MorphJEPA: Morphology-Aware Latent Prediction for Hyperspectral Images

Amartya Ray[1], Tanmay Mandaliya[1], Muhammad Haris Khan[2], and Biplab Banerjee[1]

[1] Indian Institute of Technology Bombay, Mumbai, India 24d1383@iitb.ac.in, tanmaymandaliya067@gmail.com, getbiplab@gmail.com

[2] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates muhammad.haris@mbzuai.ac.ae

Abstract. We study *learning from limited supervision* (LLS) for hyperspectral imagery (HSI): learning transferable representations from abundant unlabeled HSIs in the target sensing domain and then adapting with only a handful of labels per class. Joint-Embedding Predictive Architectures (JEPAs) are appealing for LLS because they learn via latent prediction rather than pixel reconstruction; however, they remain unexplored for HSI. Generic masking or cropping is ill-suited to HSI: band masking corrupts high-dimensional spectra, spatial masking removes entire pixel-level semantics, and dense spatial–spectral extraction can yield highly correlated, degenerate prediction targets. We propose **MorphJEPA**, the first JEPA tailored to HSI, which uses morphology as the pretext signal by predicting embeddings of pixel-aligned targets generated by per-band morphological opening/closing. This induces a directional structural bottleneck that preserves core topology while suppressing high-frequency nuisances. MorphJEPA couples this objective with a compact disentangled spatial–spectral encoder fused by cross-attention and a context-anchored Sketched Isotropic Gaussian Regularizer (SIGReg) to prevent collapse and maintain a well-conditioned latent manifold. Extensive 1/5/10-shot evaluations show consistent gains in intra-scene settings (Houston 2013, Trento) and under cross-scene temporal/spatial shifts (Houston 2013 → 2018, HyRank) for both zero-shot transfer and low-shot adaptation. Project page: https://github.com/amartya-ray/MorphJEPA.

Keywords: Hyperspectral Imaging · Self-Supervised Learning · Joint-Embedding Predictive Architecture · Mathematical Morphology

1 Introduction

Hyperspectral Imaging (HSI) enables fine-grained material understanding for applications such as precision agriculture, environmental monitoring, and urban surveillance [2]. Unlike RGB, HSI captures a dense spatial–spectral cube with numerous contiguous bands [25]. Although this richness supports precise material discrimination, annotation is expensive and expertise-driven, leading to

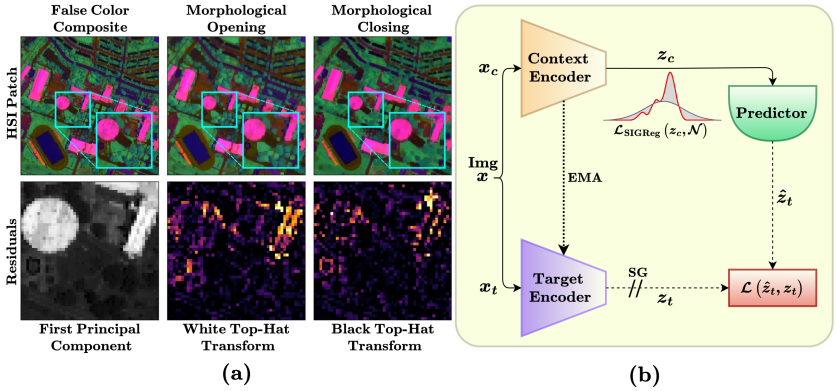

Fig. 1: MorphJEPA rationale and framework. (a) Morphological opening/closing preserve macro-geometry (e.g., building footprints) while attenuating unpredictable high-frequency residuals. **(b)** MorphJEPA predicts embeddings of morphologically smoothed, pixel-aligned targets from a context embedding, enforcing a structural bottleneck that favors transferable topological invariants, while a context-anchored regularizer prevents dimensional collapse.

severe label scarcity [4]. We therefore study *learning from limited supervision (LLS)*: learning transferable representations from abundant unlabeled HSIs in the target sensing domain, then adapting with only a few labeled samples per class [40]. Unlike few-shot learning [29, 32, 47, 48] or zero-shot learning [17, 21], LLS must extract transferable structure directly from unlabeled target data. This is particularly challenging in HSI, where high dimensionality aggravates overfitting (Hughes phenomenon) [45] and dense patch extraction induces latent redundancy that can obscure subtle class cues [44]. This raises a central question: *how can self-supervised pretraining learn representations that remain semantic and transferable under extreme low-shot supervision and scene shifts?*

Most HSI self-supervised learning (SSL) methods follow either contrastive [14, 20] or reconstruction-based [9, 19] paradigms. Contrastive methods rely on multiple *views*, but view design is fragile in HSI because spectral signatures themselves encode class identity [33]; augmentations that are benign in RGB may distort semantics in HSI. Reconstruction-based methods avoid negatives but encourage recovery of dataset-specific details that reconstruct well yet transfer poorly across scenes, seasons, or sensors. Pixel-level objectives may therefore overfit atmospheric or sensor noise rather than class-discriminative structure [25].

A key missing ingredient in prior HSI-SSL is an explicit notion of *structure* stable across sensing conditions. In many HSI scenes, class identity depends not only on spectra but also on macroscopic morphology—connectivity, boundaries, and footprint-scale geometry—which is often more invariant than high-frequency textures or sensor-specific noise [12]. Morphological operators [38] such as opening and closing provide a principled way to isolate this invariance by removing small-scale spatial fluctuations while preserving dominant topology. As shown in Fig. 1(a), they suppress residuals while retaining the structural "skeleton" of

semantic entities. This makes morphology a natural self-supervision signal for LLS: it yields *pixel-aligned* views that are controlled, semantically stable, and biased toward transferable structure.

Joint-Embedding Predictive Architectures (JEPAs) [6] learn by predicting in latent space, avoiding contrastive negatives and pixel-level reconstruction. Yet they remain unexplored for HSI because generic masking or cropping can erase material signatures and disrupt spatial–spectral coherence, yielding underdetermined targets. Furthermore, dense spatial–spectral patch extraction induces latent redundancies, encouraging shortcut correlations and representation collapse [35]. HSI therefore requires prediction targets that are *pixel-aligned* and *structurally simplified*, so nuisance variability is suppressed and the task remains well-posed. Morphology provides such targets, while JEPA offers the appropriate principle: predict *stable structure* in embedding space rather than reconstruct pixels or enforce invariance to brittle augmentations. At the same time, this spatial–spectral redundancy can push embeddings toward a low-rank, anisotropic subspace [10], motivating explicit regularization of the latent manifold.

These observations expose a clear gap: *Can morphology be coupled with JEPA to create pixel-aligned, structurally simplified targets that make nuisances hard to predict, thereby inducing representations that transfer under low-shot supervision and scene shift?*

Our approach. We introduce **MorphJEPA**, the first JEPA-style framework tailored to HSI. MorphJEPA predicts embeddings of *pixel-aligned* targets generated by per-band morphological opening/closing, together with lightweight geometry-preserving transforms. As illustrated in Fig. 1(b), predicting these morphologically smoothed targets from raw context imposes a *directional structural bottleneck*: nuisance details suppressed in the targets become difficult to predict, forcing the encoder to emphasize transferable topology and core material cues.

