on all six segmentation tasks, by 0.9 to 2.6 micro-IoU. Those two runs differ only in the masking. CoMask-NMF also exceeds every published Spec-ViT-S result in Table 3 on all six tasks, by 0.6 to 2.9 micro-IoU, at a ViT-Tiny encoder against their ViT-Small.

Table 3: Spectral Earth segmentation under a frozen encoder, micro-IoU (%) on the six released tasks. CoMask pretrains with NMF at $K = 10$, random selection, $\rho_c = 0.6$ and $\rho_s = 0.8$. The Spec-ViT-S rows are taken directly from Braham et al. (2025), Table II, and use a ViT-Small encoder. The two rows below the rule share the ViT-Tiny encoder, the pretraining corpus, the training budget and the segmentation head. Bold marks the best value per task.

Method	CDL	NLCD	EuroCrops	TreeMap	BDForet	BNETD
Spec-ViT-S, random init	65.96 ± 0.72	36.94 ± 0.30	55.78 ± 0.52	38.72 ± 0.25	61.89 ± 0.85	41.55 ± 0.22
Spec-ViT-S, MoCo-v2	71.40 ± 0.32	39.99 ± 0.08	60.21 ± 0.39	41.76 ± 0.06	68.50 ± 0.12	43.50 ± 0.05
Spec-ViT-S, DINO	71.56 ± 0.35	40.47 ± 0.06	58.76 ± 0.27	41.89 ± 0.13	68.16 ± 0.43	43.12 ± 0.10
Spec-ViT-S, MAE	72.01 ± 0.28	41.08 ± 0.05	62.14 ± 0.18	41.19 ± 0.15	68.92 ± 0.73	43.34 ± 0.26
MaskedSST-Tube, ViT-Tiny	70.83 ± 0.21	41.79 ± 0.04	62.32 ± 0.20	42.53 ± 0.07	70.29 ± 0.23	44.07 ± 0.11
CoMask-NMF, ViT-Tiny	$\mathbf{72.57 \pm 0.13}$	$\mathbf{42.86 \pm 0.10}$	$\mathbf{64.90 \pm 0.29}$	$\mathbf{43.50 \pm 0.08}$	$\mathbf{71.77 \pm 0.27}$	$\mathbf{45.01 \pm 0.10}$

6 Attribution of the Gain

Section 5 holds the encoder, the tokenizer and the training recipe fixed, so the gain it reports belongs to the masking rather than to the architecture. Which property of the masking produces that gain is a separate question, and this section answers it by replacing one property at a time and measuring what each replacement costs, a comparison no prior method that masks in a transformed basis runs.

At matched input degradation, CoMask exceeds an input-space mask by 9.09 points of linear-probe overall accuracy. Masking ratios do not support that comparison, because zeroing a fraction of a pixel's components removes a different amount of signal than zeroing the same fraction of its bands. Degradation is therefore measured on the encoder input itself, as normalised input RMSE, $\mathrm{NRMSE} = \|\tilde{\mathbf{X}} - \mathbf{X}\|_F / \|\mathbf{X}\|_F$, where both norms run over every pixel and band of a sample and $\tilde{\mathbf{X}}$ is the cube the encoder receives. Measuring degradation before encoding rather than on encoder activations is what keeps the two strategies comparable, since an input-space mask and CoMask hand the encoder different quantities. Every arm in Table 4 is matched on this quantity to within 0.006 of CoMask's 0.796, with CoMask's spatial masking fraction held fixed and only the number of masked coefficients calibrated per arm. CoMask itself runs at the fixed configuration of Tables 1 and 2, whose selection grid Appendix D reports. The input-space control, which zeros every band inside a subset of patches and reaches degradation 0.791 through its spatial ratio alone, probes to 48.82 against CoMask-NMF at 57.91. The control masks fewer locations than CoMask, at $\rho_s = 0.625$ against 0.70. It therefore reconstructs with more intact spectra available than CoMask has, not fewer.

CoMask's advantage comes from the structure of what the encoder still sees, not from spectral completeness alone. Two of the arms keep every wavelength and remove that structure. The first adds noise drawn from the fitted band covariance of the training pixels, scaled to CoMask's degradation, so the encoder receives a complete spectrum whose shape carries no fitted component. The second keeps the fitted components and permutes the band entries within each one, which preserves every component's distribution of values and destroys its spectral shape. CoMask leads the covariance corruption by 3.15 points and the permuted basis by 3.33, margins of 1.7 and 1.6 sd units across three pretraining seeds. Both are well inside the 9.09 over the input-space control, so part of the gain is available from spectral completeness alone and the rest requires structure the decomposition supplies.

CoMask's basis has to carry real spectral structure. A random non-negative basis of the same size, at the same degradation and with coefficients of its own, costs 4.82 points of probing accuracy, a result that holds for Houston under probing rather than generally. Fixed transform bases cost more than the random one. Masking a complete orthonormal discrete cosine basis (Ahmed et al., 1974) costs 5.46, and a complete Daubechies wavelet basis (Daubechies, 1992) costs 8.20 over three pretraining seeds. Those are the fixed cosine and wavelet families that prior transformed-basis methods mask (Mohamed et al., 2025; Bernuzzi et al., 2025), so the comparison covers bases the field already uses rather than synthetic ones alone. They are also the costliest basis substitutions in Table 4, falling below the random basis and below the corruptions.

The attribution repeats on EnMAP-DFC under finetuning, where macro accuracy is the primary metric. CoMask leads the matched input-space control there by 3.37 points of macro accuracy, and leads the cosine and wavelet bases by 2.51 and 2.73. Those runs use a second configuration, dominant selection at $\rho_c = 0.875$ and $\rho_s = 0.60$, and every EnMAP arm's degradation is matched to it to

Table 4: Attribution arms on Houston2018 at ViT-Tiny. Each arm replaces one property of the masking at matched normalised input RMSE. Values are mean ± sd over three pretraining seeds. Bold marks the best value per column.

Arm	Input NRMSE	LP-OA	LP-Macro	FT-OA	FT-Macro
CoMask-NMF	0.796	**57.91 ± 1.13**	**54.47 ± 2.05**	**60.48 ± 0.70**	**57.48 ± 0.39**
Input-space mask	0.791	48.82 ± 2.41	46.08 ± 2.09	59.90 ± 1.79	57.12 ± 1.34
White noise	0.796	53.21 ± 0.83	49.67 ± 0.75	58.70 ± 0.97	55.57 ± 0.69
Covariance noise	0.796	54.76 ± 2.40	51.40 ± 1.59	59.75 ± 0.92	57.43 ± 1.08
Random non-negative basis	0.791	53.08 ± 1.05	50.62 ± 1.08	59.18 ± 1.22	56.89 ± 1.55
Band-permuted NMF basis	0.797	54.58 ± 2.79	51.45 ± 1.98	58.27 ± 0.52	56.40 ± 0.48
Fixed frequency basis	0.799	52.45 ± 2.15	49.90 ± 2.04	59.82 ± 0.79	57.41 ± 0.62
Wavelet basis	0.800	49.71 ± 3.05	47.27 ± 3.09	58.15 ± 1.97	56.05 ± 1.71

within 0.027. Appendix B reports the full set. The comparison that does not repeat is the fitted-basis one, where a random non-negative basis and white noise tie CoMask at 0.26 and 0.29 points of macro accuracy. What holds across both datasets is the comparison against input-space masking and against fixed transform bases.

CoMask does not need a basis fitted to the corpus it pretrains on. On EnMAP a basis fitted on one region, holding 7% of the corpus pixels, and two bases drawn from a curated spectral library (Kokaly et al., 2017) all sit within one sd of the corpus fit under probing. The corpus basis itself transfers across the regions it was fitted over, staying at or within 0.004 R^2 of the best region-specific basis on every region. These arms use one pretraining seed, probe rather than finetune, and run on EnMAP at a configuration that is not the fixed one, and Appendix C reports them. The requirement is therefore that the basis carry spectral structure rather than that it be fitted to the target corpus. Which decomposition supplies that structure best still varies with the dataset and the encoder scale.

7 Limitations and Conclusion

CoMask's advantage is a property of the pretrained representation, so it is largest where downstream training supplies the least. At 1% of Houston2018 labels CoMask leads supervised training from scratch by 20.7 points of overall accuracy, and that margin falls to 5.0 at full supervision. Full-label finetuning compresses the differences between masking strategies in the same way, to a three-way tie on overall accuracy at ViT-Tiny and a loss to the input-space baselines on macro accuracy. Labels are not the only thing that narrows the gap. The downstream objective does too. Under class-weighted cross-entropy on EnMAP-DFC, CoMask's macro-accuracy lead at ViT-Tiny does not hold, though its leads at the two larger scales do (Appendix E).

CoMask is architecture-agnostic by construction rather than by evidence. The mask acts before tokenization and leaves the encoder and decoder untouched, so any model that consumes hyperspectral patches can adopt it, but every result reported here uses a vision transformer. On Houston2018 with a Conv3D backbone, CoMask at the fixed configuration finetunes below grid-matched supervised training from scratch at all three scales. The input-space control does the same, so the shortfall belongs to masked pretraining on that backbone in our setup rather than to the decomposition (Appendix F). Scale is bounded in the same way. Pretraining at foundation scale ran at one encoder size, because fitting the basis and pretraining there carry the cost reported in Section 4. The basis is fitted to one sensor's wavelengths, so transfer across sensors is untested.

