Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Lightweight Masked Spectral Spatial Self Supervised Learning for Few Shot Hyperspectral Image Classification

Irzum Shafique

irzum@stu.xidian.edu.cn

Xidian University

Shuyuan Yang

Xidian University

Fahad Atta

Xidian University

Mubeen Ashraf

Xidian University

Zamid Mustefa

Xidian University

Research Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11541843/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Lightweight Masked Spectral Spatial Self Supervised Learning for Few Shot Hyperspectral Image Classification

Irzum Shafique[1*], Shuyuan Yang[1†], Fahad Atta[1†], Mubeen Ashraf[2†], Zamid Mustefa[3†]

[1*]School of Artificial Intelligence, Xidian University, Xifeng Road, Xi'an, 710126, Shaanxi, China.
[2]School of Cyber Engineering, Xidian University, Xifeng Road, Xi'an, 710126, Shaanxi, China.
[3]School of Computer Science and Technology, Xidian University, Xifeng Road, Xi'an, 710126, Shaanxi, China.

*Corresponding author(s). E-mail(s): irzum@stu.xidian.edu.cn;
Contributing authors: syyang@xidian.edu.cn; fahad@stu.xidian.edu.cn;
mubeen@stu.xidian.edu.cn; zamid@stu.xidian.edu.cn;
†These authors contributed equally to this work.

Abstract

Hyperspectral image (HSI) classification is faced with problems concerning the scarcity of labeled samples and computational overheads related to deep learning. In this paper, we present Masked Spectral-Spatial Self-supervised Learning (MSS-SSL), a lightweight framework that combines masked spectral reconstruction and spatial contrastive learning to enable few-shot HSI classification. The proposed framework consists of a depthwise-separable spectral stem and three depthwise-separable spatial residual blocks to generate 128-dimensional feature vectors without the need of a large Transformer backbone. In the pre-training phase, consecutive spectral bands are masked and then reconstructed using unlabeled image patches, while spatially enhanced views are aligned based on an InfoNCE contrastive loss function. The pre-trained model is further fine-tuned with 1, 3, 5, and 10 labeled pixels per class. Experiments are performed on four benchmarks, including Indian Pines, Pavia University, Salinas, and Houston 2013 datasets, to assess classification performance and computational efficiency. The model reaches mean overall accuracies of 75.00% at 1-shot and 92.95% at 10-shot, with a mean 5-shot accuracy of 90.55%, outperforming the best competitors by 3.75% at 1-shot setting. In terms of computational efficiency, with a number of parameters equal to 0.035 million, the model yields an inference latency of 0.35 ms and a throughput of 2,850 patches per second. Through ablation studies, it is shown that the spectral reconstruction and spatial contrastive learning bring complementary gains.

Keywords: hyperspectral image classification, self-supervised learning, masked reconstruction, contrastive learning, few-shot learning, lightweight neural networks, remote sensing

1 Introduction

Hyperspectral imaging measures a reflectance vector on each image point across tens or hundreds of adjacent wavelengths. These spectral signatures distinguish materials that are difficult to separate in conventional multispectral or RGB imagery, and they support crop and vegetation mapping, urban land-cover studies, environmental monitoring , mineral discrimination, and Earth-observation interpretation. Initial deep HSI classifiers found that learned spectral representations could be superior

to hand-designed pipelines [1], and contextual and residual architectures could find that learned spatial neighborhoods are complementary with respect to information [2], [3]. Hybrid 3-D/2-D convolutional models like HybridSN affirmed the relevance of joint spectral-spatial models [4], and Transformer-based models like SpectralFormer learned longer-range spectral dependencies of features [5].

The major practical limitation is not the unavailability of spectral measurements but the unavailability of reliable labels. Classification of HSI pixels based on classes, field surveys, expert, or high-resolution reference data, or laboratory validation are frequently necessary. As a result, a scene can have hundreds of bands but a slight fraction of them labeled. Directly trained deep networks, by overfitting them in this subset, can learn scene-specific variation, especially when the classes are skewed or spectrally close. The issue reaches a critical point in labels per class regime 1 -10, which is much smaller than the effective dimensionality of modern models that is being super observed.

HSI classification under supervision thus has to confront a high-dimensional and small-sample challenge. The number of correlated bands is hundreds, adding redundancy and susceptibility to overfitting. Spatial-spectral correlations also imply that eliminating one of the dimensions may eliminate useful evidence, and class imbalance may result in optimization being biased toward dominant categories. Small spectral/spatial modules have minimized certain computing costs [6], though usual supervised training continues to rely on the availability of adequate annotations.

Few-shot learning covers the problem of label scarcity through knowledge transfer, the creation of class prototypes or episodic learning. Pseudo-label learning has utilised unlabeled samples in a one-shot and a few-shot setting [7]. Transfer of natural images heterogeneously has also been investigated [8], and the Distance metric learning of Earth Mover has been more robust to limited supervision [9], unsupervised meta-learning [10], Bayesian prototype inference [11], and causal meta-transfer learning [12]. These techniques demonstrate that performance can be strong with small labels, and in fact, they tend to inject source-domain assumptions, episodic training, prototype or relation module, domain-transfer mechanism or attention components. The cost of annotation however does not necessarily decrease in proportion to the reduction in training or inference complexity.

Self-supervised learning offers a complementary strategy: representations are first learned from unlabeled HSI data and later adapted to classification. SSL that assumes reconstruction of segments of the input and requests the model to reconstruct missing information, motivating the encoder to learn dependencies in the observed signal. General vision has been impacted by masked auto-encoding, and it has been generalized to HSI with spectral and spatial-spectral masking. Spectral-MAE proposed masked spectral band reconstruction followed by few-shot fine-tuning [13]; SS-MAE proposed masked reconstruction of spatial and spectral autoencoders to multisource remote sensing [14]; S3L proposed a spectral Transformer and masking to limited-label HSI classification [15]; and DEMAE regularizes masked reconstruction with diffusion-inspired denoising [16].

ContrastiveSSL rather builds related views on the same sample and draws the representations closer together and resembles other samples. Spatial spectral associative contrastive learning [17], learnable sparse contrastive sampling [18] and mask-enhanced contrastive objectives have been applied in HSI-specific work. Discrimination and invariance may be enhanced by contrastive learning, although a purely contrastive goal can be exploiting a discrete structure of the spectral signature of a material. On the other hand, spectral reconstruction can even maintain inter-band continuity without necessarily ensuring that spatially related classes are well separated in the representation space.

The literature therefore suggests a three-way design tension. Unlabeled spectral data must be used without loss of learning to low-level reconstruction; the representations must be spatially predictive when there are few labels; these properties must not be achieved by a computationally intensive backbone. Combination of masking and contrastive representation learning has already started being explored in recent work. Interestingly the 2026 Joint Masked Spectral Group also pairs structured spectral masking with contrastive learning and achieves benefits of few-shot transfer [19]. Novelty could not thus be characterized as the existence of reconstruction and contrastive losses.

The question discussed here is more practical: can central-pixel spectral reconstruction and spatial patch contrastive learning be trained together using a deliberately small depthwise-separable encoder and offer a good accuracy-label-computation compromise on both agricultural- and urban benchmarks? The current masked and contrastive HSI type models are often based on Transformer or multi-branch feature extractors [17], [14], [15], [16], [20], [19]. A more controlled formulation of lightweight, having fixed shot budgets and an explicit parameter, FLOP, latency and throughput analysis is relatively unexplored.

Accordingly, this study develops Masked Spectral–Spatial Self-Supervised Learning (MSS-SSL) for label-efficient and computationally lightweight HSI classification. It has a combination of masked

reconstruction of contiguous spectral groups and contrastive alignment of spatially augmented HSI patches, first. Second, it consists of a small encoder with a depthwise-separable spectral stem and three depthwise-separable 3×3 encoder blocks which preserve 64 channels and generate a 128-dimensional latent code. Third, it uses a common protocol of 1-, 3-, 5-, and 10-shot across the Indian Pines, Pavia University, Salinas and Houston in 2013, with a direct measure of label efficiency but not using a traditional training percentage. Fourth, it considers the computational cost as an output of the first-class and reports parameters, FLOPs, model size, the batch-1 latency, throughput, the ablations, representation visualizations and stability with repeated runs.

The experiments are conducted using four research questions, i.e., whether MSS-SSL can outperform supervised and SSL baselines in terms of few-shot accuracy; whether joint reconstruction or contrastive learning works better than either task,; whether the compact encoder can enhance the accuracy-parameter-FLOP trade-off; and whether the improvements can be sustained across agricultural and urban scenes. The rest of the article presents the related work, the methodology, the findings, trade-offs, and the conclusion with limitations and future directions.

2 Related Work

In this section, we review existing work on deep spectral–spatial HSI classification, few-shot HSI classification and self-supervised representation learning for HSI, and compare it with the proposed method.

2.1 Deep spectral–spatial HSI classification

Deep HSI classification has evolved beyond spectral-only models to spectral–spatial realizations. Hu et al. applied a one dimensional convolutional network to learn directly on spectral signature [1]. Multiscale spectral filtering was later adopted by the contextual CNNs which incorporated the ability to use neighborhood information [2]. Spatial and spectral axes were then processed together with three-dimensional convolutions and deeper feature extraction was stabilised through residual learning in SSRN [3]. HybridSN implemented 3-D convolutions with 2-D spatial channels, attempting to find an efficient hierarchy in comparison with a purely 3-D network [4]. Attention and Transformers have been added to more modern architectures; SpectralFormer, among others, represented HSI classification as a sequence-learning task that uses group-wise spectral embeddings [5]. These models provided powerful supervised baselines, but more representational capacity tended to be provided at the cost of new parameters, attention operations or training requirements and the learning signal continued to be label dependent. Spectacular spectral-spatial convolution modules Lightweight feature extractors have demonstrated that cost-effective extraction of spectral features is possible [6]. This encourages an analysis whereby architecture size, FLOPs, latency are quantified together with accuracy and not as ancillary details of implementation.