To support this, we use a compact disentangled spatial–spectral encoder fused by *cross-attention*. Unlike monolithic 3D-CNNs, this preserves the distinction between spatial structure and spectral identity; unlike simple concatenation, it enables conditional fusion between topological geometry and spectral evidence. We further stabilize training with a momentum (EMA) teacher [41] and a context-anchored Sketched Isotropic Gaussian Regularizer (SIGReg) [8] applied only to student context embeddings. Unlike LeJEPA's symmetric SIGReg, our asymmetric design keeps semantic targets anchored under severe spatial–spectral latent correlation. We also provide a compact theoretical analysis showing that morphology prediction reduces nuisance-induced representation drift, while SIGReg promotes well-conditioned, near-isotropic embeddings that improve low-shot generalization. Our **main** contributions are:

- We propose **MorphJEPA**, the first JEPA-style latent predictive SSL framework for hyperspectral imagery.
- We develop morphology-driven, pixel-aligned prediction targets for LLS, coupled with a lightweight disentangled spatial–spectral encoder with cross-attention fusion.

- We introduce a context-anchored distributional regularizer to mitigate collapse from spatial–spectral feature correlation, yielding a better-conditioned near-isotropic latent space and improved low-shot sample complexity.
- We validate MorphJEPA under 1-, 5-, and 10-shot protocols across intra-scene and cross-scene generalization under temporal and spatial shifts, consistently outperforming strong supervised and SSL baselines.

2 Related Works

(i) Self-Supervised Learning for HSI. HSI-SSL is dominated by *contrastive* and *reconstruction-based* objectives. Contrastive methods form positive pairs via spectral transformations (e.g., PCA splitting) [28] or multi-scale spatial sampling [49, 50], with guided variants introducing label-driven grouping [26]. Two limitations recur: *view fragility*—augmentations benign in RGB can alter material signatures (e.g., band permutation), violating semantic identity and suppressing discriminative spectra—and *training overhead* from large batches/negatives and extra heuristics. Reconstruction-based SSL, largely MAE-style [19], predicts masked bands [30] or spatial patches [24] with pixel-level losses, sometimes using LiDAR priors [42]. While effective, these objectives often encode dataset-specific correlations and high-frequency noise that fail under sensor/scene shifts; dependence on co-registered LiDAR also limits applicability.
(ii) Morphological Structure in HSI. Mathematical morphology has long been recognized as a powerful structural prior in hyperspectral analysis (see `Supp. Mat.` for a brief review of mathematical morphology operators). Extended Morphological Profiles (EMP) [16, 34] and their multiband extensions [11] capture spatial connectivity, region geometry, and boundary information that complement spectral features. More recently, morphological operators have been embedded as fixed or learnable layers in supervised networks [1, 37], demonstrating improved robustness to small-scale variations. These methods, however, employ morphology primarily as an *input-side* transformation or architectural bias. They do not leverage morphological evolution as a self-supervised objective, nor do they couple structural filtering with representation-level predictive learning.
(iii) Predictive Architectures in Remote Sensing. JEPA [6] offers an alternative to both contrastive and reconstruction-based SSL by predicting target embeddings in latent space (see `Supp. Mat.` for a brief primer on JEPA). This paradigm has recently been explored in remote sensing, including optical retrieval [15], multimodal multi-resolution modeling [7], and SAR imagery [27]. Most adaptations inherit spatial masking or cropping strategies from RGB image pretraining. However, for HSI, aggressive masking can remove semantically critical material evidence and sever spatial–spectral coherence, yielding underdetermined targets. Moreover, standard predictive targets do not explicitly account for severe latent correlations induced by dense spatial–spectral patch extraction.

In summary, prior HSI-SSL either enforces invariance to handcrafted perturbations or reconstructs raw pixels, while morphology-based methods remain largely supervised, and predictive SSL has not been tailored to HSI. We bridge

these lines of work by elevating morphology from an *input prior* to a *latent predictive pretext* within a JEPA framework: morphology shifts from "what is processed" to "what must be predicted" in embedding space.

3 Problem Setting

Given an HSI scene $\mathcal{X} \in \mathbb{R}^{H\times W\times B}$, we study LLS: we first pretrain on a large set of *unlabeled* spatial–spectral patches extracted from $\mathcal{X}$ in the *target* sensing domain, then learn a downstream classifier using only N labeled patches per class ($N \in \{1, 5, 10\}$). We evaluate both in-distribution adaptation (intra-scene) and out-of-distribution generalization (cross-scene). Our aim is to learn a transferable patch encoder $f_\theta : \mathbb{R}^{S\times S\times B} \to \mathbb{R}^D$, where $x_{i,j} \in \mathbb{R}^{S\times S\times B}$ denotes the $S \times S$ patch around pixel (i, j) over all B bands; symmetric padding by $\lfloor S/2 \rfloor$ along the spatial dimensions enables valid patch extraction near image boundaries.

Pretraining Data Construction. Unlabeled patch centers are sampled systematically across the scene using a fixed spatial stride. To form correlated, *pixel-aligned* views without heuristic spectral perturbations, we apply scene-level morphological opening/closing to $\mathcal{X}$, yielding $\mathcal{X}^{(o)}$ and $\mathcal{X}^{(c)}$, and extract aligned triplets $\{x_{i,j}, x^{(o)}_{i,j}, x^{(c)}_{i,j}\}$. During training, we additionally apply lightweight geometric transforms on-the-fly to the anchor $x_{i,j}$ to obtain $x^{(g)}_{i,j}$.

4 MorphJEPA in Detail

MorphJEPA learns f_θ via JEPA-style *latent prediction* across the aligned views described above. We first instantiate f_θ with a lightweight disentangled spatial–spectral encoder fused by cross-attention (Sec. 4.1). We then train it with a multi-view student–teacher prediction objective in which a predictor maps the context embedding $f_\theta(x_{i,j})$ to target embeddings computed from $x^{(g)}_{i,j}$, $x^{(o)}_{i,j}$, and $x^{(c)}_{i,j}$ (Sec. 4.2). The morphology-based targets provide controlled *structural simplification* while preserving spectral identity, making nuisance variations harder to predict and biasing the representation toward transferable topology and material cues. Training is further stabilized using an exponential moving average (EMA) teacher together with a context-anchored regularizer (SIGReg; Sec. 4.3). For evaluation, we freeze f_θ and train a linear head on N-shot labeled patches (Sec. 4.4). Fig. 2 summarizes the overall framework, covering both self-supervised pretraining and downstream low-shot classification.

4.1 Disentangled Spatial–Spectral Encoder

HSI contains two complementary cues: local spatial neighborhoods encode topological geometry, while spectra encode material semantics. Monolithic 3D-CNN and ViT backbones often entangle these factors [5, 22] and can overfit in low-label regimes; moreover, our morphology-based views primarily perturb *spatial* structure and should not be absorbed by implicitly altering spectral identity.