CoMask is a preprocessing step, so adopting it means changing what the mask removes and nothing else in a pipeline. The attribution arms say what that mask needs. The basis has to carry real spectral structure, while where that structure comes from matters little, so a curated spectral library serves as a starting basis where no corpus fit is available. Some locations should keep their original spectra, since nearly every configuration that masks all of them falls below the random-init probing floor (Appendix D). Among the configurations that keep some, accuracy is insensitive to the component ratio, so there is little to tune. Which decomposition supplies the structure best still varies with the dataset and the encoder scale, and that is the one choice a new deployment has to make.

AI USE STATEMENT

We used an LLM assistant to suggest alternative phrasings for the title, to tighten prose for clarity and concision in the abstract and main text, and to format the LaTeX source, including its tables.

An LLM assistant also produced the initial drafts of two appendices, Per-Dataset Details (Appendix J) and Implementation Details (Appendix I), which the authors then reviewed, revised and verified against the experimental record.

Beyond the writing, we used an LLM assistant to prepare the supplementary artifact and to check that it reproduces the paper. It curated the released code, removing author-identifying information and adding unit tests, and it ran the released checkpoints to confirm that every number reported here reproduces from them. The experiments themselves, and the code that produced the reported results, are the authors' own.

We did not use generative AI tools for research ideation, for retrieval and discovery of related work, for generating synthetic datasets, or for proving mathematical claims. The method, the experimental design, and the analysis and interpretation of results are the authors' own work. All related work was identified and read by the authors. We have reviewed all AI-assisted text and revised it to reflect our own claims and framing.

We take responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code and its unit tests, the configuration files for every reported run, the fitted decomposition bases, and the pretrained ViT-Tiny checkpoints for all three benchmarks. Checkpoints at the larger encoder scales are omitted so that the archive fits the supplementary size limit. Per-pixel coefficients are not shipped either, since they are a dense transform of the imagery that the released scripts regenerate. The included README says what each directory holds and gives the commands that reproduce the headline numbers. The ViT-Tiny results reported here have been checked to reproduce from the included checkpoints.

Every in-house number in this paper comes from one downstream evaluation harness, which uses the same classification head, the same five downstream seeds and the same metric files for every row, and which runs probing with the encoder in evaluation mode throughout. Section 4 states the comparison and metric rules that harness applies. The published rows of Table 3 are the one exception, reported as they appear in their source rather than rerun here.

The detail needed to rebuild a run sits in the appendices. Appendix I gives the pretraining step as pseudocode, the per-dataset rules by which coefficients are computed, and the two further selection strategies the released code implements but no reported experiment uses. Appendix J gives the band removal, tiling and splits for each benchmark. Appendix D gives the grid the fixed configuration was selected on and how it was scored, so that the selection can be judged rather than taken on trust.

REFERENCES

N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. doi: 10.1109/T-C.1974.223784.

Vittorio Bernuzzi, Leonardo Rossi, Tomaso Fontanini, Massimo Bertozzi, and Andrea Prati. WaveMAE: Wavelet decomposition masked auto-encoder for remote sensing. *arXiv preprint arXiv:2510.22697*, 2025.

José M. Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, 2012. doi: 10.1109/JSTARS.2012.2194696.

Alice Bizeul, Thomas Sutter, Alain Ryser, Bernhard Schölkopf, Julius von Kügelgen, and Julia E Vogt. From pixels to components: Eigenvector masking for visual representation learning. *arXiv preprint arXiv:2502.06314*, 2025.

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*, 18:16780–16797, 2025. doi: 10.1109/JSTARS.2025.3581451.

K. Cawse-Nicholson, A. M. Raiho, D. R. Thompson, G. C. Hulley, C. E. Miller, K. R. Miner, B. Poulter, D. Schimel, F. D. Schneider, P. A. Townsend, and S. K. Zareh. Intrinsic dimensionality as a metric for the impact of mission design parameters. *Journal of Geophysical Research: Biogeosciences*, 127(8):e2022JG006876, 2022. doi: 10.1029/2022JG006876.

Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Alexander F. H. Goetz, Gregg Vane, Jerry E. Solomon, and Barrett N. Rock. Imaging spectrometry for Earth remote sensing. *Science*, 228(4704):1147–1153, 1985. doi: 10.1126/science.228.4704.1147.

A. A. Green, M. Berman, P. Switzer, and M. D. Craig. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74, 1988. doi: 10.1109/36.3001.

Luis Guanter, Hermann Kaufmann, Karl Segl, Saskia Foerster, Christian Rogass, Sabine Chabrillat, Theres Kuester, André Hollstein, Godela Rossner, Christian Chlebek, Christoph Straif, Sebastian Fischer, Stefanie Schrader, Tobias Storch, Uta Heiden, Andreas Mueller, Martin Bachmann, Helmut Mühle, Rupert Müller, Martin Habermeyer, Andreas Ohndorf, Joachim Hill, Henning Buddenbaum, Patrick Hostert, Sebastian Van der Linden, Pedro J. Leitão, Andreas Rabe, Roland Doerffer, Hajo Krasemann, Hongyan Xi, Wolfram Mauser, Tobias Hank, Matthias Locherer, Michael Rast, Karl Staenz, and Bernhard Sang. The EnMAP spaceborne imaging spectroscopy mission for Earth observation. *Remote Sensing*, 7(7):8830–8857, 2015. ISSN 2072-4292. doi: 10.3390/rs70708830.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000. doi: 10.1016/S0893-6080(00)00026-5.

Damian Ibañez, Ruben Fernandez-Beltran, Filiberto Pla, and Naoto Yokoya. Masked auto-encoding spectral–spatial transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. doi: 10.1109/TGRS.2022.3217892.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, NY, second edition, 2002. ISBN 978-0-387-95442-4. doi: 10.1007/b98835.

Xudong Kang, Xuanlin Xiang, Shutao Li, and Jón Atli Benediktsson. PCA-based edge-preserving features for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7140–7151, 2017. doi: 10.1109/TGRS.2017.2743102.

N. Keshava and J. F. Mustard. Spectral unmixing. *IEEE Signal Processing Magazine*, 19(1):44–57, 2002. doi: 10.1109/79.974727.

Raymond F. Kokaly, Roger Clark, Gregg Swayze, K. Eric Livo, Todd Hoefen, Neil Pearson, Richard Wise, William Benzel, Heather A. Lowers, Rhonda Driscoll, and Anna Klein. USGS spectral library version 7. Data Series 1035, U.S. Geological Survey, Reston, VA, 2017.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. doi: 10.1038/44565.

Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jón Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019. doi: 10.1109/TGRS.2019.2907932.

Giorgio Licciardi, Prashanth Reddy Marpu, Jocelyn Chanussot, and Jon Atli Benediktsson. Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles. *IEEE Geoscience and Remote Sensing Letters*, 9(3):447–451, 2012. doi: 10.1109/LGRS.2011.2172185.

Weiwei Liu, Kai Liu, Weiwei Sun, Gang Yang, Kai Ren, Xiangchao Meng, and Jiangtao Peng. Self-supervised feature learning based on spectral masking for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. doi: 10.1109/TGRS.2023.3310489.

Abdul Matin, Rupasree Dey, Tanjim Bin Faruk, Shrideep Pallickara, and Sangmi Lee Pallickara. Knowledge-guided masked autoencoder with linear spectral mixing and spectral-angle-aware reconstruction. *arXiv preprint arXiv:2512.12445*, 2025. Accepted to the KGML Bridge at AAAI 2026 (non-archival).

Shaheer Mohamed, Tharindu Fernando, Sridha Sridharan, Peyman Moghadam, and Clinton Fookes. Dual-domain masked image modeling: A self-supervised pretraining strategy using spatial and frequency domain masking for hyperspectral data. In *IGARSS 2025 — 2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2440–2444, 2025. doi: 10.1109/IGARSS55030.2025.11242390.

Saurabh Prasad and Lori Mann Bruce. Limitations of principal components analysis for hyperspectral target recognition. *IEEE Geoscience and Remote Sensing Letters*, 5(4):625–629, 2008. doi: 10.1109/LGRS.2008.2001282.

Craig Rodarmel and Jie Shan. Principal component analysis for hyperspectral image classification. *Surveying and Land Information Science*, 62(2):115–123, 2002.

Michael E. Schaepman, Susan L. Ustin, Antonio J. Plaza, Thomas H. Painter, Jochem Verrelst, and Shunlin Liang. Earth system science related imaging spectroscopy—an assessment. *Remote Sensing of Environment*, 113:S123–S137, 2009. ISSN 0034-4257. doi: 10.1016/j.rse.2009.03.001. Imaging Spectroscopy Special Issue.

Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2166–2176, June 2023.