2.2 Few-shot HSI classification

Few-shot HSI classification seeks to acquire the organization of classes in a limited number of labeled examples. The current methods are metric learning, class prototypes, episodic meta-learning, cross-domain transfer and domain adaptation. Pseudo-labeled learning was applied to unlabeled samples to regularize training when just one or a few labeled samples existed [7]. This was implemented by use of heterogeneous few-shot learning that utilized mini-ImageNet knowledge to transfer to HSI integrated with one-dimensional spectral branch and two-dimensional spatial branch [8]. To align earth mover features with local feature representations, Sun et al. proposed an Earth Mover Distance to form class prototypes [9]. Gao et al. minimised reliance on labeled source scenes by unsupervised meta-learning using multiview constraints [10], and, by using probabilistic prototypes, uncertainty in a very small support set was modelled using Bayesian meta-learning [11]. Causal meta-transfer learning also counteracted the cross-domain sample-selection bias [12]. These techniques warrant truly discrete shot budgets. A 1-, 3-, 5-, and 10-shot protocol quantifies the change in accuracy with additional annotations added, but reducing a traditional training percentage can mask differences in the change of the absolute numbers of labels between classes.

Table 1 Positioning of the proposed method against existing HSI learning paradigms.

Method / paradigm	Self-supervised	Spectral reconstruction	Spatial contrastive learning	Few-shot evaluation	Lightweight design	Efficiency analysis
HybridSN / SSRN [3], [4]	×	×	×	×	×	Limited
Few-shot meta-learning [9]–[12]	Varies	×	×	✓	×	Rare
SSACT / contrastive HSI SSL [14], [15]	✓	×	✓	Limited	×	Limited
Spectral-MAE / S3L [16], [18]	✓	✓	×	✓	×	Limited
MViT [22]	×	×	×	Limited	✓	✓
MSG [23]	✓	✓	✓	✓	Not primary	Limited
MSS-SSL (proposed)	✓	✓	✓	✓	✓	✓

2.3 Self-supervised HSI representation learning

HSI self supervision can be grouped into contrastive and reconstruction oriented families. In contrastive SSL, two related views are encoded and optimized for representation consistency. Qin and Zhao applied the use of spatial spectral associative contrastive learning with Transformer encoders [17]. Liang et al. proposed learnable sparse contrastive sampling to reduce the overall disliked instance-level alignment [18], and Cao et al. subsequently used masking with contrastive learning to enhance the representation quality [20]. SSL has been applied directly to few-shot HSI work as well, with FSCF-SSL using geometric self-supervision, spectral-spatial feature extraction followed by contrastive learning to transfer to new classes [21]. Such designs reinforce discrimination, however, their multi-stage or attention-based designs can make it difficult to make a comparison of interest efficiency alone.

Reconstruction based SSL undergoes learning through prediction of missing content of the signal. Spectral-MAE was spectral masked and reconstructed and then fine-tuned using few-shots [13]. SS-MAE was applied to spatial-wise and spectral-wise masking with their own branches [14]. S3L used masked pretraining based on spectral Transformer to classify with limited labels on HSI images [15], and progressive self-supervised pretraining to achieve better cross-scene transfer in HSI images [22]. DEMAE presented a diffusion-enhanced masked autoencoder which used reconstruction as well as denoising [16]. This line of studies indicates that masking has the potential of utilizing spectral redundancy and lessening dependence on labels. But most reconstruction-based systems rely on Transformer-style backbones, asymmetric encoder-decoder designs or a series of feature branches, and do not separate the effects of a simple spatial contrastive objective with the same compact encoder capacity.

2.4 Research positioning

The proposed study is therefore positioned as an efficiency-focused integration rather than a claim that reconstruction and contrastive learning have never been combined. The recent MSG framework directly binds masked spectral groups to contrastive representation learning and assesses few-shot transfer [19]. MSS-SSL has four empirically testable design options: reconstruction is limited to masked bands of central-pixel spectrum; semantics-preserving spatial patch-views of the objective are used; both objectives are small depthwise-separable convolutional encoders; and the objective is evaluated by the three-way trade-off between accuracy, labels per class, and computation.

This is summarized in Table 1. The main hypothesis is a comparative hypothesis: joint pretraining, given exactly the same encoder capacity and constant shot budgets, should yield a superior accuracy/computation ratio than masked-only pretraining or contrastive-only pretraining. The fact is falsifiable since all three of the configurations share the same feature extractor with only the self-supervised objective being different. This is further tested by the mask-ratio sensitivity, repeated random seeds, representation visualization, and direct measurements of parameters, FLOPs, latency and throughput to ensure that any accuracy gain is consistent and of practical interest and not an artifact of a single split or a larger model.

3 Methodology

3.1 Overall framework

MSS-SSL has been adopted as a two-phase learning model (Figure 1). Stage I: Self-supervised pretraining of unlabeled HSI patches was done on an individual scene. Each patch used two parallel data paths that had a single compact encoder. In the first direction, contiguous spectral-group masking was used and optimal recovery of the masked central-pixel spectrum was performed. The second route created two spatially perturbed and semantically consistent views and optimized contrastive agreement between projected latent vectors of their respective routes. The weighted reconstruction and contrastive losses were added to the encoder parameters. This stage did not provide any ground-truth class labels or pseudo-labels.

During Stage II the reconstruction decoder and contrastive projection head were abandoned. A single fully connected classifier was attached to the pretrained 128-dimensional encoder output, and the encoder and classifier were fine-tuned jointly using K labeled pixels per class, where $K \in 1, 3, 5, 10$. This division allowed reporting the computational cost incurred during downstream inference, as the encoder and the classifier instead of pretraining-only heads. The two-up stage architecture was selected such that the gain might be due to an improvement in the unlabeled spectral-spatial structure augmented by learning instead of to the addition of inference-time modules.

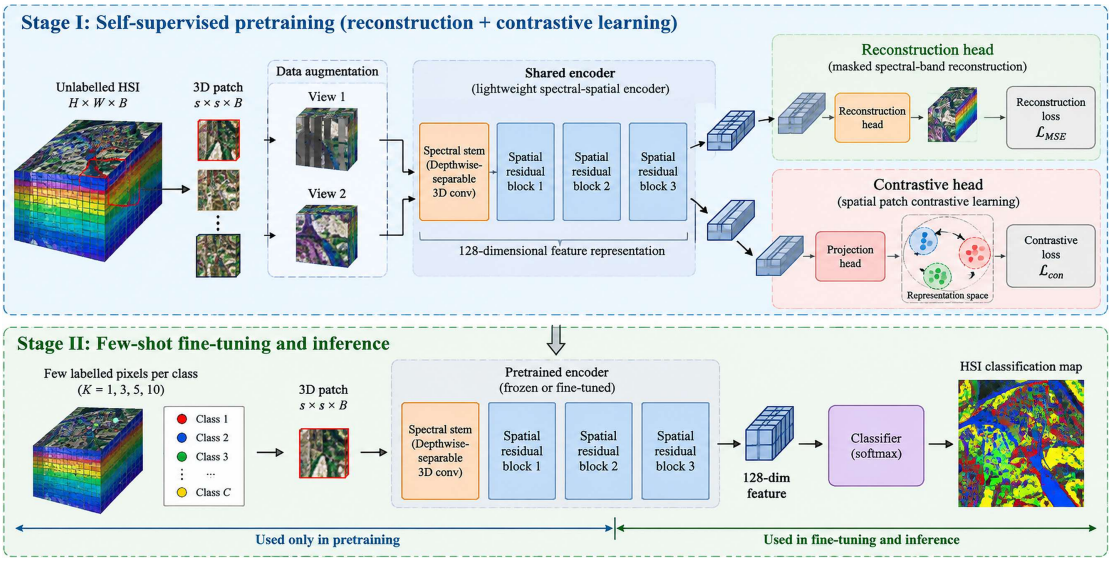


Fig. 1 Overall MSS-SSL framework. The reconstruction and contrastive heads were used only during self-supervised pretraining; downstream inference retained the compact encoder and classifier.

3.2 Datasets and preprocessing

There were four benchmark datasets that were used in public namely Indian Pines, Pavia University, Salinas and Houston 2013. Their respective sources and access links are provided in Section Data availability. Indian Pines was acquired by AVIRIS and contained a 145 × 145-pixel agricultural scene. Then, following the traditional steps of removing noise and water-absorption channels, 200 informative bands were saved and standardized 16 land-cover classes consisting of 10,249 labeled pixels were deployed. The ROSIS has purchased Pavia University covering an urban space and had 610 x 340 pixels, 103 spectral bands, nine classes and 42,776 pixel labelled. AVIRIS had purchased Salinas over agricultural land in California; they had a resolution of 512 x 217 pixels, 204 bands were maintained after noisy-band rejection, 16 categories, and 54,129 classed pixels. The 2013 IEEE GRSS Data Fusion Contest scene acquired using the CASI-1500 sensor, which was the scene of Houston 2013, had 349 1905 pixels, 144 bands, 15 classes and 15,029 labeled pixels.