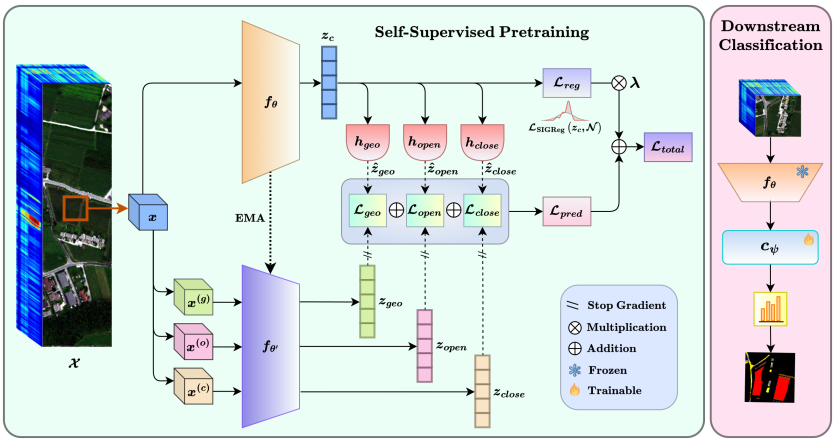

Fig. 2: Overview of the MorphJEPA architecture. *Self-Supervised Pretraining:* Given a hyperspectral patch x, we generate a geometric target ($x^{(g)}$) and morphology-filtered targets ($x^{(o)}$, $x^{(c)}$). A student encoder f_θ maps x to a context embedding z_c, while an EMA teacher encoder $f_{\theta'}$ generates stop-gradient target embeddings (z_{geo}, z_{open}, z_{close}). Task-specific predictors (h_{geo}, h_{open}, h_{close}) map z_c to these targets, enforcing a directional structural bottleneck optimized via $\mathcal{L}_{pred}$. To prevent manifold collapse, a context-anchored SIGReg ($\mathcal{L}_{reg}$) regularizes z_c toward an isotropic Gaussian prior. *Downstream Classification:* The pretrained encoder f_θ is frozen, and a linear classifier c_ψ is trained using a few labeled samples.

We therefore use a compact disentangled encoder f_θ that factorizes spatial and spectral processing and fuses them via cross-attention. This design provides (i) spectrum-preserving features under morphological views, (ii) an explicit mechanism to modulate spatial structure using spectrally informative cues, and (iii) a lightweight architecture suitable for dense HSI cubes.

Spatial Branch. The spatial pathway captures multi-scale structure using a standard and a dilated 3×3 convolution (each producing $D/2$ channels), which are concatenated and projected to D channels:

$$F_{\text{spa}} = \text{GELU}\Big(\text{BN}\Big(\Big[\text{Conv}_{\text{2D}}^{3\times3}(x) \parallel \text{Conv}_{\text{2D}}^{3\times3,d=2}(x)\Big]\Big)\Big), \quad F_{\text{spa}} \in \mathbb{R}^{S\times S\times D}. \tag{1}$$

Spectral Branch. The spectral pathway performs per-pixel spectral mixing via a 1×1 convolution, without spatial aggregation:

$$F_{\text{spec}} = \text{GELU}\big(\text{BN}\big(\text{Conv}_{\text{2D}}^{1\times1}(x)\big)\big), \quad F_{\text{spec}} \in \mathbb{R}^{S\times S\times D}. \tag{2}$$

Cross-Attention Fusion. We flatten both maps into $T = S^2$ tokens and fuse them through a multi-head cross-attention bottleneck, where spatial tokens query spectral tokens. Let $Q \in \mathbb{R}^{T\times D}$ be derived from F_{spa} (with positional embeddings) and $K, V \in \mathbb{R}^{T\times D}$ from F_{spec}. The fused tokens are:

$$H = Q + \text{softmax}\left(\frac{QK^\top}{\sqrt{D_h}}\right)V, \quad D_h = D/h, \quad H \in \mathbb{R}^{T\times D}. \tag{3}$$

A lightweight FFN followed by global average pooling produces patch embeddings. Overall, this factorization assigns topology (the component perturbed by morphology) to the spatial pathway, while the spectral pathway anchors semantics, improving robustness under spatial perturbations and scene shifts.

4.2 Multi-View Prediction Objective

Unlike prior HSI-SSL that contrasts heuristically augmented views or reconstructs masked pixels/bands, we learn via *latent prediction* over a controlled, structure-preserving view family. For each anchor patch, we use the aligned targets: a geometric view $x^{(g)}$ (random rotations/flips) and morphology-filtered views $x^{(o)}$ and $x^{(c)}$.

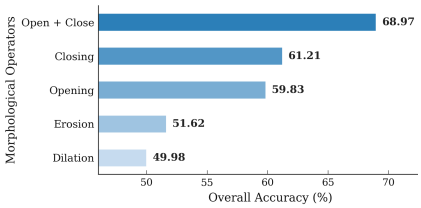

Fig. 3: Ablation of morphological operators on Houston 2013 under 5-shot regime.

We use opening/closing (rather than atomic erosion/dilation) since they are idempotent filters that attenuate small-scale spatial variations while preserving dominant structure, yielding semantically aligned targets without altering spectral identity. As shown in Fig. 3, standard erosion/dilation underperforms the compound filters. Furthermore, integrating both opening and closing into the predictive objective yields the strongest results; their complementary nature captures a complete morphological profile, corroborating our hypothesis that robust, topology-preserving operators are essential for extracting semantically aligned targets.

To model these asymmetric transitions, we adopt a teacher–student JEPA formulation with an EMA teacher. Let $z_c = f_\theta(x)$ denote the student context embedding. The teacher produces target embeddings (no gradients):

$$z_{geo} = \mathrm{sg}\Big(f_{\theta'}(x^{(g)})\Big), \quad z_{open} = \mathrm{sg}\Big(f_{\theta'}(x^{(o)})\Big), \quad z_{close} = \mathrm{sg}\Big(f_{\theta'}(x^{(c)})\Big). \tag{4}$$

We use lightweight predictors h_{geo}, h_{open}, h_{close} to map z_c to each target:

$$\hat{z}_{geo} = h_{geo}(z_c), \quad \hat{z}_{open} = h_{open}(z_c), \quad \hat{z}_{close} = h_{close}(z_c), \tag{5}$$

and minimize the summed latent prediction loss:

$$\mathcal{L}_{pred} = \mathcal{L}_{geo} + \mathcal{L}_{open} + \mathcal{L}_{close}, \quad \mathcal{L}_v = \|\hat{z}_v - z_v\|_2^2,\ v \in \{geo, open, close\}. \tag{6}$$

The morphology terms encourage the encoder to capture topology that remains stable under opening/closing. However, using only these terms can admit drift toward relative structural differences; the geometric term acts as a semantic anchor by requiring consistency with an *identity-preserving* transformation, enforcing absolute material fidelity and spatial equivariance (Sec. 5.4, Tab. 4). Overall, predicting morphologically simplified targets from a raw context induces

a *directional bottleneck*: high-frequency noise attenuated by opening/closing becomes hard to predict, whereas stable topological content remains predictable. This provides a well-posed alternative to aggressive masking in dense HSIs, which can remove semantically critical structure and degrade transfer.

4.3 Latent Manifold Regularization

Although the EMA teacher prevents trivial constant collapse, it does not constrain the manifold structure of the learned representation space. In high dimensional HSI settings, predictive training alone can yield *dimensional collapse*, where embeddings concentrate in a low-rank anisotropic subspace due to highly correlated overlapping patches, thereby destabilizing self-supervised optimization [18].

While an isotropic-target SIGReg has been proposed to prevent degenerate embeddings (LeJEPA [8]), it is applied symmetrically there, eliminating the need for momentum teachers. In hyperspectral settings, however, symmetric conditioning of all embeddings can cause manifold collapse under severe spatial–spectral correlation (Sec. 5.3, Tab. 1, Fig. 4). We therefore retain the EMA teacher and apply SIGReg only to the student context embeddings, preserving stable predictive targets while regularizing the latent manifold.