Michael Schmitt, Lloyd Hughes, Pedram Ghamisi, Naoto Yokoya, and Ronny Hänsch. 2020 IEEE GRSS data fusion contest. IEEE Dataport, 2019.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. doi: 10.1162/089976698300017467.

Tobias Storch, Hans-Peter Honold, Sabine Chabrillat, Martin Habermeyer, Paul Tucker, Maximilian Brell, Andreas Ohndorf, Katrin Wirth, Matthias Betz, Michael Kuchler, Helmut Mühle, Emiliano Carmona, Simon Baur, Martin Mücke, Sebastian Löw, Daniel Schulze, Steffen Zimmermann, Christoph Lenzen, Sebastian Wiesner, Saika Aida, Ralph Kahle, Peter Willburger, Sebastian Hartung, Daniele Dietrich, Nicolae Plesia, Mirco Tegler, Katharina Schork, Kevin Alonso, David Marshall, Birgit Gerasch, Peter Schwind, Miguel Pato, Mathias Schneider, Raquel de los Reyes, Maximilian Langheinrich, Julian Wenzel, Martin Bachmann, Stefanie Holzwarth, Nicole Pinnel, Luis Guanter, Karl Segl, Daniel Scheffler, Saskia Foerster, Niklas Bohn, Astrid Bracher, Mariana A. Soppa, Ferran Gascon, Rob Green, Raymond Kokaly, Jose Moreno, Cindy Ong, Manuela Sornig, Ricarda Wernitz, Klaus Bagschik, Detlef Reintsema, Laura La Porta, Anke Schickling, and Sebastian Fischer. The EnMAP imaging spectroscopy mission towards operations. *Remote Sensing of Environment*, 294:113632, 2023. ISSN 0034-4257. doi: 10.1016/j.rse.2023.113632.

Romain Thoreau, Laurent Risser, Véronique Achard, Béatrice Berthelot, and Xavier Briottet. Toulouse hyperspectral data set: A benchmark data set to assess semi-supervised spectral representation learning and pixel-wise classification techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212:323–337, 2024. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2024.05.003.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010. URL http://jmlr.org/papers/v11/vincent10a.html.

Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. HyperSIGMA: Hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6427–6444, 2025. doi: 10.1109/TPAMI.2025.3557581.

Hongqiang Wang, S. Derin Babacan, and Khalid Sayood. Lossless hyperspectral-image compression using context-based conditional average. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):4187–4193, 2007. doi: 10.1109/TGRS.2007.906085.

Yue Wang, Ming Wen, Hailiang Zhang, Jinyu Sun, Qiong Yang, Zhimin Zhang, and Hongmei Lu. HSIMAE: A unified masked autoencoder with large-scale pretraining for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:14064–14079, 2024. doi: 10.1109/JSTARS.2024.3432743.

Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *International Conference on Learning Representations (ICLR)*, 2023.

Yonghao Xu, Bo Du, Liangpei Zhang, Daniele Cerra, Miguel Pato, Emiliano Carmona, Saurabh Prasad, Naoto Yokoya, Ronny Hänsch, and Bertrand Le Saux. Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(6):1709–1724, 2019. doi: 10.1109/JSTARS.2019.2911113.

Alina Zare and Paul Gader. Sparsity promoting iterated constrained endmember detection in hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 4(3):446–450, 2007. doi: 10.1109/LGRS.2007.895727.

Lingxuan Zhu, Jiaji Wu, Wang Biao, Yi Liao, and Dandan Gu. SpectralMAE: Spectral masked autoencoder for hyperspectral remote sensing image reconstruction. *Sensors*, 23(7):3728, 2023. ISSN 1424-8220. doi: 10.3390/s23073728.

A Label Efficiency

Appendix A reports the two label-efficiency sweeps in full. Section 5 quotes margins from the sweeps at selected label fractions, and Table 5 gives every fraction, with the from-scratch and input-space rows beside CoMask at each one.

B EnMAP Attribution Arms

Appendix B reports the EnMAP-DFC attribution arms in full. Section 6 prints the margins it draws on, and Table 6 adds the remaining arms, the probing columns and the random-init floor.

Table 5: Label efficiency under finetuning. Houston2018 is reported on overall accuracy and EnMAP-DFC on macro accuracy, following the per-dataset metric rule. Input-space rows are one fixed pretrained checkpoint per architecture, applied at every fraction and selected on validation macro accuracy at full labels, which differs from the test-scored cell selection in Tables 1 and 2. No input-space run exists at 1% of Houston labels. Bold marks the best value per fraction.

(a) Houston2018, FT-OA

Labels	From-scratch	Block	Tube	CoMask-NMF
1%	30.38 ± 2.04	n/a	n/a	**51.10 ± 3.04**
5%	39.20 ± 1.68	57.81 ± 1.21	56.89 ± 1.30	**60.12 ± 3.97**
10%	52.90 ± 1.32	62.49 ± 1.39	59.61 ± 3.86	**64.61 ± 1.42**
20%	56.81 ± 0.73	62.70 ± 1.48	58.67 ± 1.48	**63.87 ± 0.72**
50%	57.21 ± 1.87	62.89 ± 1.80	61.00 ± 2.11	**62.93 ± 1.07**
100%	56.15 ± 1.92	60.82 ± 1.94	59.09 ± 2.22	**61.19 ± 0.45**

(b) EnMAP-DFC, FT-Macro

Labels	From-scratch	Tube	CoMask-NMF
10%	42.39 ± 0.53	45.23 ± 2.26	**47.28 ± 1.78**
25%	46.04 ± 1.12	45.98 ± 1.11	**47.68 ± 1.34**
50%	48.44 ± 0.30	47.03 ± 0.91	**49.95 ± 1.17**
100%	56.88 ± 1.96	51.80 ± 2.18	**57.43 ± 1.37**

Table 6: Attribution arms on EnMAP-DFC at ViT-Tiny. CoMask runs at the anchor configuration for these arms, dominant selection with $\rho_c = 0.875$, $\rho_s = 0.60$, $K = 8$ and NMF, which is not the fixed configuration of Tables 1 and 2. Each arm replaces one property of the masking, and every arm is matched to the anchor's input degradation of 0.7434 normalised input RMSE, to within 0.027. Every arm shares the anchor's spatial mask fraction of 0.5625, the input-space control included, so the control differs from the component arms in what it does at the masked locations rather than in how many it touches. The two $K = 8$ bases mask 7 of 8 components, the fixed frequency basis 8 of 200 and the wavelet basis 24 of 200, and the two noise arms use scale 0.99. Values are mean ± sd over five downstream seeds at one pretraining seed. The random-init row is a frozen, never-pretrained encoder, reported as a reference. Bold marks the best value in the two finetuning columns only, following Table 2, because no method separates from the random-init floor under probing here.

Arm	Input NRMSE	LP-OA	LP-Macro	FT-OA	FT-Macro
CoMask-NMF anchor	0.7434	81.92 ± 0.55	49.57 ± 1.35	**85.00 ± 0.54**	**57.19 ± 1.67**
Input-space mask	0.7500	82.88 ± 0.09	51.46 ± 1.31	83.14 ± 2.08	53.82 ± 3.91
White noise	0.7425	83.03 ± 0.23	50.90 ± 1.69	84.72 ± 0.84	56.90 ± 1.90
Covariance noise	0.7425	82.24 ± 0.31	49.91 ± 0.90	84.00 ± 1.20	55.15 ± 3.73
Random non-negative basis	0.7284	82.73 ± 0.58	50.13 ± 2.04	84.92 ± 0.74	56.93 ± 1.49
Band-permuted NMF basis	0.7172	82.29 ± 1.28	50.30 ± 1.36	84.29 ± 0.71	55.77 ± 2.44
Fixed frequency basis	0.7425	82.79 ± 0.51	50.81 ± 1.40	83.76 ± 0.93	54.68 ± 2.80
Wavelet basis	0.7471	82.32 ± 0.51	49.29 ± 1.27	84.15 ± 1.40	54.46 ± 4.18
Random-init floor	n/a	81.45 ± 0.72	45.30 ± 2.46	n/a	n/a

CoMask leads every arm in Table 6 on both finetuning columns. The margin over the matched input-space control is 3.37 points of macro accuracy, and the two fixed transform bases follow at 2.51 for the cosine basis and 2.73 for the wavelet basis. The two spectrally complete corruptions cost less, 2.04 for covariance noise and 1.42 for the band-permuted basis. Both of those margins fall inside the standard deviations of the rows they compare, so they carry less weight than the transform results. Two arms tie the anchor instead, a random non-negative basis at 0.26 points of macro accuracy and white noise at 0.29. Section 6 therefore scopes the fitted-basis result to Houston under linear probing.

Linear probing does not separate the arms on EnMAP, which reproduces the result that sets this dataset's primary metric in Section 4. Every arm in Table 6 falls in a 1.11-point band of overall accuracy, from 81.92 to 83.03, and that band includes the anchor. The band sits 0.5 to 1.6 points above the random-init floor of 81.45. Macro accuracy separates the arms from the floor more clearly, at 45.30 against a band of 49.29 to 51.46, but it does not order them either. That band is 2.17 points wide against standard deviations of 0.90 to 2.04.