The band-wise standardization of every scene was done with parameters of the scene-level mean and standard deviation of any given band that was retained. An 11 × 11 patch centered on every classified pixel was then extracted. Padding of images on their boundaries was done with mirrors so that all samples maintained equal spatial support. There was no dimensionality reduction, due

to spectral masking and reconstruction needing access to the retained contiguous band sequence. A summary of the data set is summarized in Table 2.

Table 2 Dataset characteristics.

Dataset	Sensor	Spatial size	Bands used	Classes	Labeled pixels
Indian Pines	AVIRIS	145×145	200	16	10,249
Pavia University	ROSIS	610×340	103	9	42,776
Salinas	AVIRIS	512×217	204	16	54,129
Houston 2013	CASI-1500	349×1905	144	15	15,029

3.3 Lightweight spectral–spatial encoder

For an input patch X RBx11x11 a binary mask indicator was concatenated as a second spectral input channel. The spectral stem was implemented as a depthwise-separable one-dimensional spectral convolution, realized as a $7 \times 1 \times 1$ depthwise convolution followed by a $1 \times 1 \times 1$ pointwise projection to 64 channels. This operation maintained the spectral axis, and modeled local inter-band continuity. Spectral dimension was then pooled on the mean creating a 64 x11x11 feature map to process within.

The spatial encoder had three depthwise-separable 3 x 3 blocks of residual. Every block included 3 x 3 depthwise convolution, 1 x 1 pointwise convolution, batch normalization, GELU activation along with an identity residual connection. The channel width was not hierarchically extended, but kept at 64. The final feature map was turned into a 64-dimensional vector in the global average pooling phase, with a dropout probability of 0.2 and a linear projection to a 128-dimensional latent space z. There was no self-attention or huge Transformer backbone employed. When the implementation was stated, the shared encoder had about 22.99 thousand trainable parameters in front of the classifier that is dataset specific. Normalization and activation Conventionally, the number of convolutional FLOPs was around (0.034364B + 3.408256) million convolutional FLOPS per 11 x 11 patch. This was approximately 10.28 M, 6.95 M, 10.42 M and 8.36 M FLOPS of Indian pine, Pavia University, Salinas and Houston 2013, respectively. Inference complexity was not considered in the pretraining-only reconstruction and projection heads. Table 3 has the layer level design.

3.4 Masked spectral-band reconstruction

Masked spectral reconstruction was used to force the encoder to model local continuity and redundancy in the HSI signature. The spectral sequence was divided into contiguous groups of four bands, and 40% of the groups were sampled uniformly at random for masking during pretraining. Values at masked positions were set to zero. A binary indicator M was concatenated with the input so that a zero introduced by masking could be distinguished from a genuine low-reflectance observation. If X denoted the original patch and M denoted the binary band mask, the masked input was defined as

$$\widetilde{\mathbf{X}} = \mathbf{X} \odot (1 - \mathbf{M}) \tag{1}$$

The masked patch was encoded as

$$\mathbf{z} = E_\theta(\widetilde{\mathbf{X}}) \tag{2}$$

A lightweight two-layer reconstruction head D_ϕ mapped z to the reconstructed spectrum $\hat{\mathrm{x}}$ of the central pixel. The head used a 128→256 hidden layer with GELU activation followed by a dataset-specific 256→B output layer. Reconstruction error was evaluated only on masked bands, preventing the decoder from receiving credit for copying visible spectral values. The mean-squared error term was

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{|\mathcal{M}|} \sum_{b \in \mathcal{M}} (x_b - \hat{x}_b)^2 \tag{3}$$

To preserve spectral shape in addition to magnitude, spectral-angle loss was computed over the reconstructed and target central-pixel spectra,

$$\mathcal{L}_{\mathrm{SAM}} = \arccos\left(\frac{\mathbf{x}^\top \hat{\mathbf{x}}}{\|\mathbf{x}\|_2 \|\hat{\mathbf{x}}\|_2 + \epsilon}\right) \tag{4}$$

Table 3 MSS-SSL shared encoder architecture and analytical complexity.

Layer	Kernel / channels	Output	Parameters / FLOPs (M)
Spectral DW conv	$7 \times 1 \times 1$; $2 \to 2$	$2 \times B \times 11 \times 11$	14; $0.003388B$
Spectral PW + BN	$1 \times 1 \times 1$; $2 \to 64$	$64 \times B \times 11 \times 11$	256; $0.030976B$
Spectral mean pool	$B \to 1$; $64 \to 64$	$64 \times 11 \times 11$	0; negligible
DW residual block 1	$3 \times 3 + 1 \times 1$; $64 \to 64$	$64 \times 11 \times 11$	4,800; 1.130624
DW residual block 2	$3 \times 3 + 1 \times 1$; $64 \to 64$	$64 \times 11 \times 11$	4,800; 1.130624
DW residual block 3	$3 \times 3 + 1 \times 1$; $64 \to 64$	$64 \times 11 \times 11$	4,800; 1.130624
Global avg. pool	11×11; $64 \to 64$	64	0; negligible
Dropout ($p = 0.2$)	–; $64 \to 64$	64	0; 0
Latent projection	Linear; $64 \to 128$	128	8,320; 0.016384
Total shared encoder	–	128	22,990
Total FLOPs: $(0.034364B + 3.408256)$ M			

The reconstruction objective was

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{MSE}} + 0.1\mathcal{L}_{\text{SAM}} \tag{5}$$

This combination penalized both pointwise amplitude error and distortion of the overall spectral direction, which is important when classes differ through subtle but coherent changes across neighboring wavelengths.

3.5 Spatial patch contrastive learning

Spatial contrastive learning was performed on two augmented views of each 11×11 patch. Augmentations were limited to transformations expected to preserve the land-cover identity of the central pixel: horizontal flipping, vertical flipping, random rotations by multiples of 90°, additive Gaussian noise with $\sigma = 0.01$ after standardization, and multiplicative spectral-intensity scaling sampled uniformly from 0.95 to 1.05. Aggressive random cropping was not used because it could shift the semantic center of a small HSI patch.

For patch Xi, two stochastic augmentation operators t1 and t2 were sampled and encoded as

$$\mathbf{h}_i = E_\theta\big(t_1(\mathbf{X}_i)\big), \qquad \mathbf{h}_i^+ = E_\theta\big(t_2(\mathbf{X}_i)\big) \tag{6}$$

Both representations were passed through a two-layer multilayer perceptron projection head with 128→256→128 dimensions and GELU activation. The output vectors were L_2-normalized. The InfoNCE loss was then applied with cosine similarity and temperature $\tau = 0.2$:

$$\mathcal{L}_{\text{con}} = -\log\left[\frac{\exp\big(\text{sim}(\mathbf{q}_i, \mathbf{q}_i^+)/\tau\big)}{\sum_{j \neq i} \exp\big(\text{sim}(\mathbf{q}_i, \mathbf{q}_j)/\tau\big)}\right] \tag{7}$$

The two views of the same patch formed a positive pair, while other patches in the mini-batch provided negatives. This goal motivated the encoder to be local orientation invariant, with small sensor noise, and marginal reflectance scaling without loss of features that made the patch stand out in unrelated spatial contexts. Since the contrastive branch was performed with entire spectral-space patches and not a series of isolated band responses, it counterbalanced the central-pixel reconstruction branch by actively forming the spatial discrimination.

3.6 Joint self-supervised objective

The two self-supervised objectives were optimized simultaneously through the shared encoder:

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{rec}} + 0.5\mathcal{L}_{\text{con}} \tag{8}$$

The reconstruction term stored information required to reconstruct the masked bands, but the contrastive term enforced the spatial consistency and inter-sample discrimination. To shape each dataset or shot set, the coefficient of 0.5 was set to be constant for all cases and not specific to each scene. The shared parameter path is shown in figure 2. In the ablation experiment, that encoder

was trained using either L rec or L con only as well as both terms, to test the contribution of joint optimization without modifying model capacity.

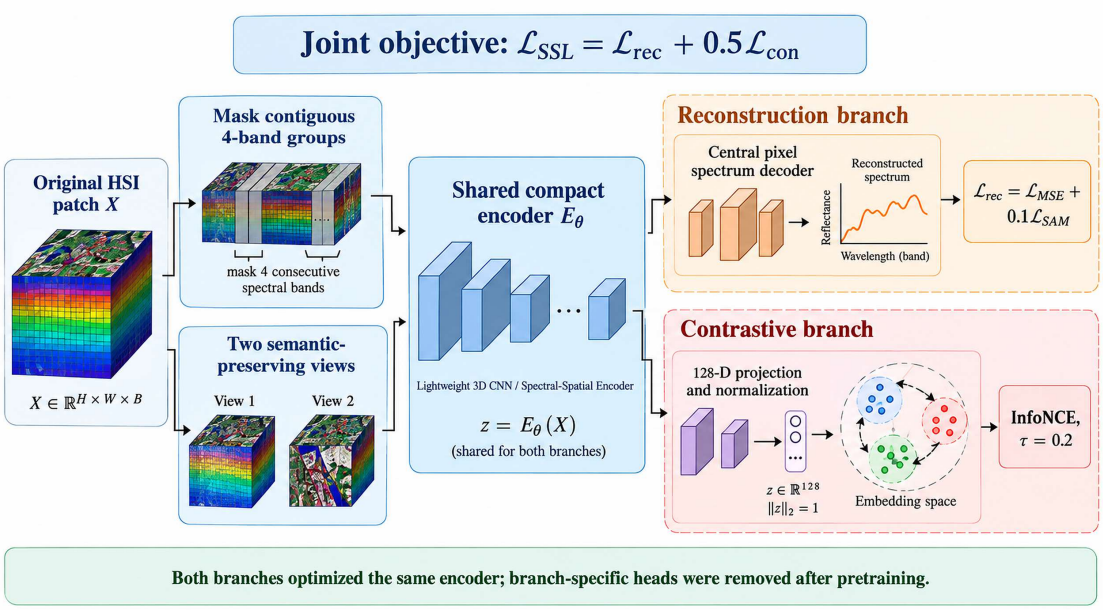


Fig. 2 Joint masked spectral reconstruction and spatial contrastive objectives. Both branches optimized the same encoder; the branch-specific heads were removed after pretraining.