Let $Z_c = \{z_c^{(k)}\}_{k=1}^{M} \subset \mathbb{R}^D$ denote a batch of student context embeddings. We impose SIGReg by penalizing the discrepancy between the empirical batch distribution and an isotropic Gaussian prior $\mathcal{N}(0, I)$ using an empirical characteristic-function matching objective, yielding $\mathcal{L}_{reg} = \mathcal{L}_{\text{SIGReg}}(Z_c)$ (see `Supp. Mat.` for detailed mathematical formulation). By anchoring the empirical distribution of Z_c to $\mathcal{N}(0, I)$, SIGReg enforces bounded trace and near-isotropic covariance (Sec. 4.5), ensuring predictive alignment in a well-conditioned latent space. The final objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda \mathcal{L}_{reg}, \tag{7}$$

where λ controls the strength of the regularization.

4.4 Downstream Low-shot Classification

After pretraining, we freeze f_θ and train a linear classifier $c_\psi : \mathbb{R}^D \to \mathbb{R}^K$ on N-shot ($N \in \{1, 5, 10\}$) labeled patches using cross-entropy:

$$\min_{\psi} \; \mathcal{L}_{\text{sup}} = \mathbb{E}_{(x,y)} \left[-\log \frac{\exp(\,[c_\psi(f_\theta(x))]_y\,)}{\sum_{k=1}^{K} \exp(\,[c_\psi(f_\theta(x))]_k\,)} \right]. \tag{8}$$

4.5 Why MorphJEPA Transfers and Probes Better

Let $x = (s, n)$, where s denotes class-relevant structure and n nuisance variability, and let $z = f_\theta(x)$ be the learned representation. We consider intra- and

cross-scene shifts that preserve s but alter nuisance statistics. Following standard optimal transport bounds [36], a downstream predictor φ with L-Lipschitz loss satisfies:

$$\mathcal{R}_{\mathcal{P}'}(\varphi \circ f_\theta) \;\leq\; \mathcal{R}_{\mathcal{P}}(\varphi \circ f_\theta) \;+\; L\, W_1(\mathcal{P}_z, \mathcal{P}'_z), \tag{9}$$

so transfer depends on the representation drift in latent space. Since MorphJEPA predicts embeddings of morphologically filtered aligned targets that suppress nuisance-sensitive details while preserving dominant structure, it encourages invariance to n and thus reduces $W_1(\mathcal{P}_z, \mathcal{P}'_z)$.

Proposition 1. *Under mild regularity assumptions, MorphJEPA yields smaller nuisance-induced representation drift than standard JEPA objectives with nuisance preserving targets.*

For low-shot linear probing with N labeled samples and representation covariance Σ, the expected Rademacher complexity [23] scales as:

$$\widehat{\mathcal{R}}(c_\psi \circ f_\theta) \;\leq\; \mathcal{O}\left(B\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} \right). \tag{10}$$

Our context-anchored SIGReg promotes an approximately isotropic latent space, which controls $\mathrm{Tr}(\Sigma)$ and improves conditioning for low-shot linear separation.

Proposition 2. *If SIGReg enforces approximate isotropy, then the low-shot complexity term is tightened through improved covariance control.*

Detailed assumptions and proofs for both propositions are provided in the `Supp. Mat.`

5 Experiments

5.1 Datasets and Evaluation Setup

We evaluate MorphJEPA on two intra-scene (Houston 2013, Trento) [42] and two cross-scene (Houston: Houston 2013 → 2018; HyRank: Dioni → Loukia) [46] benchmarks (see `Supp. Mat.` for detailed dataset specifications).

Intra-Scene Evaluation. MorphJEPA is pretrained on the entire unlabeled scene (e.g., Houston 2013, Trento). Representation quality is then assessed by training a linear classifier on the frozen latent features of that same scene under $1, 5,$ and 10-shot regimes. This evaluates the backbone's intrinsic discriminative power without further weight updates.

Cross-Scene Generalization. We evaluate robustness for cross-sensor temporal transfer across aligned spectral overlap (Houston 2013 → 2018) and pronounced urban-to-rural structural shifts (Dioni → Loukia), where the former in each pair denotes the *source* and the latter the *target* scene. We employ two distinct protocols:

1. **Zero-Shot Target Transfer:** This assesses the transferability of topological invariants. Both the encoder (via self-supervision) and the classifier (via $1, 5,$ and 10-shot supervision) are trained exclusively on source-scene data. The model observes no target-scene samples (labeled or unlabeled) during any training stage prior to evaluation.
2. **Low-Shot Target Adaptation:** This assesses representation adaptability. The pretrained encoder remains frozen, and a linear classifier is adapted using $1, 5,$ and 10 labeled target-scene samples before target evaluation.

5.2 Experimental Protocol

Following standard HSI protocols [1, 3], we apply PCA (15 components) to all scenes, to isolate structural semantics from raw sensor noise. MorphJEPA is implemented with a patch size of 15×15 and pretrained for 100 epochs using AdamW [31] (learning rate 2×10^{-4}, weight decay 10^{-4}) with cosine annealing to 10^{-6}. The EMA momentum is set to 0.996, the regularization weight to $\lambda = 0.05$, and morphological views use a 3×3 structuring element (Sec. 5.4, Fig. 7).

For downstream evaluation, the encoder is frozen, and a linear classifier is trained for 100 epochs using AdamW (learning rate 10^{-3}, weight decay 0.01) under N-shot settings ($N \in \{1, 5, 10\}$). We use a batch size of 32 (or all available labeled samples when smaller), with linear warmup followed by linear decay. Performance is reported using overall accuracy (OA), average accuracy (AA), and the Kappa coefficient (κ) [4], averaged over 10 independent runs.

5.3 Comparison to the Literature

Quantitative Analysis. We evaluate MorphJEPA against some standard SSL baselines (SimCLR [14], MAE [19], I-JEPA [6], and LeJEPA [8]) using our proposed encoder. We also compare MorphJEPA against a diverse set of recent state-of-the-art (SOTA) HSI-specific methodologies. These encompass fully and semi-supervised models (MorpMamba [1], Res-CP [39]), self-supervised frameworks (DMVL [28], SC-EADNet [50], SSLSM [30], HSI-MAE [24], GGCL [26], SpectralDINO [13]), and a self-supervised multi-modal framework utilizing joint HSI and LiDAR data (CM-SCON [42]).

Tab. 1 reports the comparative results on the Houston 2013 and Trento datasets. MorphJEPA consistently achieves SOTA performance, outperforming MAE and SC-EADNet by +6.28% and +2.88% OA, respectively, under 5-shot Houston 2013 evaluation. Furthermore, the catastrophic collapse of LeJEPA (24.27% OA) empirically proves our morphological pretext is necessary to prevent manifold degeneration and learn discriminative features.

Tab. 2 evaluates cross-sensor temporal transfer (Houston 2013 $\rightarrow$ 2018), where MorphJEPA secures the highest OA across most zero-shot and low-shot regimes, outperforming GGCL by +1.20% OA under 5-shot target adaptation. This confirms that predicting morphological evolution successfully isolates transferable features that are robust to severe temporal and sensor shifts.

Table 1: 1, 5, and 10-shot classification performance comparison of MorphJEPA and SOTA methods on Houston 2013 and Trento datasets. Best results are highlighted in red; second-best results are in blue.