C Basis Provenance

Appendix C reports where a CoMask basis can come from. Two results sit behind Section 6's provenance paragraph. The first asks how the fitted basis transfers across the corpus it was fitted over, and the second asks whether the source of the basis changes downstream accuracy.

The corpus basis transfers across the geography it was fitted over. Table 7 evaluates it against three region-specific donors on held-out pixels from each region, where it stays at or within 0.004 R^2 of

Table 7: Held-out R^2 of EnMAP bases across corpus regions. Rows are the basis and columns the evaluation region, with regions assigned from each scene's latitude and longitude. Each region donor is one NMF fit at $K = 8$ on 400,000 pixels from that region's fit tiles, evaluated on 150,000 pixels drawn from disjoint held-out tiles. The region splits are 3,964 fit and 699 held-out tiles for Africa and the Middle East, 1,349 and 238 for Asia, and 10,849 and 1,914 for Europe. The corpus basis is the 5M-pixel fit used for every EnMAP row in this paper, so donor against corpus cells vary sample size along with region. Bold marks the best value per column.

Basis	Africa, Middle East	Asia	Europe	Global
Africa and Middle East donor	**0.9871**	0.9617	0.9696	0.9771
Asia donor	0.7416	**0.9885**	0.8312	0.8383
Europe donor	0.9782	0.9700	**0.9866**	0.9826
Corpus basis	0.9832	0.9740	0.9860	**0.9846**

Table 8: Downstream effect of where the masking basis comes from. EnMAP-DFC at ViT-Tiny, linear probing in eval mode, one pretraining seed, mean $\pm$ sd over five downstream seeds. These arms use random selection with $\rho_c = 0.875$ and $\rho_s = 0.60$, which is neither the fixed configuration of Tables 1 and 2 nor the anchor of Table 6. They form their own matched-degradation family, within 0.005 of each other and about 0.077 below the anchor, so values here are not on a common degradation scale with Table 6. The random-init row is a frozen, never-pretrained encoder, reported as a reference. No value is bolded, because the four arms lie within one standard deviation of each other.

Basis used for masking	LP-OA	LP-Macro
Full-corpus NMF, $K = 8$	83.22 ± 0.18	51.00 ± 1.37
NMF fitted on Asia only	82.75 ± 0.36	50.88 ± 2.02
Curated USGS spectral library	82.42 ± 0.39	50.85 ± 1.85
Greedy USGS spectral library	82.78 ± 0.38	50.29 ± 1.65
Random-init floor	81.45 ± 0.72	45.30 ± 2.46

the best donor on every region, at 0.9832, 0.9740 and 0.9860, and holds the highest global value at 0.9846. A basis fitted on a single region carries no such guarantee. The Asia donor reaches 0.9885 on its own region and falls to 0.7416 and 0.8312 off it, while the other two donors stay within 0.01 of the corpus basis on every region. One-region fits therefore differ in how far they carry, and a donor that fits its own region best can still be the weakest elsewhere.

The three region bases differ from each other more than a corpus basis differs across fit sample sizes. Each basis's components are matched one to one against another's by Hungarian assignment on cosine similarity (Kuhn, 1955). No pair of region bases reaches a minimum matched cosine of 0.90. Refitting the corpus basis at sample sizes from 5 million to 30 million pixels holds that minimum between 0.844 and 0.931, as Appendix G reports. Geography therefore moves the components further than sample size does.

Where the basis comes from does not measurably change downstream accuracy. Table 8 replaces the corpus fit with a basis fitted on Asia alone and with two bases drawn from a curated spectral library (Kokaly et al., 2017). The full-corpus fit leads the other three by 0.44 to 0.80 points of overall accuracy and by 0.12 to 0.71 points of macro accuracy, and every one of those margins falls inside one standard deviation. All four also clear the random-init floor, by 1.0 to 1.8 points of overall accuracy and 5.0 to 5.7 points of macro accuracy. The four bases therefore tie as bases that work rather than as bases that do not.

Reconstruction quality does not track downstream quality on EnMAP. The curated library reconstructs held-out EnMAP spectra at 0.1411 normalised RMSE against the corpus fit's 0.0304, and it ties the corpus fit downstream all the same. These arms run on one dataset at one encoder scale, so the result holds for EnMAP rather than showing that fit quality never matters. Section 6 makes the related point that which decomposition supplies the structure best varies with the dataset and the encoder scale.

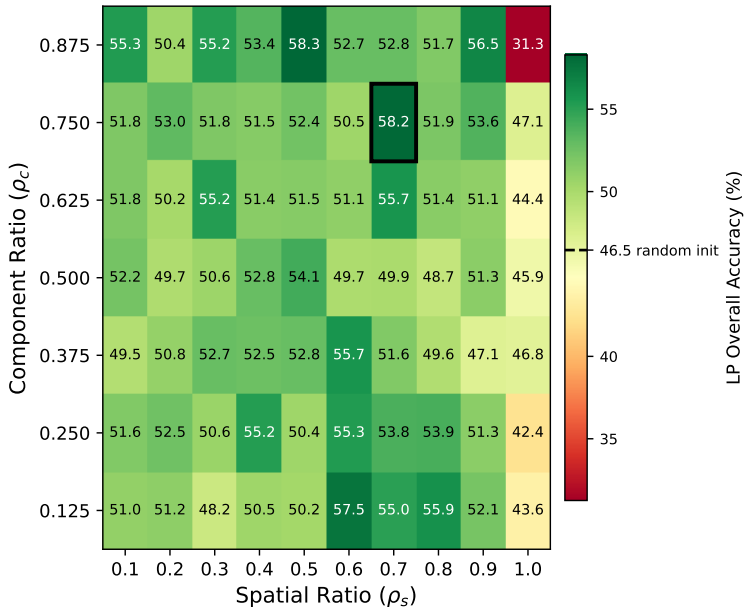


Figure 3: Linear-probe overall accuracy across the component ratio ρ_c and the spatial ratio ρ_s on Houston2018 at ViT-Tiny, for CoMask-NMF with dominant selection and $K = 8$. The fixed configuration at $\rho_c = 0.75$ and $\rho_s = 0.70$ is outlined, and the random-init probing floor of 46.45 is marked on the colour bar.

D CONFIGURATION SENSITIVITY

Appendix D reports how CoMask's masking configuration was chosen and how much the method depends on it. The grid is Houston2018 at ViT-Tiny with NMF at $K = 8$. Seven component ratios from 0.125 to 0.875 cross with ten spatial ratios from 0.1 to 1.0, giving 70 cells for each selection strategy. Each cell is one pretraining run at seed 5, probed with the same five downstream seeds and the same harness as the main tables.

The fixed configuration used in every Houston2018 and EnMAP row, dominant selection with $\rho_c = 0.75$ and $\rho_s = 0.70$, was chosen on this grid, which is scored on test accuracy. It ranks second among the 63 dominant cells with $\rho_s < 1$, at 58.19 ± 0.62. The median over those cells is 51.80, and three of them exceed the strongest input-space cell of Table 1 at 57.11.

Removing every intact spectrum costs more than any other configuration choice. The fourteen cells at $\rho_s = 1.0$ average 41.5 across the two selection strategies, against 52.1 for the cells with $\rho_s < 1$. Twelve of those fourteen fall below the random-init probing floor of 46.45, and the two that do not sit within 0.7 of it. EnMAP repeats the pattern on macro accuracy. Its ten cells at $\rho_s = 1.0$ average 39.70, against 49.86 for the 52 cells at lower spatial ratios. All ten fall under the EnMAP floor of 45.30.

Within $\rho_s < 1$ the grid shows no gradient in either ratio. Overall accuracy ranges from 47.08 to 58.29 across the 63 dominant cells, and neighbouring cells differ by 3 to 7 points. Each cell is a single pretraining run, and the spread across three pretraining seeds at the fixed configuration is 1.13, so part of the cell-to-cell variation is seed noise.

SPICE was also pretrained on EnMAP at ViT-Small and ViT-Base, at a configuration that is not the fixed one. Table 9 reports those two runs beside the corresponding NMF rows from Table 2. Because the two arms differ in the decomposition, the selection strategy and both ratios at once, no difference between them is attributable to the basis. The table is therefore coverage of which decompositions were run on EnMAP rather than a comparison between them.

Table 9: Further SPICE runs on EnMAP-DFC at ViT-Small and ViT-Base. These two runs use random selection with $\rho_c = 0.625$ and $\rho_s = 0.60$, so they differ from the CoMask rows of Table 2 in the decomposition, the selection strategy and both ratios at once. The CoMask-NMF rows are the fixed configuration, reproduced from Table 2 for context. No value is bolded, because the rows are not a basis comparison.