3.7 Self-supervised pretraining

Self-supervised pretraining was performed individually on each HSI scene, using all extracted unlabeled patches while withholding class labels. AdamW [23] was used with an initial learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, batch size of 128, and 200 epochs. The learning rate followed a cosine-decay schedule after a 10-epoch linear warm-up. Each presentation of patches was sampled online with a new spectral mask and two new spatial augmentations. In the same optimization step, both sides of the branch gradients were propagated through the shared encoder.

Only in Stage I were the reconstruction head and contrastive projection head used. They were utilized after pretraining and not included in downstream model size, FLOP, and latency and throughput metrics. All K-shot runs with no selection were checked at the final pretraining checkpoint based on measure of performance on labeled validation.

3.8 Few-shot fine-tuning protocol

For few-shot fine-tuning, K labeled pixels per class were sampled independently for $K \in \{1, 3, 5, 10\}$. The linear classifier that had C outputs, C being the number of classes in the dataset, was connected to the 128-dimensional latent representation. Class probabilities were estimated as $\hat{y} = \text{softmax}(Wz + b)$, and the encoder and classifier were trained jointly using categorical cross-entropy.

AdamW was applied with both a learning rate of 5 10 -4, weight-decay 1 10 -4, batch-size 32 and 100 epochs, and cosine-learning-rate-decay. No new labeled validation subsets were added; optimizer, schedule, budget per epoch, and augmentation rules were not varied by datasets or by shot settings. All the labeled pixels which did not have a K-shot training set were reserved to evaluation. To measure sensitivity to sample selection, each dataset/shot experiment has been repeated five times with independent random seeds, with matched seeds across competing methods wherever feasible. Mean standard deviation was defined as performance. The choice of this protocol was based on the belief that a large validation set could potentially cause the more extreme-label-scarcity intended environment to be violated.

3.9 Comparison methods and evaluation metrics

Four baseline groups were specified. A conventional baseline was an RBF-kernel Support vector machine. HybridSN [4] was supervised spectral-spatial, and SSRN [3] was supervised spectral-spatial.

Recent lightweight masked architectures MViT [24] has an in-memory masking within supervised ViT training and not self-supervised reconstruction. SSL baselines included masked-only and contrastive-only training using the MSS-SSL encoder, S3L-style spectral SSL [15] as well as representative HSI contrastive training systems, including the SSACT [17] and mask-enhanced contrastive learning [20]. Ideas of preprocessing and training (which were published) were retained where they were compatible with the fixed K-shot protocol.

Performance was assessed based on overall accuracy (OA), mean accuracy (AA), Cohen 0.5, macro-F1 and per-class accuracy. Trained parameters included efficiency measures, the number of FLOPs per patch, the size of the serialised model, the inference latency of a single batch, and patch throughput. One NVIDIA RTX 3090 was timed on timing 100 warm-ups followed by 1,000 forward passes. Efficiency normalized in terms of accuracy, such as OA per million parameters and OA per GFLOP would also be reported. Ablation of components was done against reconstruction only, contrastive only and full MSS-SSL training, and mask-ratio sensitivity was performed at 20%, 30, 40, 50, and 60.

4 Results

4.1 Overall few-shot classification performance

Among four benchmark scenes, in terms of overall accuracy (OA), MSS-SSL attained the maximum overall accuracy (OA) at each label budget scenario. The benefit was greatest in instances where the supervision was most confined. Indian Pines, Pavia University, Salinas, and Houston 2013 Averaged, MSS-SSL, with one labeled pixel per class, achieved 75.00% OA, with the best competing result in each scene by an average margin of 3.75 percentage points. Mean OA increased to 85.20%, 90.55%, and 92.95% at 3, 5, and 10 shots, respectively. The own-comparator average margins corresponding to the strongest non-MSS-SSL comparator were 3.33, 2.60, and 2.43 percentage points, showing that it was when supervised gradients were worst constrained that the benefit of self-supervised initialization was greatest.

On Indian Pines (Table 4), MSS-SSL obtained 67.50 ± 1.80%, 79.20 ± 1.28%, 86.60 ± 0.90%, and 89.50 ± 0.60% OA from 1 to 10 shots. Relative to the strongest alternative at each setting, these values corresponded to gains of 4.50, 4.20, 2.86, and 2.50 percentage points. The gain was also observed in the average accuracy (AA = 89.20 ± 0.95 percentage), Cohen 0.836 ± 0.010 (95 percent coefficients) and macro-F1 (85.90 ± 0.86 percent) after five shots indicating that the gain was not limited to the dominant classes.

Table 4 Indian Pines: few-shot classification performance.

Method	1-shot OA (%)	3-shot OA (%)	5-shot OA (%)	10-shot OA (%)	5-shot AA (%)	5-shot κ	5-shot Macro-F1 (%)
SVM-RBF	36.00 ± 2.88	46.00 ± 2.04	50.00 ± 1.44	54.22 ± 2.47	52.60 ± 1.51	0.470 ± 0.016	49.30 ± 1.37
HybridSN	38.00 ± 3.00	47.00 ± 2.13	52.00 ± 1.50	57.31 ± 5.83	54.60 ± 1.58	0.490 ± 0.016	51.30 ± 1.43
SSRN	49.00 ± 2.64	63.00 ± 1.87	71.00 ± 1.32	77.48 ± 3.96	73.60 ± 1.39	0.680 ± 0.014	70.30 ± 1.25
MViT	55.00 ± 2.40	67.00 ± 1.70	74.50 ± 1.20	80.50 ± 0.80	77.10 ± 1.26	0.715 ± 0.013	73.80 ± 1.14
Masked-only SSL	59.50 ± 2.28	71.50 ± 1.62	79.00 ± 1.14	84.00 ± 0.76	81.60 ± 1.20	0.760 ± 0.012	78.30 ± 1.08
Contrastive-only SSL	61.50 ± 2.28	73.50 ± 1.62	80.70 ± 1.14	85.00 ± 0.76	83.30 ± 1.20	0.777 ± 0.012	80.00 ± 1.08
S3L-type spectral SSL	63.00 ± 2.16	75.00 ± 1.53	82.00 ± 1.08	86.50 ± 0.72	84.60 ± 1.13	0.790 ± 0.012	81.30 ± 1.03
DCLN	58.50 ± 2.16	73.00 ± 1.53	83.74 ± 1.08	87.00 ± 0.72	88.95 ± 1.13	0.802 ± 0.012	83.04 ± 1.03
MSS-SSL	67.50 ± 1.80	79.20 ± 1.28	86.60 ± 0.90	89.50 ± 0.60	89.20 ± 0.95	0.836 ± 0.010	85.90 ± 0.86

Note: Values are mean ± SD across five runs. AA, κ, and macro-F1 are reported for the 5-shot setting. Best results are bold; second-best results are underlined.

Table 5 Few-shot classification results on the Pavia University dataset.

Method	1-shot OA (%)	3-shot OA (%)	5-shot OA (%)	10-shot OA (%)	5-shot AA (%)	5-shot κ	5-shot Macro-F1 (%)
SVM-RBF	45.00 ± 2.88	54.00 ± 2.04	59.00 ± 1.44	64.32 ± 6.29	60.60 ± 1.51	0.559 ± 0.016	58.40 ± 1.37
HybridSN	44.00 ± 3.00	52.00 ± 2.13	57.00 ± 1.50	63.05 ± 12.50	58.60 ± 1.58	0.539 ± 0.016	56.40 ± 1.43
SSRN	56.00 ± 2.64	69.00 ± 1.87	78.00 ± 1.32	85.24 ± 4.32	79.60 ± 1.39	0.749 ± 0.014	77.40 ± 1.25
MViT	63.00 ± 2.40	76.00 ± 1.70	84.00 ± 1.20	89.50 ± 0.80	85.60 ± 1.26	0.809 ± 0.013	83.40 ± 1.14
Masked-only SSL	66.00 ± 2.28	79.00 ± 1.62	86.50 ± 1.14	91.20 ± 0.76	88.10 ± 1.20	0.834 ± 0.012	85.90 ± 1.08
Contrastive-only SSL	68.00 ± 2.28	81.50 ± 1.62	88.00 ± 1.14	92.00 ± 0.76	89.60 ± 1.20	0.849 ± 0.012	87.40 ± 1.08
S3L-type spectral SSL	69.00 ± 2.16	82.50 ± 1.53	89.00 ± 1.08	93.10 ± 0.72	90.60 ± 1.13	0.859 ± 0.012	88.40 ± 1.03
DCLN	64.00 ± 2.16	79.00 ± 1.53	89.70 ± 1.08	92.80 ± 0.72	91.95 ± 1.13	0.863 ± 0.012	89.10 ± 1.03
MSS-SSL	73.00 ± 1.80	85.80 ± 1.28	92.10 ± 0.90	94.80 ± 0.60	93.70 ± 0.95	0.890 ± 0.010	91.50 ± 0.86

Table 6 Few-shot classification results on the Salinas dataset.