Method	Houston 2013									Trento								
	1-shot			5-shot			10-shot			1-shot			5-shot			10-shot		
	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
HSI-tailored supervised/semi-supervised methods																		
MorpMamba [1]	22.90	21.97	21.41	44.95	45.71	43.19	59.06	57.21	52.32	39.30	37.05	35.33	61.09	58.33	59.28	68.81	66.93	67.21
Res-CP [39]	43.54	45.60	39.06	65.08	74.58	62.06	72.90	77.43	70.58	64.75	57.91	56.19	78.05	74.99	76.21	88.88	83.73	86.20
Standard SSL baselines (with our spatial–spectral encoder)																		
SimCLR [14]	35.43	38.19	30.54	60.18	62.45	55.82	65.11	67.28	60.47	62.53	55.38	52.76	79.48	74.22	75.83	83.24	78.47	80.12
MAE [19]	41.24	44.47	38.15	68.45	71.23	65.78	73.38	76.12	70.95	68.14	60.48	58.22	83.17	78.44	80.13	86.52	81.28	83.19
I-JEPA [6]	29.83	32.38	24.15	48.57	51.22	43.51	53.42	55.68	48.24	52.08	48.64	45.28	68.43	64.18	65.77	72.48	69.13	70.36
LeJEPA [8]	18.52	20.08	12.33	24.27	26.54	18.38	28.14	29.77	22.48	25.38	22.84	19.53	32.13	28.58	25.37	35.58	32.15	29.77
HSI-tailored SSL frameworks																		
DMVL [28]	28.66	30.50	23.21	43.47	42.54	39.19	47.92	47.81	43.74	65.84	60.94	57.61	87.26	77.09	83.44	89.89	79.02	86.78
SC-EADNet [50]	50.73	55.72	46.18	71.85	75.51	69.28	76.04	78.32	74.06	64.36	55.77	53.47	82.39	78.73	79.91	87.81	84.07	85.02
SSLSM [30]	29.94	34.09	25.12	43.75	45.79	39.26	49.39	49.20	45.03	45.92	43.43	36.97	84.25	82.80	79.09	89.06	86.47	87.07
HSI-MAE [24]	36.86	38.22	32.74	65.13	69.11	61.82	74.69	74.47	69.67	64.62	59.39	57.79	82.56	80.91	80.16	89.55	86.09	87.35
GGCL [26]	37.83	36.70	32.32	51.21	52.71	47.09	60.67	57.48	57.18	77.92	65.65	63.09	80.91	69.86	73.44	81.51	74.35	75.28
SpectralDINO [13]	28.75	34.86	24.25	56.13	60.97	52.63	66.95	69.52	64.18	71.28	65.25	61.13	85.99	80.11	81.75	88.27	84.98	85.34
CM-SCON [42]	40.64	41.82	36.41	68.95	71.25	64.92	76.45	78.34	74.92	66.38	62.33	59.23	87.22	82.89	84.28	90.48	86.92	87.59
MorphJEPA	52.06	59.65	48.25	74.73	77.53	72.61	79.15	81.33	77.38	73.33	68.23	64.97	89.62	84.40	86.20	92.30	88.79	89.77

Table 2: 1, 5, and 10-shot cross-scene cross-sensor temporal transfer performance comparison of MorphJEPA and SOTA methods on Houston (Houston 2013 → 2018) dataset under zero-shot target transfer and low-shot target adaptation regimes.

Method	2013 → 2018 (Zero-shot Target Transfer)									2013 → 2018 (Low-shot Target Adaptation)								
	1-shot			5-shot			10-shot			1-shot			5-shot			10-shot		
	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
HSI-tailored supervised/semi-supervised methods																		
MorpMamba [1]	7.34	8.92	3.81	10.20	13.44	9.66	17.52	15.34	10.73	14.14	16.31	8.42	26.06	24.12	18.28	34.51	36.91	24.64
Res-CP [39]	29.41	34.08	15.54	47.33	39.44	30.86	61.73	40.32	43.22	34.53	22.11	14.47	58.08	41.47	39.99	61.75	45.34	46.18
HSI-tailored SSL frameworks																		
DMVL [28]	15.87	14.83	7.37	19.38	26.65	11.63	29.27	32.62	14.39	21.80	16.51	11.89	25.78	19.13	13.85	27.25	20.95	15.57
SC-EADNet [50]	30.36	23.02	13.02	63.20	23.57	21.51	63.48	24.08	23.29	32.61	35.07	21.14	63.32	61.10	47.28	71.04	66.69	56.74
SSLSM [30]	14.21	28.59	8.68	26.74	35.54	17.40	38.85	38.24	21.43	15.09	30.17	9.48	30.55	41.39	19.33	43.09	45.36	31.94
HSI-MAE [24]	19.43	19.50	10.29	32.82	25.45	16.29	37.99	41.90	28.84	25.52	26.44	14.55	37.53	37.22	26.53	50.13	52.46	37.70
GGCL [26]	22.71	32.25	12.75	42.02	34.74	22.60	42.35	43.12	29.91	22.95	34.91	13.35	66.01	28.03	36.10	67.80	37.19	42.54
SpectralDINO [13]	16.02	18.02	8.66	18.25	31.59	11.05	40.05	38.14	29.65	39.28	28.69	16.67	50.03	51.54	34.13	55.20	52.35	38.67
CM-SCON [42]	21.86	23.39	12.98	34.74	37.47	24.04	41.54	38.31	30.84	27.32	28.85	17.47	40.87	43.29	28.28	48.87	51.30	36.28
MorphJEPA	33.24	32.48	16.71	62.36	39.07	41.19	64.30	42.41	44.45	38.06	40.05	22.18	67.21	62.43	51.33	71.97	67.84	57.57

Tab. 3 details geographic generalization on HyRank (Dioni → Loukia). While competitive baselines occasionally approach our OA, MorphJEPA overwhelmingly dominates AA and κ across both regimes, proving the morphological pretext yields spatially invariant representations that transfer robustly across all semantic categories rather than overfitting to dominant target-scene samples.

Qualitative Analysis. Fig. 4 visualizes the purely unsupervised t-SNE embeddings of our proposed encoder pretrained using different JEPA objectives on Houston 2013. Generic spatial masking in I-JEPA causes severe semantic interleaving, while LeJEPA suffers from catastrophic dimensional collapse, reducing the entire representation manifold to an inseparable 1D curve ($\kappa = 1.35$). In contrast, MorphJEPA's structural bottleneck successfully filters out high-frequency spatial nuisances, enforcing a stable, well-conditioned ($\kappa = 1.14$), and class-discriminative representation space without any label guidance. Fig. 5 further

Table 3: 1, 5, and 10-shot cross-scene spatial transfer performance comparison of MorphJEPA and SOTA methods on HyRank (Dioni → Loukia) dataset under zero-shot target transfer and low-shot target adaptation regimes.