Architecture	Method	LP-OA	LP-Macro	FT-OA	FT-Macro
ViT-Small	CoMask-SPICE	81.69 ± 2.43	51.69 ± 1.96	84.39 ± 0.60	56.47 ± 1.27
	CoMask-NMF	82.40 ± 0.76	52.11 ± 2.05	84.70 ± 0.75	56.05 ± 2.01
ViT-Base	CoMask-SPICE	83.59 ± 0.45	53.40 ± 1.17	84.60 ± 0.33	54.55 ± 1.31
	CoMask-NMF	82.70 ± 1.03	52.67 ± 1.49	84.20 ± 0.58	54.09 ± 1.33

Table 10: EnMAP-DFC under class-weighted cross-entropy, across encoder scales. Both the linear probe and the finetuning use inverse-frequency class-weighted cross-entropy, so every row is a separately trained downstream model rather than a re-evaluation of Table 2, and every row loads the same pretraining checkpoint as its counterpart there. Methods, encoder scales and the cell-selection rule are Table 2's, and parameter counts are omitted because they are unchanged. CoMask uses the fixed configuration (dominant selection, $\rho_c = 0.75$, $\rho_s = 0.70$, $K = 8$). Input-space rows use block masking at ratio 0.5 and tube masking at 0.7. Values are mean ± sd over five downstream seeds at one pretraining seed. The random-init row is a frozen, never-pretrained encoder, reported as a reference. Bold marks the best finetuning value per column and architecture. The probing columns carry no bold, following Table 2.

Architecture	Method	LP-OA	LP-Macro	FT-OA	FT-Macro
ViT-Tiny	From-scratch	n/a	n/a	78.62 ± 0.88	61.71 ± 0.99
	Random-init floor	75.55 ± 0.81	50.90 ± 1.24	n/a	n/a
	MaskedSST-Block	77.43 ± 1.69	54.74 ± 1.47	77.48 ± 1.38	**63.28 ± 0.97**
	MaskedSST-Tube	74.09 ± 2.13	52.31 ± 2.10	75.09 ± 2.27	56.41 ± 3.38
	CoMask-PCA	76.49 ± 0.71	54.30 ± 1.18	77.72 ± 1.03	62.82 ± 0.33
	CoMask-NMF	74.22 ± 1.80	54.91 ± 1.66	**78.80 ± 0.77**	63.07 ± 0.53
	CoMask-SPICE	76.61 ± 1.63	54.74 ± 1.53	77.81 ± 1.51	63.19 ± 0.90
ViT-Small	From-scratch	n/a	n/a	77.64 ± 1.60	60.55 ± 0.89
	Random-init floor	75.06 ± 2.31	56.10 ± 2.10	n/a	n/a
	MaskedSST-Tube	77.14 ± 1.01	55.36 ± 0.29	75.41 ± 3.40	58.27 ± 2.78
	CoMask-NMF	77.13 ± 1.51	58.17 ± 1.75	**78.45 ± 0.85**	**61.54 ± 0.73**
ViT-Base	From-scratch	n/a	n/a	72.75 ± 2.78	47.45 ± 2.05
	Random-init floor	73.97 ± 3.79	56.54 ± 2.58	n/a	n/a
	MaskedSST-Tube	77.72 ± 0.51	50.75 ± 0.80	**76.68 ± 1.90**	57.07 ± 2.13
	CoMask-NMF	75.06 ± 1.88	57.47 ± 2.45	76.36 ± 1.06	**57.62 ± 5.64**

E CLASS-WEIGHTED ENMAP

Appendix E reports EnMAP-DFC under inverse-frequency class-weighted cross-entropy at the downstream stage, an objective Table 2 does not use. The change lowers overall accuracy on every row and raises macro accuracy on every row but one, supervised training from scratch at ViT-Base, so Table 10 reports a second protocol rather than a better version of the first.

CoMask-NMF reaches the highest finetuning overall accuracy at ViT-Tiny under this objective, at 78.80 ± 0.77, while the macro-accuracy lead it holds in Table 2 does not survive the change. MaskedSST-Block takes finetuning macro accuracy at 63.28 ± 0.97 against CoMask-NMF's 63.07 ± 0.53, a margin of 0.21 that sits well inside both standard deviations, and it leads probing overall accuracy by 3.21. Which decomposition supplies the basis does not change that, since the three CoMask rows span 0.37 points of finetuning macro accuracy, at 62.82, 63.07 and 63.19. Four of the five pretrained rows still exceed supervised training from scratch on macro accuracy, which reaches 61.71 ± 0.99, with MaskedSST-Tube the exception at 56.41.

CoMask-NMF leads input-space masking at both larger scales under this objective, and by wider margins than at ViT-Tiny. At ViT-Small it leads MaskedSST-Tube by 2.81 points of probing macro accuracy, 3.04 of finetuning overall accuracy and 3.27 of finetuning macro accuracy. At ViT-Base it leads by 6.72 points of probing macro accuracy, while the two finetuning columns split, with MaskedSST-Tube ahead by 0.32 on overall accuracy and CoMask-NMF ahead by 0.55 on macro

Table 11: Houston2018 with a Conv3D backbone, across encoder scales. CoMask uses the fixed configuration (dominant selection, $\rho_c = 0.75$, $\rho_s = 0.70$, $K = 8$), and the input-space control zeros whole pixel spectra at $\rho_s = 0.625$. From-scratch is trained with the same dense grid patching that the pretrained rows are pretrained and evaluated with, so the three families differ only in pretraining. No input-space control was pretrained at Conv3D-Base. No random-init floor is reported, because a frozen, never-pretrained Conv3D read at chance in eval mode, 4.98 overall accuracy, its batch-normalization statistics never having been estimated from data, which makes the value a measurement artifact rather than a floor. These probing columns are therefore not comparable with the ViT floors of Table 1. †Off-rule, using random rather than dominant selection at the same two ratios, reported because it is the row an earlier version of this work printed. Values are mean ± sd over five downstream seeds at one pretraining seed. Bold marks the best value per column and architecture.

Architecture	Method	LP-OA	LP-Macro	FT-OA	FT-Macro
Conv3D-Tiny	From-scratch	n/a	n/a	**53.52 ± 1.32**	**50.89 ± 1.27**
	Input-space control	38.11 ± 1.70	36.26 ± 1.83	52.93 ± 1.17	49.71 ± 1.29
	CoMask-NMF	**46.09 ± 1.72**	**41.73 ± 1.40**	51.39 ± 1.98	48.33 ± 1.48
Conv3D-Small	From-scratch	n/a	n/a	**53.61 ± 1.75**	**50.86 ± 1.84**
	Input-space control	35.62 ± 0.56	31.48 ± 0.40	51.36 ± 0.94	47.45 ± 1.55
	CoMask-NMF	**36.98 ± 0.82**	33.70 ± 0.73	50.85 ± 2.00	47.73 ± 1.32
	CoMask-NMF†	36.45 ± 0.30	**33.80 ± 0.15**	53.55 ± 2.86	49.14 ± 3.40
Conv3D-Base	From-scratch	n/a	n/a	**55.23 ± 0.75**	**52.61 ± 1.13**
	CoMask-NMF	**33.79 ± 1.54**	**30.42 ± 1.40**	53.66 ± 3.45	49.16 ± 3.69

accuracy against a standard deviation of 5.64. The random-init floor moves under this objective as well, rising to 56.10 and 56.54 points of probing macro accuracy at the two scales. CoMask-NMF clears that floor by 2.07 at ViT-Small and by 0.93 at ViT-Base, the second inside the floor's own standard deviation of 2.58, while MaskedSST-Tube falls 0.74 and 5.79 below it.

F CONV3D BACKBONE

Appendix F reports CoMask with a Conv3D backbone, the one architecture family outside the ViT encoders of Sections 5 and 6. Houston2018 carries both an input-space control and a from-scratch baseline at three scales, while EnMAP carries neither a control nor a pretraining run at the fixed configuration, so Tables 11 and 12 answer different questions and are read separately.

CoMask teaches a frozen Conv3D encoder more than an input-space mask does, at both scales where a control was pretrained. Under probing it leads the control by 7.98 points of overall accuracy at Conv3D-Tiny and by 1.36 at Conv3D-Small, with the off-rule run at Small leading by 0.83. Neither margin can be placed against an absolute reference, because no usable random-init floor exists for this backbone, so these are comparisons between two pretraining strategies rather than evidence that either clears chance.

CoMask finetunes close to grid-matched from-scratch training at every Conv3D scale on Houston2018 without reaching it. The gap runs 2.13, 2.76 and 1.57 points of overall accuracy from Conv3D-Tiny to Conv3D-Base, and 2.56, 3.13 and 3.45 of macro accuracy. The input-space control does not reach from-scratch at either scale where it was pretrained. The off-rule run at Conv3D-Small is the exception among the pretrained rows, finetuning to 53.55 ± 2.86 against from-scratch's 53.61 ± 1.75, and nothing on record accounts for why random selection should do better there than the dominant selection the fixed configuration uses.

CoMask leads supervised training from scratch on finetuning macro accuracy at both EnMAP Conv3D scales under the unweighted objective, by 1.84 at Conv3D-Small and 3.63 at Conv3D-Base, while from-scratch holds overall accuracy by 4.45 and 5.95. Class weighting brings overall accuracy to a tie, with CoMask ahead by 1.54 at Conv3D-Small and from-scratch ahead by 1.36 at Conv3D-Base, both margins inside the standard deviations of the rows they compare. Macro accuracy splits under that objective, with from-scratch ahead by 0.78 at Conv3D-Small, inside its own standard deviation of 1.67, and CoMask ahead by 3.17 at Conv3D-Base. Each margin is taken against the strongest of the three CoMask configurations in that column. No Conv3D pretraining run exists at the fixed configuration on EnMAP, and no input-space Conv3D baseline was pretrained there, so the three rows are the configurations available rather than a comparison between them.