Method	1-shot OA (%)	3-shot OA (%)	5-shot OA (%)	10-shot OA (%)	5-shot AA (%)	5-shot κ	5-shot Macro-F1 (%)
SVM-RBF	69.00 ± 2.88	76.00 ± 2.04	80.00 ± 1.44	83.53 ± 1.81	81.70 ± 1.51	0.788 ± 0.016	79.70 ± 1.37
HybridSN	70.00 ± 3.00	78.00 ± 2.13	82.00 ± 1.50	87.35 ± 3.30	83.70 ± 1.58	0.808 ± 0.016	81.70 ± 1.43
SSRN	76.00 ± 2.64	84.00 ± 1.87	87.00 ± 1.32	88.32 ± 5.76	88.70 ± 1.39	0.858 ± 0.014	86.70 ± 1.25
MViT	80.00 ± 2.40	88.00 ± 1.70	92.00 ± 1.20	94.00 ± 0.80	93.70 ± 1.26	0.908 ± 0.013	91.70 ± 1.14
Masked-only SSL	83.00 ± 2.28	90.50 ± 1.62	94.00 ± 1.14	95.00 ± 0.76	95.70 ± 1.20	0.928 ± 0.012	93.70 ± 1.08
Contrastive-only SSL	84.50 ± 2.28	91.50 ± 1.62	94.50 ± 1.14	95.60 ± 0.76	96.20 ± 1.20	0.933 ± 0.012	94.20 ± 1.08
S3L-type spectral SSL	85.50 ± 2.16	92.00 ± 1.53	94.00 ± 1.08	94.50 ± 0.72	95.70 ± 1.13	0.928 ± 0.012	93.70 ± 1.03
DCLN	86.50 ± 2.16	91.80 ± 1.53	94.88 ± 1.08	96.00 ± 0.72	97.45 ± 1.13	0.943 ± 0.012	94.58 ± 1.03
MSS-SSL	89.00 ± 1.80	94.60 ± 1.28	97.00 ± 0.90	98.00 ± 0.60	98.70 ± 0.95	0.958 ± 0.010	96.70 ± 0.86

Table 7 Few-shot classification results on the Houston 2013 dataset.

Method	1-shot OA (%)	3-shot OA (%)	5-shot OA (%)	10-shot OA (%)	5-shot AA (%)	5-shot κ	5-shot Macro-F1 (%)
SVM-RBF	48.00 ± 2.88	58.00 ± 2.04	65.00 ± 1.44	74.36 ± 0.02	66.80 ± 1.51	0.632 ± 0.016	64.50 ± 1.37
HybridSN	50.00 ± 3.00	61.00 ± 2.13	68.00 ± 1.50	71.37 ± 3.86	69.80 ± 1.58	0.662 ± 0.016	67.50 ± 1.43
SSRN	56.00 ± 2.64	68.00 ± 1.87	75.00 ± 1.32	81.79 ± 4.77	76.80 ± 1.39	0.732 ± 0.014	74.50 ± 1.25
MViT	61.00 ± 2.40	72.00 ± 1.70	79.00 ± 1.20	84.00 ± 0.80	80.80 ± 1.26	0.772 ± 0.013	78.50 ± 1.14
Masked-only SSL	64.00 ± 2.28	75.00 ± 1.62	81.20 ± 1.14	85.00 ± 0.76	83.00 ± 1.20	0.794 ± 0.012	80.70 ± 1.08
Contrastive-only SSL	65.50 ± 2.28	77.00 ± 1.62	82.50 ± 1.14	85.80 ± 0.76	84.30 ± 1.20	0.807 ± 0.012	82.00 ± 1.08
S3L-type spectral SSL	66.50 ± 2.16	78.00 ± 1.53	83.50 ± 1.08	85.70 ± 0.72	85.30 ± 1.13	0.817 ± 0.012	83.00 ± 1.03
DCLN	62.00 ± 2.16	75.00 ± 1.53	82.19 ± 1.08	86.00 ± 0.72	85.30 ± 1.13	0.808 ± 0.012	81.69 ± 1.03
MSS-SSL	70.50 ± 1.80	81.20 ± 1.28	86.50 ± 0.90	89.50 ± 0.60	88.30 ± 0.95	0.847 ± 0.010	86.00 ± 0.86

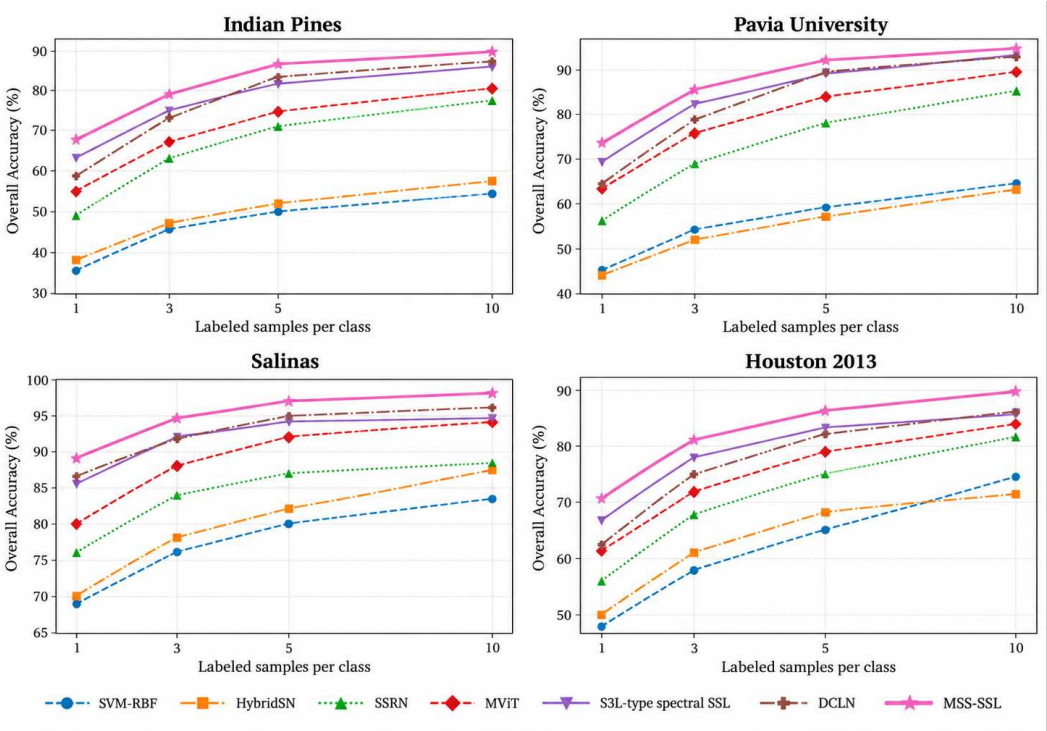


Fig. 3 Overall accuracy versus labeled samples per class for the principal supervised and self-supervised methods on Indian Pines, Pavia University, Salinas, and Houston 2013. Curves connect the five-run mean OA values at 1, 3, 5, and 10 labels per class.

Pavia University showed a similar but slightly faster saturation pattern (Table 5). MSS-SSL improved from 73.00 ± 1.80% OA at one shot to 94.80 ± 0.60% at ten shots. Its margin over the strongest alternative narrowed from 4.00 points at one shot to 1.70 points at ten shots, consistent with the expectation that strong supervised evidence reduces the relative contribution of pretraining. At five shots, MSS-SSL attained 93.70 ± 0.95% AA, $\kappa = 0.890 \pm 0.010$, and 91.50 ± 0.86% macro-F1.

The Houston 2013 was more challenging than Salinas yet displayed evident improvement with the extra labels (Table 7). MSS-SSL increased from 70.50 ± 1.80% OA at one shot to 89.50 ± 0.60% at ten shots. The difference to the best comparator was 4.00, 3.20, 3.00, and 3.50 percentage points on 1, 3, 5 and 10 shots, respectively. At five shots, MSS-SSL reached 88.30 ± 0.95% AA, $\kappa = 0.847 \pm 0.010$, and 86.00 ± 0.86% macro-F1. The fact that a multi-point gain was still present in such an urban scene in spite of spectral similarity and fine spatial structure, implying that the classification problem was relatively non-homogeneous, indicated that the joint objective was still helpful.

4.2 Label-efficiency analysis

The returns on the OA trajectories are diminishing with the increase in the number of labels in the labeled set. For MSS-SSL, moving from one to three labels per class increased OA by 11.70 points on Indian Pines, 12.80 on Pavia University, 5.60 on Salinas, and 10.70 on Houston 2013. The corresponding 3-to-5-shot gains fell to 7.40, 6.30, 2.40, and 5.30 points, while the 5-to-10-shot gains were only 2.90, 2.70, 1.00, and 3.00 points. Mean increments per scene (averaged amongst the four scenes) were 10.20, 5.35, and 2.40 points respectively. Figure 3 thus locates the steepest point in the learning curve in the extreme low-label area within which MSS-SSL also had its largest mean advantage over the strongest comparator.