Method	Dioni → Loukia (Zero-shot Target Transfer)									Dioni → Loukia (Low-shot Target Adaptation)								
	1-shot			5-shot			10-shot			1-shot			5-shot			10-shot		
	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
	HSI-tailored supervised/semi-supervised methods																	
MorpMamba [1]	11.24	9.05	5.88	13.38	9.88	8.56	21.36	15.52	13.82	17.65	18.90	11.30	29.84	27.50	20.15	39.12	40.55	28.74
Res-CP [39]	30.05	28.92	24.03	43.14	32.94	31.72	46.36	35.62	37.13	31.06	32.17	24.72	56.76	56.87	51.09	60.19	59.57	53.50
	HSI-tailored SSL frameworks																	
DMVL [28]	26.89	26.40	21.10	32.03	29.57	24.81	37.65	33.16	28.16	30.04	33.01	23.49	35.17	35.69	29.00	41.42	41.15	34.68
SC-EADNet [50]	28.44	13.37	12.97	34.54	20.20	19.18	38.73	22.30	21.44	43.23	46.82	32.99	61.43	63.53	52.04	65.50	67.72	58.57
SSLSM [30]	19.77	20.90	12.28	24.37	32.36	19.63	33.58	32.84	26.57	32.27	39.09	25.47	48.26	58.95	40.71	52.03	62.2	44.63
HSI-MAE [24]	22.15	21.80	14.51	35.63	28.27	26.85	40.25	34.67	36.21	28.75	30.10	19.45	43.80	44.50	32.60	55.45	57.20	45.30
GGCL [26]	22.36	13.32	12.43	25.41	21.91	15.98	27.50	22.13	16.17	33.08	38.45	28.36	50.79	62.44	42.29	60.25	65.56	56.01
SpectralDINO [13]	18.80	22.07	11.57	38.63	33.16	28.42	39.70	33.37	30.58	38.34	42.14	31.01	53.30	68.01	46.75	55.46	71.32	49.41
MorphJEPA	31.49	30.09	24.18	43.02	35.54	33.82	46.68	38.00	38.36	41.30	48.87	33.70	60.37	71.22	54.11	65.88	76.06	60.09

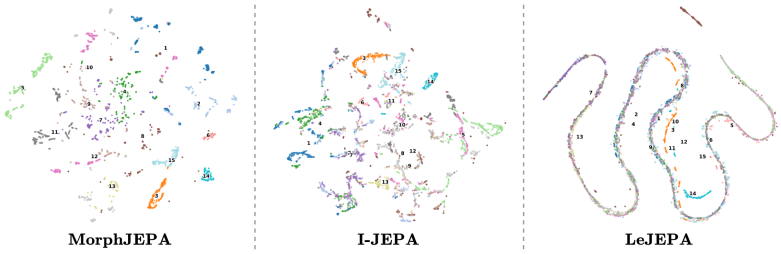

Fig. 4: t-SNE visualization of the learned representation spaces on Houston 2013 dataset. MorphJEPA produces well-separated and semantically meaningful class distributions. I-JEPA struggles with massive class overlap due to semantically destructive pretext tasks, while LeJEPA experiences severe manifold degeneration, collapsing into a continuous 1D structure.

corroborates this observation through PCA projections of the pretrained MorphJEPA features, revealing coherent spatial–spectral clustering and boundary-aware organization across both urban and geographical scenes.

Fig. 6 demonstrates that MorphJEPA successfully suppresses the erratic, localized prediction errors seen in GGCL and CM-SCON, yielding highly stable macro-structures. While our strong structural bottleneck inherently trades micro-fidelity on ultra-thin features (e.g., roads) for this regional stability, this deliberate regularization is precisely what drives our superior generalization under extreme label scarcity.

5.4 Ablation Studies

Analysis of Loss Components. The ablation in Tab. 4 reveals a clear trend across both intra-scene (Houston 2013) and cross-scene (HyRank) tasks in the 5-shot regime. Geometry-only supervision (i) performs the weakest, while morphological objectives (open and close) (ii) yield substantial gains (+15.7% OA on Houston 2013 and +18.3% on HyRank over geo-only), confirming that topology-

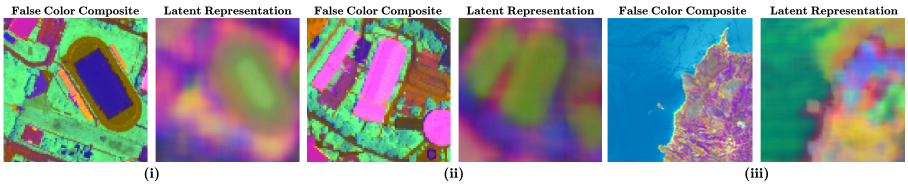

Fig. 5: PCA visualization of final-layer features from our spatial–spectral encoder pretrained with MorphJEPA on Houston (urban) and Dioni (geographical) scenes. For each image, features are independently projected to RGB using the top three principal components. Without any supervision, the representations exhibit clear semantic organization, with consistent clustering of cohesive regions sharing similar material signatures across urban structures **(i, ii)** and coastal terrain **(iii)**, reflecting emergent topology-aware and boundary-sensitive feature formation.

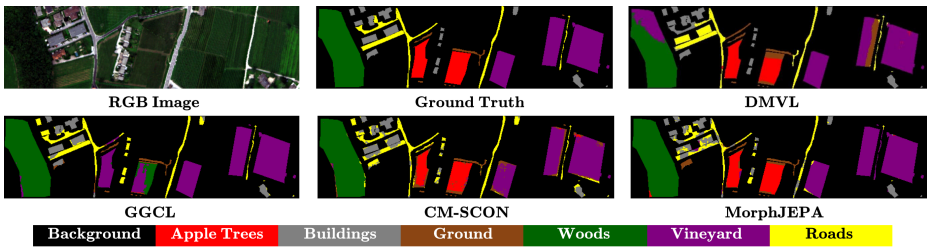

Fig. 6: Visual comparison of 5-shot classification maps on Trento. MorphJEPA produces remarkably clean, cohesive regions (e.g., Apple Trees, Vineyard) and eliminates the severe "salt-and-pepper" semantic confusion prevalent in SOTA methods.

preserving targets provide more transferable structure than appearance-based equivariances. When geometric and morphological components are jointly optimized (iii), performance increases further, and the standard deviation drops significantly (Houston 2013: 3.72 $\rightarrow$ 1.62), indicating that geometric cues act as a semantic anchor stabilizing morphological prediction. Finally, adding the context-anchored $\mathcal{L}_{reg}$ (iv) yields a modest but consistent OA gain alongside additional deviation reduction ($\approx 31\%$ on Houston 2013 and $\approx 19\%$ on HyRank), demonstrating improved manifold conditioning and reduced sensitivity to initialization in low-shot regimes.

Impact of Pretraining and Encoder Architecture. The ablation in Tab. 5 shows that our spatial–spectral encoder trained from scratch collapses under extreme low-shot conditions (A: $31.92\% \pm 13.20$ OA on Houston 2013) compared to its SSL-pretrained counterpart (F: $74.73\% \pm 1.11$), confirming the necessity of our proposed morphological pretraining. Beyond a +42.8% absolute OA gain on Houston 2013, self-supervision reduces the standard deviation by nearly 12×, indicating vastly improved optimization stability. Furthermore, MorphJEPA serves as a backbone-agnostic framework: both 3D-CNN (B) and SpectralFormer (C) achieve strong performance (64.12% and 69.50% OA on Houston 2013, respec-

Table 4: Ablation of loss components under 5-shot regime on intra-scene (Houston 2013) and cross-scene (HyRank) datasets.