Table 12: EnMAP-DFC with a Conv3D backbone, under both downstream objectives. No Conv3D pretraining run exists at the fixed configuration on EnMAP and no input-space Conv3D baseline was pretrained there, so every CoMask row uses random selection and the table compares the three available configurations against supervised training from scratch. Ratio pairs are (ρ_c, ρ_s), all at $K = 8$ with NMF. No random-init floor is reported, for the reason given in Table 11; the class-weighted probe on a never-pretrained Conv3D read 16 macro accuracy, the same artifact. Values are mean ± sd over five downstream seeds at one pretraining seed. Bold marks the best finetuning value per column, architecture and objective. The probing columns carry no bold, following Table 2.

Architecture	Method	LP-OA	LP-Macro	FT-OA	FT-Macro
Unweighted cross-entropy					
Conv3D-Small	From-scratch	n/a	n/a	**81.95 ± 2.13**	48.44 ± 2.60
	CoMask-NMF (0.75, 0.70)	78.76 ± 1.27	40.11 ± 1.32	73.00 ± 5.40	48.40 ± 1.75
	CoMask-NMF (0.625, 0.60)	79.66 ± 1.10	42.15 ± 1.60	71.43 ± 6.00	49.14 ± 1.95
	CoMask-NMF (0.625, 0.80)	79.90 ± 0.98	40.26 ± 1.23	77.50 ± 3.19	**50.28 ± 2.29**
Conv3D-Base	From-scratch	n/a	n/a	**82.46 ± 1.32**	46.52 ± 4.00
	CoMask-NMF (0.75, 0.70)	79.21 ± 1.28	40.07 ± 1.69	76.40 ± 5.13	46.04 ± 2.61
	CoMask-NMF (0.625, 0.60)	79.90 ± 1.10	42.29 ± 1.16	71.83 ± 2.93	**50.15 ± 1.40**
	CoMask-NMF (0.625, 0.80)	79.34 ± 1.65	39.53 ± 1.64	76.51 ± 2.84	49.66 ± 2.02
Class-weighted cross-entropy					
Conv3D-Small	From-scratch	n/a	n/a	73.05 ± 3.04	**52.84 ± 1.67**
	CoMask-NMF (0.75, 0.70)	69.38 ± 4.43	44.24 ± 0.90	73.09 ± 2.21	48.69 ± 1.54
	CoMask-NMF (0.625, 0.60)	70.30 ± 0.78	46.03 ± 1.08	74.39 ± 2.61	52.06 ± 0.76
	CoMask-NMF (0.625, 0.80)	73.52 ± 2.84	44.63 ± 0.91	**74.59 ± 1.48**	48.26 ± 2.33
Conv3D-Base	From-scratch	n/a	n/a	**74.08 ± 2.45**	48.86 ± 2.35
	CoMask-NMF (0.75, 0.70)	71.71 ± 3.05	45.04 ± 0.68	69.96 ± 5.34	49.54 ± 1.46
	CoMask-NMF (0.625, 0.60)	72.37 ± 2.14	46.40 ± 0.68	72.10 ± 2.70	**52.03 ± 1.07**
	CoMask-NMF (0.625, 0.80)	74.19 ± 1.90	45.22 ± 0.78	72.72 ± 2.83	50.33 ± 1.03

Table 13: Fit quality of the EnMAP NMF decomposition by fit sample size and component count. Each basis is fitted on a uniform random pixel sample of the stated size and evaluated on all 19,013 EnMAP training tiles, 77.9 million pixels, which include the fit sample, 6.4% of the evaluation pixels at the 5M fit. The values are therefore in-sample fit quality and not held-out generalisation. Coefficients for evaluation come from the clipped pseudoinverse of Appendix I.1. The 5M fit at $K = 8$ is the basis used for every EnMAP row in this paper. No value is bolded, because the two panels disagree on which component count fits best.

Fit sample	$K = 4$	$K = 6$	$K = 8$	$K = 10$	$K = 12$
Global R^2					
5M	0.9588	0.9769	0.9860	0.9854	0.9743
10M	0.9588	0.9781	0.9857	0.9843	0.9729
20M	0.9588	0.9781	0.9854	0.9791	0.9768
30M	0.9588	0.9786	0.9853	0.9775	0.9382
Mean per-pixel RMSE ± sd across pixels, in units of 10^{-3}					
5M	3.89 ± 2.69	2.65 ± 2.35	2.17 ± 1.70	1.92 ± 2.06	2.03 ± 3.13
10M	3.89 ± 2.69	2.70 ± 2.14	2.18 ± 1.73	1.96 ± 2.16	2.12 ± 3.19
20M	3.89 ± 2.69	2.70 ± 2.15	2.20 ± 1.76	2.07 ± 2.66	1.90 ± 3.00
30M	3.89 ± 2.69	2.69 ± 2.09	2.20 ± 1.77	2.09 ± 2.80	2.79 ± 5.08

G Decomposition Fit by Sample Size and Component Count

Appendix G reports how the EnMAP decomposition depends on the number of pixels it is fitted on and on the number of components it uses. The paper's EnMAP basis is the 5 million pixel fit at $K = 8$, so the question is what would change if either quantity moved. Both tables evaluate every basis on all EnMAP training tiles, including the fit sample, which makes them a measure of fit rather than of generalisation.

The $K = 8$ fit is insensitive to how many pixels it is fitted on. Global R^2 moves from 0.9860 at 5 million pixels to 0.9853 at 30 million, and mean per-pixel RMSE from 2.17 to 2.20, so a

Table 14: Stability of the EnMAP basis across fit sample sizes. Each atom of the 5M basis is matched one to one against an atom of the larger-sample basis at the same component count, by the Hungarian assignment on cosine similarity used in Appendix C. Values are the minimum and mean matched cosine over atoms, where the minimum is what says whether every atom survives refitting. The $K = 8$ row is the component count used for every EnMAP row in this paper.

Components	5M against 10M	5M against 20M	5M against 30M
$K = 4$	1.000 / 1.000	1.000 / 1.000	1.000 / 1.000
$K = 6$	0.279 / 0.860	0.256 / 0.857	0.261 / 0.858
$K = 8$	0.844 / 0.962	0.903 / 0.971	0.931 / 0.976
$K = 10$	0.879 / 0.956	0.675 / 0.898	0.401 / 0.882
$K = 12$	0.807 / 0.954	0.879 / 0.942	0.854 / 0.944

Table 15: Cost of the decomposition pipeline, per dataset. Pixel sampling on Spectral Earth ran as 20 single-CPU tasks in parallel, 22.3 CPU-hours in total. The fits ran on 8 CPUs for Houston2018 and EnMAP and on 4 for the Spectral Earth MiniBatchNMF, the repack on 32 CPUs, and the projections on one GPU for EnMAP and one B200 for Spectral Earth. Houston2018 needs no projection step because its solver returns the coefficients for the pixels it was fitted on, and neither Houston2018 nor EnMAP needs a repack. Timings for Houston2018 and EnMAP were re-measured on current hardware rather than taken from the original runs, whose timings were not kept.

Step	Houston2018	EnMAP	Spectral Earth
Pixel sampling	0.7 s	4.0 min	1.41 h
Decomposition fit	2.3 min	12.1 min	4.61 h
Data repack	n/a	n/a	0.78 h
Coefficient projection	none	1.69 h	5.04 h
Total	2.3 min	1.96 h	11.84 h

sixfold increase in the fit sample leaves the two within 0.0007 and 0.03 of each other. Neither panel establishes a best component count. $K = 8$ holds the highest R^2 at every sample size while $K = 10$ reaches the lowest mean RMSE at every sample size, so the two criteria disagree and neither selects a component count on its own. The largest count is the least stable of the five as the sample grows, with $K = 12$ falling to 0.9382 R^2 at 30 million pixels.

The $K = 8$ atoms survive refitting as well as the fit does. Every atom of the 5 million pixel basis keeps a counterpart at cosine 0.844 or above against the 10 million fit and 0.931 or above against the 30 million fit, at means of 0.962 and 0.976. That stability does not extend to every component count. At $K = 6$ one atom has no counterpart at any sample size, at a minimum cosine of 0.256 to 0.279, and at $K = 10$ the match degrades as the sample grows, from 0.879 against the 10 million fit to 0.401 against the 30 million one. The $K = 4$ basis is identical at every sample size, at cosine 1.000, which the record does not explain.

H Decomposition Compute Footprint

Appendix H breaks the cost of the decomposition into its steps. Section 4 states what that cost comes to as a share of pretraining, and this appendix gives the steps behind the totals and the pretraining times they are measured against.