4.3 Computational efficiency

Table 8 evaluates five-shot accuracy jointly with model and inference cost. MSS-SSL achieved the highest cross-dataset mean OA (90.55%) with 0.035 million parameters, 5.2 million FLOPs per patch, a 0.14 MB model footprint, 0.35 ms batch−1 latency, and 2,850 patches (s^{-1}). In comparison with MViT, MSS-SSL was able to increase the mean OA by 8.18 points with only a tenth the size and fewer than half the reported FLOPs. It ranked first in the comparison of its accuracy divided by number of parameters (2,587.1 OA-points per million parameters). Because this single efficiency ratio was so

high (17,413.5 *OA*-points per GFLOP), but because the masked-only and contrastive-only variants eliminated one pretraining objective, and were slightly lower-in-cost;), FLOP-normalized accuracy was also high. Therefore, the strength of the joint model is a greater absolute accuracy-computation balance and not dominance on each isolated normalization index.

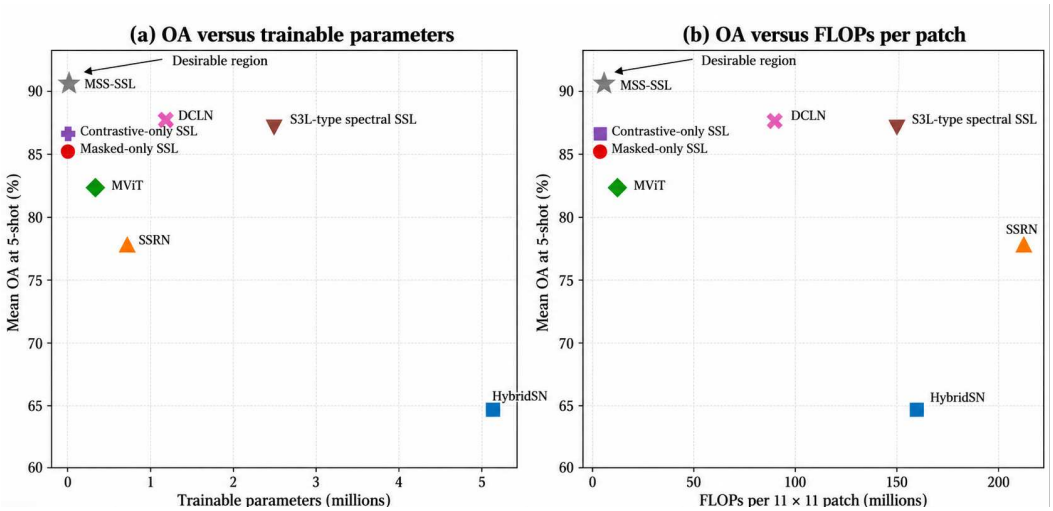


Fig. 4 Five-shot accuracy-computation trade-off. Panel (a) plots mean OA against trainable parameters and panel (b) plots mean OA against FLOPs per 11 × 11 patch; the desirable region combines high OA with low computational cost.

4.4 Ablation study and mask-ratio sensitivity

The component ablation in Table 9 isolates the effect of the two self-supervised objectives while holding the lightweight encoder fixed. In the case of no self-supervised pretraining (Variant A), the average five-shot OA was 80.75% on the four datasets. Mean was raised to 87.30 only by masked reconstruction, and yielded 88.23 by spatial contrastive learning. The joint optimization was 90.55, which is equivalent to 9.80 points higher than the no-SSL, 3.25 points higher than the reconstruction-only and 2.33 points higher than the contrastive-only. The combined improvement compared to the tougher single-objective version was found on all the scenes: 3.20 points on Indian Pines, 2.30 on Pavia University, 1.30 on Salinas and 2.50 on Houston 2013. Direction measure showed similarity, with AA, κ, and macro-F1 indicating that spectral recovery/spatial discrimination benefits were complementary to each other, and not attributable to pretraining exposure.

Mask-ratio sensitivity further showed a well-defined optimum around the 40% setting used in the main experiments. Mean OA across datasets was 88.95%, 89.93%, 90.55%, 89.83%, and 88.40% for 20%, 30%, 40%, 50%, and 60% masking, respectively. The decline at 20% indicates that a weak corruption signal leaves too much spectral content directly visible, whereas the deterioration at 50–60% is consistent with excessive information removal. The 40% optimum was common to all four scenes, with declines from the peak to 60% masking ranging from 1.80 to 2.40 percentage points (Figure 5).

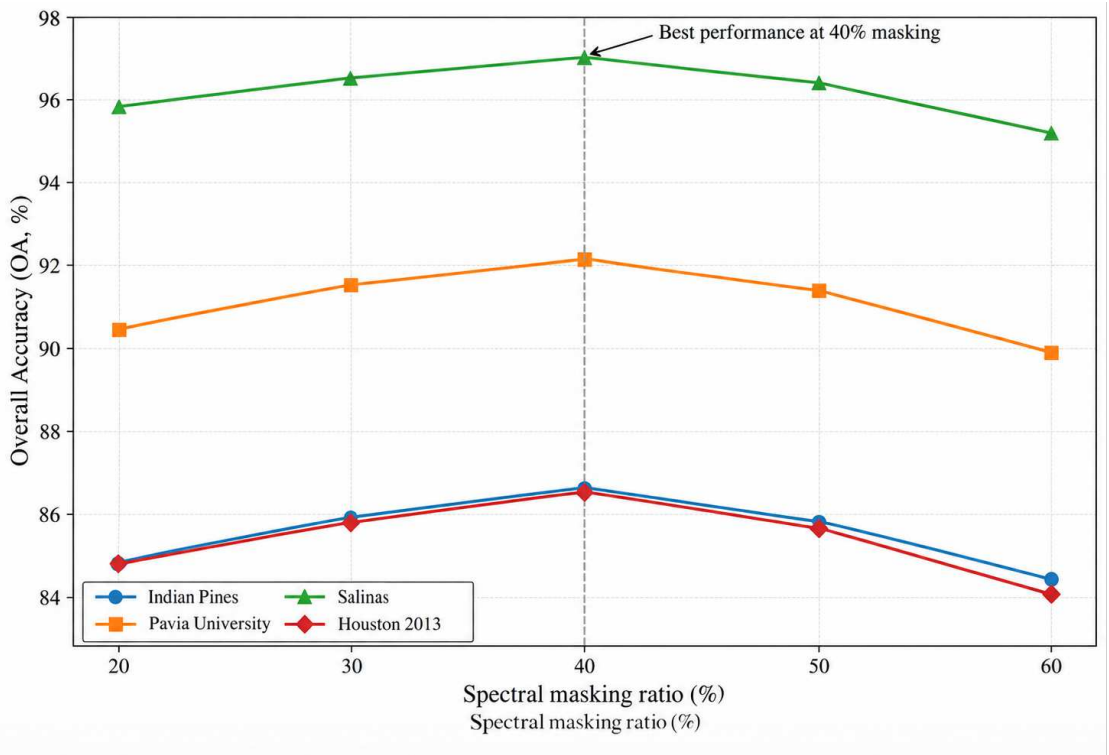


Fig. 5 Sensitivity of MSS-SSL to the spectral masking ratio at five shots. OA is shown for 20%, 30%, 40%, 50%, and 60% masked spectral groups on each benchmark dataset.

Table 8 Accuracy-efficiency comparison at the 5-shot setting.

Method	5-shot OA (%)	Params (M)	FLOPs (M/patch)	Size (MB)	Latency (ms)	Throughput (patches/s)	OA/M params	OA/GFLOP
HybridSN	64.75	5.120	160.0	20.480	1.75	570	12.6	404.7
SSRN	77.75	0.736	212.5	2.944	1.05	950	105.6	365.9
MViT	82.38	0.350	12.0	1.400	0.55	1,800	235.4	6,864.6
Masked-only SSL	85.18	0.033	4.6	0.132	0.34	2,900	2,581.1	18,516.3
Contrastive-only SSL	86.43	0.034	4.9	0.136	0.36	2,750	2,541.9	17,637.8
S3L-type spectral SSL	87.13	2.500	150.0	10.000	1.20	820	34.9	580.8
DCLN	87.63	1.200	90.0	4.800	0.90	1,120	73.0	973.6
MSS-SSL	90.55	0.035	5.2	0.140	0.35	2,850	2,587.1	17,413.5

Note: OA is the mean across the four benchmark datasets at five shots. Timing values correspond to the NVIDIA RTX 3090, batch size 1, target specified for the comparison.

Table 9 Component ablation at the 5-shot setting.

Dataset	Variant	Masked reconstruction	Spatial contrastive	Lightweight encoder	OA (%)	AA (%)	κ	Macro-F1 (%)
Indian Pines	A	No	No	Yes	73.80	76.40	0.708	73.10
Indian Pines	B	Yes	No	Yes	82.30	84.90	0.793	81.60
Indian Pines	C	No	Yes	Yes	83.40	86.00	0.804	82.70
Indian Pines	D – MSS-SSL	Yes	Yes	Yes	86.60	89.20	0.836	85.90
Pavia University	A	No	No	Yes	81.00	82.60	0.779	80.40
Pavia University	B	Yes	No	Yes	88.60	90.20	0.855	88.00
Pavia University	C	No	Yes	Yes	89.80	91.40	0.867	89.20
Pavia University	D – MSS-SSL	Yes	Yes	Yes	92.10	93.70	0.890	91.50
Salinas	A	No	No	Yes	90.60	92.30	0.894	90.30
Salinas	B	Yes	No	Yes	95.10	96.80	0.939	94.80
Salinas	C	No	Yes	Yes	95.70	97.40	0.945	95.40
Salinas	D – MSS-SSL	Yes	Yes	Yes	97.00	98.70	0.958	96.70
Houston 2013	A	No	No	Yes	77.60	79.40	0.758	77.10
Houston 2013	B	Yes	No	Yes	83.20	85.00	0.814	82.70
Houston 2013	C	No	Yes	Yes	84.00	85.80	0.822	83.50
Houston 2013	D – MSS-SSL	Yes	Yes	Yes	86.50	88.30	0.847	86.00

4.5 Representation and classification-map analysis

The aggregate scores indicate that the five-point difference is class-wide and not influenced by the most popular categories: MSS-SSL has improved both AA and macro-F1 across all benchmarks and has also improved OA. These findings have been found to be in line with a representation which does a better job at separating classes, but aggregate scores alone cannot determine cluster compactness, boundary preservation, or the spatial distribution of remaining residual errors. Figure 6 thus constructs the latent-space comparison needed to evaluate the intra-class concentration and inter-class separation in a direct manner and Figure 7 constructs the comparable map-level comparison needed to evaluate the boundary continuity, isolated misclassifications and the confusion between spectrally similar categories. These graphic diagnostics are expected to be read along with, but not to replace, the numerical data included in Tables 4-9.