ID	Loss Components				Houston 2013			Dioni → Loukia		
	$\mathcal{L}_{geo}$	$\mathcal{L}_{open}$	$\mathcal{L}_{close}$	$\mathcal{L}_{reg}$	OA	AA	κ	OA	AA	κ
(i)	✓	✗	✗	✗	53.23 ± 2.40	59.94 ± 2.30	49.54 ± 2.63	36.52 ± 4.12	44.15 ± 3.85	32.40 ± 4.05
(ii)	✗	✓	✓	✗	68.97 ± 3.72	73.79 ± 3.01	66.45 ± 4.01	54.85 ± 5.10	65.20 ± 4.80	48.90 ± 5.25
(iii)	✓	✓	✓	✗	73.97 ± 1.62	76.67 ± 0.90	71.82 ± 1.73	59.76 ± 3.90	70.50 ± 3.15	53.56 ± 3.65
(iv)	✓	✓	✓	✓	**74.73 ± 1.11**	**77.53 ± 0.97**	**72.61 ± 1.19**	**60.37 ± 3.15**	**71.22 ± 2.37**	**54.11 ± 3.37**

Table 5: Ablation of encoder architectures and spatial–spectral fusion strategies in the 5-shot regime. We evaluate performance on intra-scene (Houston 2013) and cross-scene (HyRank) benchmarks. We compare our encoder trained from scratch (A), standard HSI backbones pretrained with MorphJEPA (B–C), and our disentangled spatial–spectral encoder utilizing concatenation (D), CBAM-based fusion (E), and cross-attention fusion (F).

ID	Params (M)	Houston 2013			Dioni → Loukia		
		OA	AA	κ	OA	AA	κ
(A)	0.18	31.92 ± 13.20	35.40 ± 12.50	28.10 ± 11.80	22.15 ± 14.50	25.80 ± 13.80	12.40 ± 12.10
(B)	4.10	64.12 ± 1.85	68.40 ± 1.55	61.20 ± 2.05	48.50 ± 2.90	55.10 ± 2.40	41.80 ± 2.65
(C)	0.38	69.50 ± 1.40	72.15 ± 1.25	67.30 ± 1.45	52.40 ± 2.55	60.80 ± 2.15	46.20 ± 2.45
(D)	**0.12**	72.10 ± 1.15	74.80 ± 1.05	69.90 ± 1.20	56.90 ± 2.20	65.40 ± 1.95	49.80 ± 2.10
(E)	0.15	73.05 ± 1.19	75.62 ± 0.98	70.43 ± 1.12	58.05 ± 2.35	66.90 ± 2.53	51.20 ± 2.95
(F)	0.18	**74.73 ± 1.11**	**77.53 ± 0.97**	**72.61 ± 1.19**	**60.37 ± 3.15**	**71.22 ± 2.37**	**54.11 ± 3.37**

tively) under the same objective. Within our disentangled design, simple concatenation (D) provides a highly parameter-efficient baseline (0.12M), while CBAM fusion [43] (E) yields modest gains, suggesting that generic reweighted attention is insufficient. Cross-attention fusion (F) delivers the best performance across both datasets, validating the importance of structured spatial–spectral querying for delineating morphological boundaries.

Fig. 7 evaluates MorphJEPA across its core design parameters: data regime, $\mathcal{L}_{reg}$ strength, morphological scale, and EMA dynamics. Consistent trends and aligned performance peaks across intra- and cross-scene transfers demonstrate stable behavior, proving MorphJEPA's optimal hyperparameters generalize robustly without brittle tuning dependencies (additional results in `Supp. Mat.`). We also compare the floating-point operations (FLOPs), parameter counts, and inference time of MorphJEPA with other SOTA HSI methods in Tab. 6. Notably, MorphJEPA delivers the lowest inference time (0.98 ms) while keeping both FLOPs (0.04 G) and trainable parameters (0.18 M) exceptionally low compared to most SOTA models.

6 Takeaways

We propose MorphJEPA, the first JEPA-style framework designed specifically for HSI. By predicting morphologically simplified targets in latent space, our method enforces a structural bottleneck that filters high-frequency nuisances while preserving transferable topology. Coupled with a cross-attention spatial–

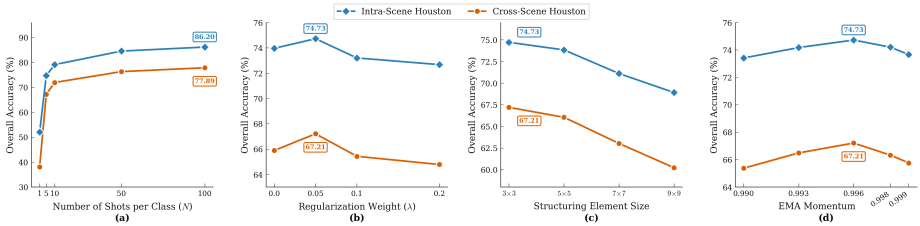

Fig. 7: Sensitivity analyses comparing intra-scene and cross-scene robustness on Houston dataset. Except for **(a)** which sweeps the number of labeled samples N, all evaluations use the 5-shot regime. **(a)** Performance scales consistently with increased labeled data. **(b)** Intermediate regularization ($\lambda = 0.05$) yields optimal conditioning. **(c)** Larger structuring elements degrade performance by over-smoothing fine structures. **(d)** EMA momentum shows a distinct optimum for teacher stability.

Table 6: Computational complexity comparison of MorphJEPA with SOTA HSI methods in terms of FLOPs, trainable parameters, and inference time. Best results are highlighted in red.

Method	FLOPs (G)	Params (M)	Inference (ms)
MorpMamba [1]	0.03	0.07	1.15
Res-CP [39]	0.63	0.28	1.48
DMVL [28]	2.12	24.62	5.87
SC-EADNet [50]	0.02	1.51	13.52
SSLSM [30]	0.39	25.65	6.02
HSI-MAE [24]	1.52	3.18	2.15
GGCL [26]	0.05	0.19	1.08
SpectralDINO [13]	0.75	21.55	12.16
CM-SCON [42]	1.85	1.45	4.35
MorphJEPA	0.04	0.18	0.98

spectral encoder and context-anchored SIGReg conditioning to prevent collapse, MorphJEPA achieves SOTA performance in intra-scene and cross-scene settings under extreme low-shot regimes. Future work will explore directional morphological filters and hybrid predictive-reconstructive objectives to better preserve ultra-fine linear structures without sacrificing macro-semantic stability of scenes.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the Anusandhan National Research Foundation (ANRF), Government of India, under Project No. **CRG/2023/004389**, and the IITB-HRTI Centre of Excellence on Computer Vision, Multimodal AI, and Geo-Intelligence.