Projecting the coefficients costs more than fitting the basis, on both datasets where a projection step exists. On EnMAP the projection takes 1.69 of the pipeline's 1.96 hours, against 12.1 minutes for the fit itself. On Spectral Earth the two are closer, at 5.04 and 4.61 hours of an 11.84-hour pipeline. Houston2018 needs no projection at all, because its solver returns the coefficients for the pixels it was fitted on, which leaves a 2.3-minute fit as the whole cost.

The share of pretraining that the decomposition takes is largest where pretraining is shortest. It falls from 2.8% at ViT-Tiny to 0.5% at ViT-Base on Houston2018, and from 19.7% to 6.8% on EnMAP, in both cases because the pipeline cost is fixed while pretraining lengthens with encoder capacity.

Table 16: Pretraining wall time and the share of it that the decomposition takes. Each pretraining time is for the run this paper reports, on one B200 except Spectral Earth, which used four. The Spectral Earth figure is estimated from checkpoint timestamps at about 24 minutes per epoch over 300 epochs, excluding a 1.5-day idle gap, rather than read from a single scheduler record. ‡On Spectral Earth the share of GPU-hours is about 1.0%, because the sampling, fit and repack steps run on CPUs while pretraining occupies four B200s.

Dataset	ViT-Tiny	ViT-Small	ViT-Base
Pretraining wall time			
Houston2018	1.35 h	4.02 h	7.32 h
EnMAP	9.95 h	13.66 h	28.92 h
Spectral Earth	about 120 h	n/a	n/a
Decomposition as a share of pretraining wall time			
Houston2018	2.8%	1.0%	0.5%
EnMAP	19.7%	14.4%	6.8%
Spectral Earth	9.8%‡	n/a	n/a

Algorithm 1 One CoMask pretraining step.

Require: Encoder–decoder $f_{\boldsymbol{\theta}}$
Require: Component matrix $\boldsymbol{B} \in \mathbb{R}^{K \times C}$, fitted once per dataset and held fixed
Require: Coefficient cube $\mathbf{A}$, precomputed per tile and loaded as a fixed buffer; the rule used for each dataset is given below
Require: Ratios ρ_{c}, ρ_{s}, selection strategy π

Sample a batch $\{(\mathbf{X}_b, \mathbf{A}_b)\}_{b=1}^{B}$
for each sample b in the batch **do**
  $\boldsymbol{m}_b \leftarrow \pi(\mathbf{A}_b, \rho_{\mathrm{c}})$ ▷ one K-vector per sample
  $\tilde{\mathbf{A}}_b \leftarrow \mathbf{A}_b \odot \boldsymbol{m}_b$ ▷ broadcast over the $H \times W$ locations
  $\tilde{\mathbf{X}}_b \leftarrow \tilde{\mathbf{A}}_b \boldsymbol{B}$ ▷ partial reconstruction, $H \times W \times C$
  if $\rho_{\mathrm{s}} < 1$ **then**
    Draw $\mathcal{M}_b$ uniformly with $|\mathcal{M}_b| = \lfloor \rho_{\mathrm{s}} HW \rfloor$
    $\tilde{\mathbf{X}}_b[(i,j) \notin \mathcal{M}_b] \leftarrow \mathbf{X}_b[(i,j)]$ ▷ keep the original spectrum outside $\mathcal{M}_b$
  end if
end for
$\mathcal{L} \leftarrow \dfrac{1}{BHW} \sum_{b,i,j} \left\| f_{\boldsymbol{\theta}}(\tilde{\mathbf{X}})_b(i,j) - \mathbf{X}_b(i,j) \right\|_1$
Update $\boldsymbol{\theta}$ by a gradient step on $\mathcal{L}$

On Spectral Earth the wall-time share is 9.8% and the share of GPU-hours about 1.0%, since the sampling, fit and repack steps run on CPUs while pretraining occupies four B200s.

I Implementation Details

Algorithm 1 gives one pretraining step in the notation of Section 3. The components and the per-pixel coefficients are computed offline, once per dataset, and enter the step as fixed buffers, so nothing in the loop depends on the decomposition beyond a matrix product.

I.1 Coefficient computation

Section 3 states that each decomposition supplies its coefficients through its own transform. The concrete rule differs by dataset, and the differences are recorded here rather than smoothed over, because no single expression covers all five cases.

On EnMAP the coefficients are the clipped pseudoinverse solution,

$$\boldsymbol{a} = \max\bigl(0, (\boldsymbol{B}^{\top})^{\dagger} \boldsymbol{x}\bigr), \tag{4}$$

where $(\boldsymbol{B}^{\top})^{\dagger}$ is the pseudoinverse of $\boldsymbol{B}^{\top}$. The clamp is active at 64% of pixels, and the result sits 1.55% in relative L_2 from the exact non-negative least squares solution. On Spectral Earth the

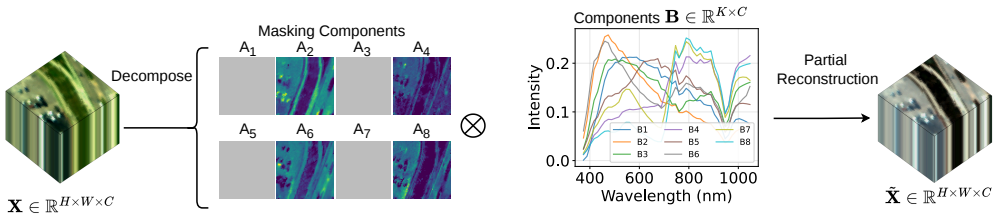


Figure 4: Component masking at the per-component level, for a Houston2018 patch with $K = 8$ and $\rho_c = 0.5$. The cube $\mathbf{X}$ is decomposed into eight coefficient maps $\mathbf{A}_k$ and the component matrix $\boldsymbol{B}$, whose rows are plotted at right. Four components are masked, shown greyed, and the retained four recombine into the partial reconstruction $\tilde{\mathbf{X}}$. The figure shows $\rho_s = 1$, so every location receives the partial reconstruction.

coefficients come from 50 multiplicative non-negative least-squares updates with the basis held fixed, stored in half precision. On Houston2018 the NMF coefficients are the solver's own non-negative transform, 1.7% in relative L_2 from exact non-negative least squares. The SPICE coefficients are that method's own abundances, non-negative and summing to one at every pixel. The PCA coefficients are the signed, unconstrained projection, and 47% of them are negative, which is why non-negativity cannot be stated as a property of CoMask coefficients in general.

I.2 Further selection strategies

Random and dominant selection fix the number of masked components and let the resulting degradation vary. Two strategies in the released code invert that. RMSE-weighted selection takes a target normalized RMSE and removes components in random order until the partial reconstruction is closest to it. Abundance-weighted selection takes a target fraction of the total coefficient magnitude at each pixel and accumulates components until it is reached. Both let the number of masked components vary across samples, and neither was used in any experiment reported in this paper.

J Per-Dataset Details

Appendix J gives the per-dataset detail that Table 17 summarises, together with the band handling that Section 3 refers to. Each dataset contributes a different condition to the evaluation, so each subsection states what that condition is and how the bands reach the encoder.

J.1 Houston2018

Houston2018 is the IEEE GRSS DFC 2018 scene over the University of Houston (Xu et al., 2019), acquired by a CASI 1500 sensor at 1 m ground sampling distance over 48 bands in the visible and near-infrared. The official spatial split gives 504,712 labeled training pixels and 85,437 labeled test pixels. Pretraining draws all 8×8 patches from the training region without using labels. Within the labeled training pixels, 10% are held out for validation, and that holdout is fixed across the five evaluation seeds, so model selection does not vary with the seed.

The scene's 20 classes are moderately imbalanced, in contrast to EnMAP, where one class holds two thirds of the labeled pixels. Non-Residential dominates the labeled pool while Water and Unpaved Parking hold a few thousand pixels each, as Figure 5 shows across all 20.

The 48 CASI bands need no removal, and two zero-valued bands are appended at load time so that the band count divides by the spectral patch size of 10. The decomposition's components are zero-padded along the band axis to match, which keeps the components and the encoder input on one band set.

The input-space sweep covers eleven masking ratios, 0.02, 0.03, 0.05, 0.08, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60 and 0.70, in each of the two masking modes, which gives the 22 cells of Section 5. The sweep also contains ratios 0.8 and 0.9, which are left out of that comparison. Both mask every location in the tile and leave the probe at chance, so they are the sweep's two weakest cells and leaving them out makes the comparison harder for CoMask rather than easier.