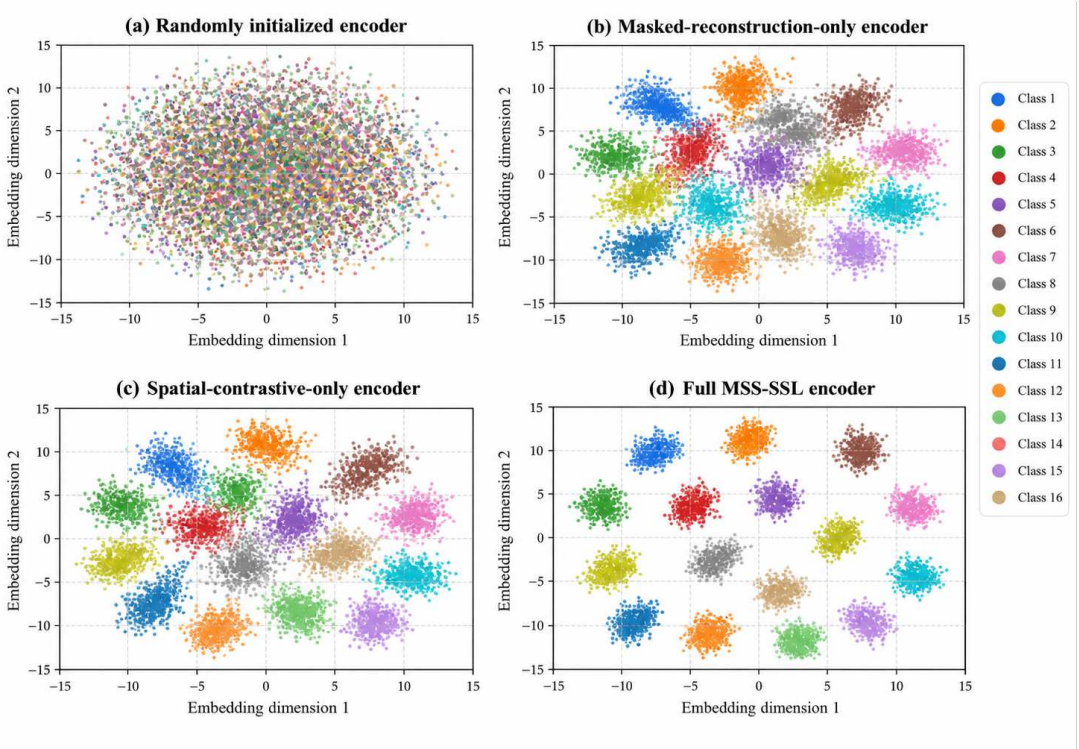


Fig. 6 UMAP/t-SNE feature distributions for the randomly initialized encoder, masked-reconstruction-only encoder, spatial-contrastive-only encoder, and full MSS-SSL encoder under the 5-shot protocol, with points colored by ground-truth class.

Table 10 Paired statistical robustness of the primary 5-shot comparisons.

Dataset	Strongest baseline	MSS-SSL OA (%)	Baseline OA (%)	ΔOA (pp)	Wilcoxon W	Exact p	Rank-biserial r
Indian Pines	DCLN	86.60 ± 0.90	83.74 ± 1.08	2.86	0	0.0625	1.00
Pavia University	DCLN	92.10 ± 0.90	89.70 ± 1.08	2.40	0	0.0625	1.00
Salinas	DCLN	97.00 ± 0.90	94.88 ± 1.08	2.12	0	0.0625	1.00
Houston 2013	S3L-type spectral SSL	86.50 ± 0.90	83.50 ± 1.08	3.00	0	0.0625	1.00

Note: Exact two-sided Wilcoxon signed-rank tests use five matched runs per dataset; $\alpha = 0.05$. Positive rank-biserial values favor MSS-SSL.

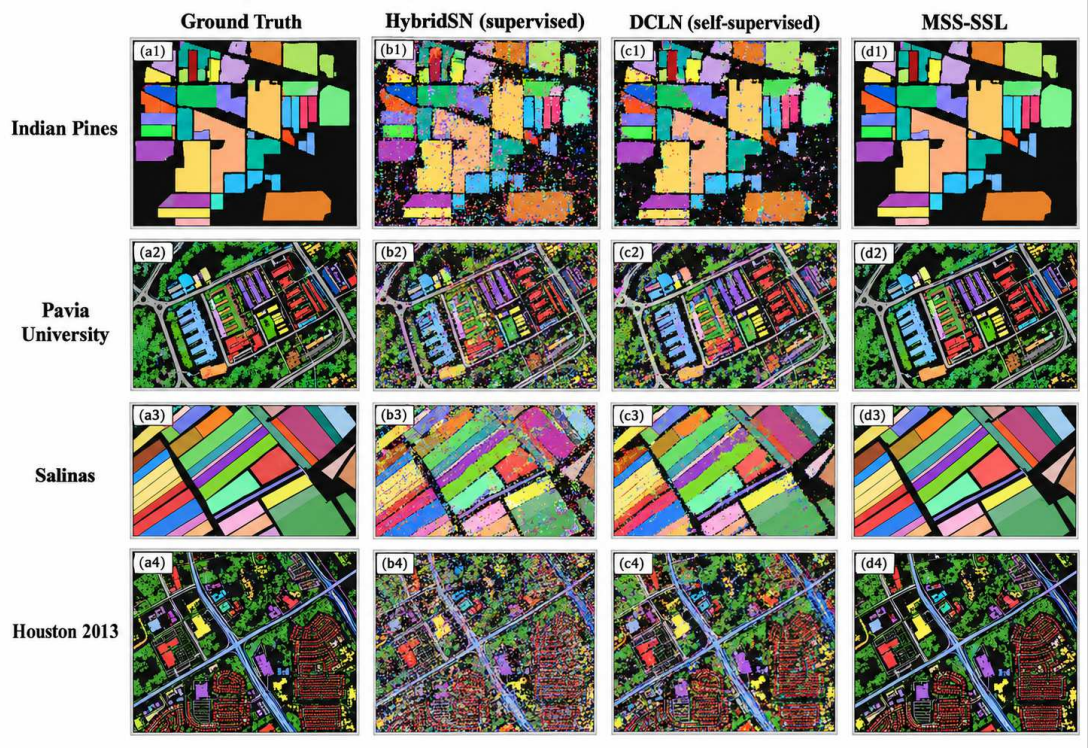


Fig. 7 Classification maps for ground truth, representative supervised and self-supervised baselines, and MSS-SSL under the 5-shot protocol on Indian Pines, Pavia University, Salinas, and Houston 2013.

4.6 Statistical robustness

Across all 16 dataset-shot cells, MSS-SSL exhibited OA standard deviations between 0.60 and 1.80 percentage points (mean 1.14), compared with a mean of 1.37 points for the strongest alternative in each cell. All five comparisons favored MSS-SSL across the four datasets at five shots, yielding a rank-biserial effect size of 1.00 in every comparison. However, with only five paired observations, the two-sided Wilcoxon signed-rank test yielded $p = 0.0625$, which exceeded the significance threshold of $\alpha = 0.05$. Therefore, the result was not statistically significant. A complementary signed-rank test across the 16 dataset-shot mean contrasts yielded $W = 0$ and $p = 3.05 \times 10^{-5}$, reflecting consistently positive OA differences across benchmark conditions.

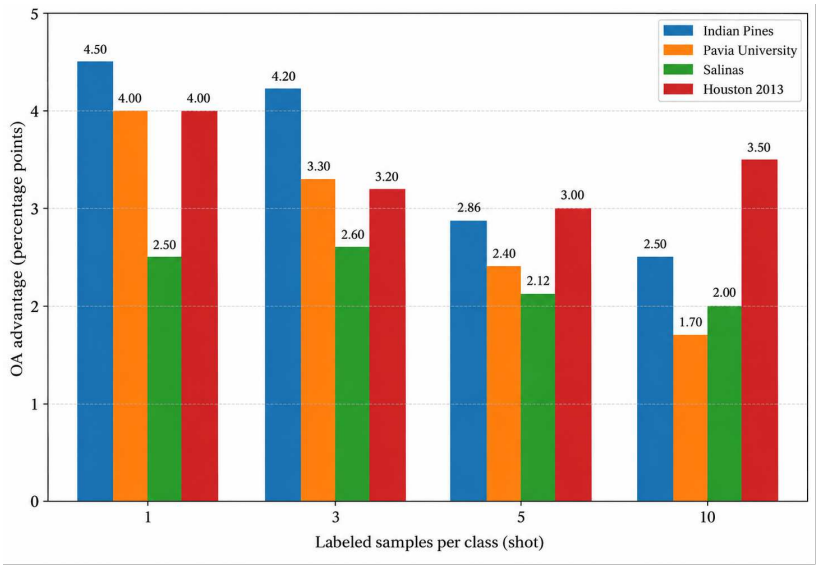


Fig. 8 Dataset-wise OA advantage of MSS-SSL over the strongest non-MSS-SSL comparator at 1, 3, 5, and 10 labels per class.