References

1. Ahmad, M., Butt, M.H.F., Khan, A.M., Mazzara, M., Distefano, S., Usama, M., Roy, S.K., Chanussot, J., Hong, D.: Spatial–spectral morphological mamba for

hyperspectral image classification. Neurocomputing **636**, 129995 (2025)
2. Ahmad, M., Distefano, S., Khan, A.M., Mazzara, M., Li, C., Li, H., Aryal, J., Ding, Y., Vivone, G., Hong, D.: A comprehensive survey for hyperspectral image classification: The evolution from conventional to transformers and mamba models. Neurocomputing p. 130428 (2025)
3. Ahmad, M., Mazzara, M., Distefano, S., Khan, A.M., Butt, M.H.F., Hong, D.: Policymamba: Localized policy attention with state space model for land cover classification. IEEE Transactions on Neural Networks and Learning Systems (2025)
4. Ahmad, M., Shabbir, S., Roy, S.K., Hong, D., Wu, X., Yao, J., Khan, A.M., Mazzara, M., Distefano, S., Chanussot, J.: Hyperspectral image classification—traditional to deep models: A survey for future prospects. IEEE journal of selected topics in applied earth observations and remote sensing **15**, 968–999 (2021)
5. Arya, S., Dubey, S.R., Moorthi, S.M., Dhar, D., Singh, S.K.: 3d-cmt: 3d-cnn meets transformer for hyperspectral image classification. In: Proceedings of the Asian Conference on Computer Vision. pp. 679–695 (2024)
6. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023)
7. Astruc, G., Gonthier, N., Mallet, C., Landrieu, L.: Anysat: One earth observation model for many resolutions, scales, and modalities. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19530–19540 (2025)
8. Balestriero, R., LeCun, Y.: Lejepa: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544 (2025)
9. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
10. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)
11. Barman, G., Sagar, B.D.: A new approach to compute pixel vector based morphological profile for classification of hyperspectral image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025)
12. Benediktsson, J.A., Palmason, J.A., Sveinsson, J.R.: Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Transactions on Geoscience and Remote Sensing **43**(3), 480–491 (2005)
13. Chen, B., Zhang, G., Chen, T., Wang, M., Liu, J., Wang, Y., Zhang, R., Li, S., Hu, B.: Spectraldino: Dual mixture-of-subspaces low-rank adaptation for cross-domain hyperspectral image few-shot classification. IEEE Transactions on Geoscience and Remote Sensing (2025)
14. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
15. Choudhury, S., Salunkhe, Y., Mehrotra, S., Banerjee, B.: Rejepa: A novel joint-embedding predictive architecture for efficient remote sensing image retrieval. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2373–2382 (2025)
16. Dalla Mura, M., Villa, A., Benediktsson, J.A., Chanussot, J., Bruzzone, L.: Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis. IEEE Geoscience and Remote Sensing Letters **8**(3), 542–546 (2010)

17. Demertzis, K., Iliadis, L.: Geoai: A model-agnostic meta-ensemble zero-shot learning method for hyperspectral image analysis and classification. Algorithms **13**(3), 61 (2020)
18. He, J., Du, J., Ma, W.: Preventing dimensional collapse in self-supervised learning via orthogonality regularization. Advances in Neural Information Processing Systems **37**, 95579–95606 (2024)
19. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
20. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
21. Huang, L., Chen, Y., Li, Z., Ghamisi, P., Du, Q.: Hzscm: Hyperspectral image zero-shot classification via vision-language models. IEEE Transactions on Geoscience and Remote Sensing (2025)
22. Jia, S., Wang, Y., Jiang, S., He, R.: A center-masked transformer for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024). https://doi.org/10.1109/TGRS.2024.3369075
23. Kakade, S.M., Sridharan, K., Tewari, A.: On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. Advances in neural information processing systems **21** (2008)
24. Kong, W., Qi, L., Liu, B., Pei, J.: A scalable self-supervised learner for hyperspectral image classification. In: Tiny Papers at ICLR 2023 (2023), https://openreview.net/forum?id=uwbyW92Sonu
25. Landgrebe, D.: Hyperspectral image data analysis. IEEE Signal processing magazine **19**(1), 17–28 (2002)
26. Li, B., Fang, L., Chen, N., Kang, J., Yue, J.: Enhancing hyperspectral image classification: Leveraging unsupervised information with guided group contrastive learning. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–17 (2024)
27. Li, W., Yang, W., Liu, T., Hou, Y., Li, Y., Liu, Z., Liu, Y., Liu, L.: Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. ISPRS Journal of Photogrammetry and Remote Sensing **218**, 326–338 (2024)
28. Liu, B., Yu, A., Yu, X., Wang, R., Gao, K., Guo, W.: Deep multiview learning for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **59**(9), 7758–7772 (2020)
29. Liu, B., Yu, X., Yu, A., Zhang, P., Wan, G., Wang, R.: Deep few-shot learning for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **57**(4), 2290–2304 (2018)
30. Liu, W., Liu, K., Sun, W., Yang, G., Ren, K., Meng, X., Peng, J.: Self-supervised feature learning based on spectral masking for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
32. Pal, D., Bose, S., Banerjee, B., Jeppu, Y.: Morgan: Meta-learning-based few-shot open-set recognition via generative adversarial network. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6295–6304 (2023)
33. Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A.: Deep learning classifiers for hyperspectral imaging: A review. ISPRS Journal of Photogrammetry and Remote Sensing **158**, 279–317 (2019)

34. Peeters, S., Marpu, P.R., Benediktsson, J.A., Dalla Mura, M.: Classification using extended morphological attribute profiles based on different feature extraction techniques. In: 2011 IEEE International Geoscience and Remote Sensing Symposium. pp. 4453–4456 (2011). https://doi.org/10.1109/IGARSS.2011.6050221
35. Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., et al.: Recent advances in techniques for hyperspectral image processing. Remote sensing of environment **113**, S110–S122 (2009)
36. Redko, I., Habrard, A., Sebban, M.: Theoretical analysis of domain adaptation with optimal transport. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 737–753. Springer (2017)
37. Roy, S.K., Mondal, R., Paoletti, M.E., Haut, J.M., Plaza, A.: Morphological convolutional neural networks for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 8689–8702 (2021)
38. Serra, J.: Image analysis and mathematical morphology. Academic Press, Inc. (1983)
39. Seydgar, M., Rahnamayan, S., Ghamisi, P., Bidgoli, A.A.: Semisupervised hyperspectral image classification using a probabilistic pseudo-label generation framework. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–18 (2022)
40. Tao, C., Qi, J., Guo, M., Zhu, Q., Li, H.: Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–26 (2023)
41. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
42. Vasim, A., Kashyap, P., Choudhury, S., Banerjee, B.: Spatially-aware cross-modal contrastive learning for low-shot hsi classification. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
43. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
44. Yao, J., Cao, X., Hong, D., Wu, X., Meng, D., Chanussot, J., Xu, Z.: Semi-active convolutional neural networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022)
45. Zhang, F., Wang, Q., Li, X.: Hyperspectral image band selection via global optimal clustering. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp. 1–4. IEEE (2017)
46. Zhang, Y., Li, W., Zhang, M., Qu, Y., Tao, R., Qi, H.: Topological structure and semantic information transfer network for cross-scene hyperspectral image classification. IEEE Transactions on Neural Networks and Learning Systems pp. 1–14 (2021). https://doi.org/10.1109/TNNLS.2021.3109872
47. Zhao, Y., Sun, J., Hu, N., Zai, C., Han, Y.: Residual channel attention based sample adaptation few-shot learning for hyperspectral image classification. Scientific Reports **14**(1), 26746 (2024)
48. Zhen, Q., Zhang, X., Li, Z., Hou, B., Tang, X., Gao, L., Jiao, L.: Few-shot hyperspectral image classification based on domain adaptation of class balance. In: IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium. pp. 2255–2258 (2022). https://doi.org/10.1109/IGARSS46834.2022.9883302

49. Zhou, H., Zhang, X., Zhang, C., Ma, Q.: Vision transformer with contrastive learning for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters **20**, 1–5 (2023)
50. Zhu, M., Fan, J., Yang, Q., Chen, T.: Sc-eadnet: A self-supervised contrastive efficient asymmetric dilated network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–17 (2022). https://doi.org/10.1109/TGRS.2021.3131152