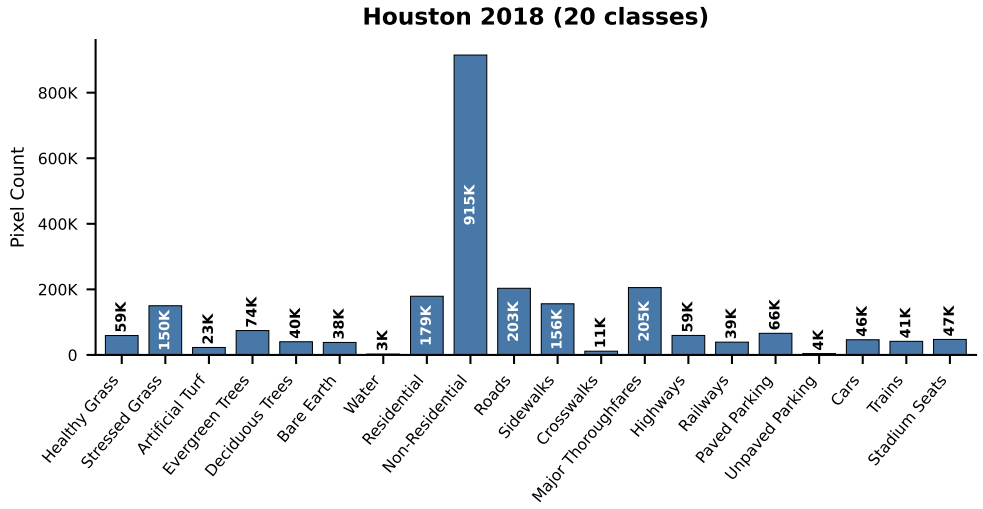


Figure 5: Per-class labeled pixel counts on Houston2018 across the full labeled scene, showing the dominant Non-Residential class and the tail of small classes.

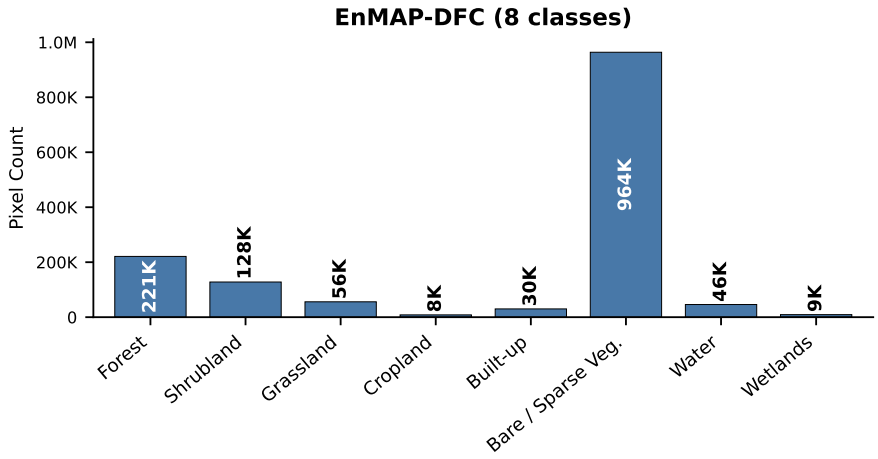


Figure 6: Per-class labeled pixel counts on EnMAP-DFC. Bare and Sparse Vegetation is an order of magnitude more frequent than any other class, which is what macro accuracy is reported against.

J.2 EnMAP and EnMAP-DFC

EnMAP and EnMAP-DFC are a paired pretraining collection and evaluation subset, both built from EnMAP satellite scenes (Guanter et al., 2015; Storch et al., 2023) and released together by Scheibenreif et al. (2023). The imaging spectrometer records 224 bands from 420 to 2450 nm at 30 m ground sampling distance. Removing 22 water-absorption bands at indices 126 to 140 and 160 to 166, together with 2 noisy edge bands at 222 and 223, leaves the 200 bands both datasets use. That removal list is applied when the decomposition is fitted and when either dataset is read, so the components and the encoder input stay on one band set.

The pretraining collection is 90 cloud-free Level-2 scenes over Europe acquired in the fourth quarter of 2022, tiled without overlap into the 19,013 tiles of Table 17 and used without labels. The labeled subset comes from elsewhere. EnMAP-DFC holds 357 tiles from two scenes over Mexico City, split into 286 training and validation tiles and 71 test tiles, with land-cover labels from the DFC 2020 challenge (Schmitt et al., 2019). Pretraining on European scenes and evaluating over Mexico City therefore places a geographic domain shift between the two stages.

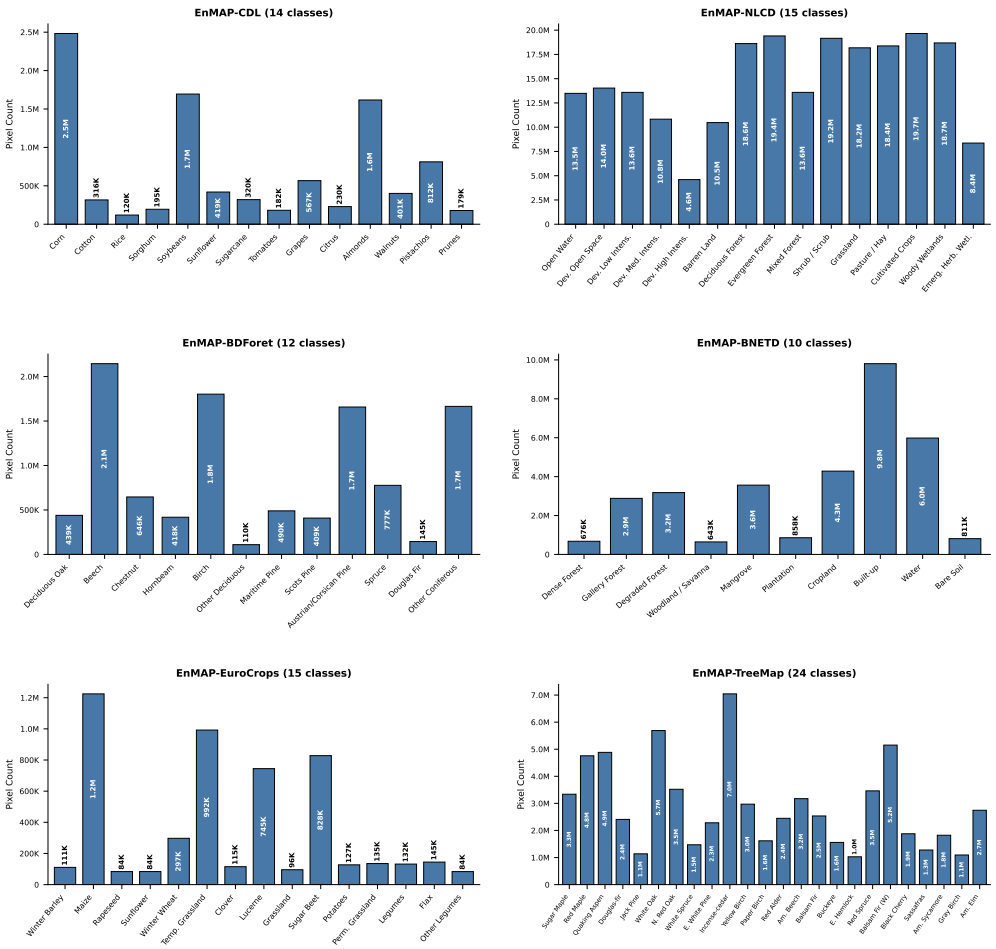


Figure 7: Per-class labeled pixel counts across the six Spectral Earth segmentation tasks. The counts and the shape of the distribution vary across tasks, with EnMAP-NLCD nearly balanced and others carrying long tails.

The eight classes are severely imbalanced, which is why Section 4 makes macro accuracy the primary metric here. Bare and Sparse Vegetation accounts for roughly two thirds of the labeled pixels, and Cropland, the smallest class, for about 8,000, a spread Figure 6 shows across all eight.

J.3 Spectral Earth

Spectral Earth (Braham et al., 2025) is built from EnMAP scenes at roughly 11,636 globally distributed locations. Its 538,000 patches of 128×128 pixels at 30 m ground sampling distance are released with the 22 water-absorption bands already removed, leaving 202. Two noisy edge bands at indices 200 and 201, which are the original EnMAP bands 222 and 223, are dropped for consistency with the EnMAP pipeline, leaving 200. This work uses the release downloaded in April 2026.

The six segmentation tasks released with the corpus span crop type, land cover, forest cover and tree species. They are EnMAP-CDL at 14 classes and EnMAP-NLCD at 15, both over the United States, EnMAP-EuroCrops at 15 over Europe, EnMAP-BNETD at 10, EnMAP-BDForet at 12 over France, and EnMAP-TreeMap at 24 over the United States. Each task uses the training, validation and test splits the corpus provides. Figure 7 shows how far the six differ in class balance, from the nearly balanced EnMAP-NLCD to the long tails of EnMAP-CDL and EnMAP-EuroCrops.

Table 17: The three benchmarks. †Pretraining on the EnMAP collection, downstream evaluation on the labeled EnMAP-DFC subset. ‡PCA and SPICE at ViT-Tiny only under the fixed configuration; further SPICE runs at ViT-Small and ViT-Base are reported in Appendix D.

Dataset	Bands	Tile size	# Tiles	Pretraining pixels	Decompositions	Downstream task
Houston2018	48	8×8	1 scene	1.4M	PCA, NMF, SPICE	20-class classification
EnMAP†	200	64×64	19,013	77M	PCA, NMF, SPICE‡	8-class classification
Spectral Earth	200	128×128	538,000	8.8B	NMF	6 segmentation tasks