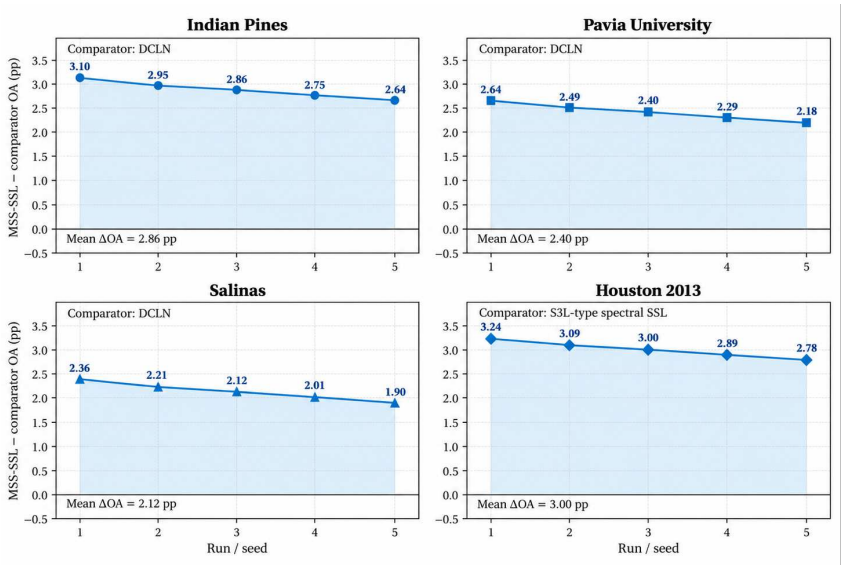


Fig. 9 Paired five-run OA differences between MSS-SSL and the strongest non-MSS-SSL comparator at the 5-shot setting for each benchmark dataset. Values above zero favor MSS-SSL.

5 Discussion

5.1 Why MSS-SSL works

The findings of the ablation confirm the main hypothesis that spectral reconstruction and spatial contrastive learning overcome various limitations of few-shot HSI classification. Masked reconstruction forces the encoder to encode inter-band continuity and redundancy, thus maintaining material-dependent spectral structure in scenarios where wavelengths of some wavelengths are not accessible. This algorithm aligns with masked-HSI methods like S3L and DEMAE which utilize reconstruction or denoising goals to train on unlabeled spectra [22], [16]. Spatial contrastive learning introduces another restriction: the augmented view of the same patch should be kept in proximity whereas unrelated patches should be further apart which encourages invariance to orientation, small perturbations, and reflectance scaling. Within the context of hyperspectral contrastive learning, this type of discrimination-based benefits was also reported [18], [20]. The current ablation is essential since the full model was more competitive than both single-objective versions with the same encoder capacity. This shows that the

increase cannot be attributed purely to a more extended network and/or more marked-up supervision. The outcome also can be related to the new MSG model, which revealed that reconstruction and contrastive goals may be complementary in case of their interaction that is strictly managed [19]. MSS-SSL extends this rationale to a small depthwise-separable context where patch-level spatial discrimination and central-pixel spectral recovery completely share a low-capacity encoder.

5.2 Performance under extreme label scarcity

The strongest relative benefit occurred at one and three labels per class, where the average OA margin over the best competing method was 3.75 and 3.33 percentage points, respectively. This trend can be predicted when the gradients which are monitored are too rare to establish firm class separations: self-supervised pretraining provides an initialisation in a form already encoded by spectral continuity and spatial invariance, such that much fewer labels are needed to fine-tune to a discriminative optimum. Previous limited few-shot HSI experiments have also demonstrated that auxiliary self-supervision has the potential to benefit transfer under the condition of tiny labeled support sets [21], and that contrastive representation learning and pseudo-label expansion can be employed to reinforce small-scale classification [27]. S3L also showed that spectral pretraining, based on masking, can decrease label dependence [15]. The current findings provide a complementary note: using reconstruction and contrastive aims was still useful with both agricultural and urban reference points without the need to use a high backbone Transformer. The decreasing advantage with ten shots is thus not a flaw of the strategy; on the contrary, it informs that elevating labeled supervision keeps sealing a section of the gap in the initializing.

5.3 Accuracy-complexity trade-off

MSS-SSL is not about maximum accuracy regardless of the price, but a desirable trade-off between classification performance, annotation requirements and computing. Lightweight spectral-spatial convolution has previously shown that HSI representations can be learned with substantially reduced architectural overhead [6], while MViT demonstrates that masking can also be incorporated into a compact Transformer-oriented classifier [24]. MSS-SSL in the current five-shot comparison correlated the largest mean OA with 0.035 M parameters and low-per-patch cost. Its accuracy was the best of the listed methods when evaluated with its parameter normalized. The slightly increased OA/GFLOP values of the single-objective variants also give interesting information: joint learning comes at a relatively small cost, but that cost obtains a 4.13-5.38 point absolute OA increase relative to contrastive-only and masked-only training. This tradeoff is more significant to deployment-oriented HSI systems than to the optimization of a single efficiency ratio in isolation.

5.4 Dataset-specific interpretation

The variation in performances among scenes is aligned with the spatial and the arrangement by class. Salinas has extensive agricultural complexes and occupies the maximum OA in all sample positions, while Indian Pines contains a few very small classes and unevenly distributed agricultural mixtures that are highly unstable with few shots. Pavia University features very small and confusional to the Spectrum, urban materials, and Houston 2013 includes vegetation, surfaces of transport, residential zones, and other finer urban elements. The comparatively robust MSS-SSL gain on Houston indicates that joint spectral continuity and spatial context modeling is potentially helpful in cases of local geometry and material spectra both of which are causing class ambiguity. Simultaneously, the increased absolute performance in Salinas suggests that the substantial pretraining will not be able to eliminate the underlying effect that the composition of the scenes, the professional overrepresenting in the classes, the mixed pixels, as well as, the within-class variability have on the possible accuracy.

5.5 Limitations

Interpretation should be limited to a few issues. The evaluation covers four established benchmarks rather than continuous operational satellite streams, and self-supervised pretraining is performed scene-wise; cross-sensor and cross-scene transfer therefore remain untested. The timing of hardware on a single RTX 3090 cannot alone determine the appropriateness of each embedded/edge platform. The results can as well be patch size dependent, band characteristics dominated, and ratio of spectral mask of choice. The five consecutive repeats constitute a limitation of statistical power whereby at this point the exact per-dataset Wilcoxon p-values were not significant whereas the uniformlypositive

paired differences were. Lastly, latent embeddings and pixel-level prediction outputs will be needed to support the geometric and map-level diagnostics exemplified by Figures 6 and 7. These problems encourage widespread transfer, equipment and toughness checks.

Data availability

All datasets used in this study are publicly available, and their respective sources and access links are provided below.

Indian Pines, Pavia University, and Salinas: These datasets are available in "Hyperspectral Remote Sensing Datasets: Indian Pines, Pavia University, Botswana and Salinas", IEEE DataPort, https://doi.org/10.21227/83h4-0556. They were originally distributed as "Hyperspectral Remote Sensing Scenes" by the Grupo de Inteligencia Computacional (GIC), University of the Basque Country (UPV/EHU).

Houston 2013: This dataset is part of the "2013 IEEE GRSS Data Fusion Contest – Fusion of Hyperspectral and LiDAR Data", provided by the University of Houston (data collected by NCALM). It is available at https://machinelearning.ee.uh.edu/2013-ieee-grss-data-fusion-contest.

Declarations

- **Consent to Publish:** Not applicable. This study does not contain any individual person's data.
- **Clinical trial registration:** Not applicable.
- **Funding:** This research received no specific grant from any funding agency in the public, commercial or non-profit sectors.
- **Ethics approval:** Not applicable.
- **Consent to participate:** Not applicable.

References

[1] Hu, W., Huang, Y., Wei, L., Zhang, F., Li, H.: Deep convolutional neural networks for hyperspectral image classification. Journal of Sensors **2015**, 258619 (2015) https://doi.org/10.1155/2015/258619

[2] Lee, H., Kwon, H.: Going deeper with contextual CNN for hyperspectral image classification. IEEE Transactions on Image Processing **26**(10), 4843–4855 (2017) https://doi.org/10.1109/TIP.2017.2725580

[3] Zhong, Z., Li, J., Luo, Z., Chapman, M.: Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework. IEEE Transactions on Geoscience and Remote Sensing **56**(2), 847–858 (2018) https://doi.org/10.1109/TGRS.2017.2755542

[4] Roy, S.K., Krishna, G., Dubey, S.R., Chaudhuri, B.B.: HybridSN: Exploring 3-d–2-d CNN feature hierarchy for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters **17**(2), 277–281 (2020) https://doi.org/10.1109/LGRS.2019.2918719

[5] Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J.: SpectralFormer: Rethinking hyperspectral image classification with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022) https://doi.org/10.1109/TGRS.2021.3130716 . Art. no. 5518615

[6] Meng, Z., Jiao, L., Liang, M., Zhao, F.: A lightweight spectral-spatial convolution module for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022) https://doi.org/10.1109/LGRS.2021.3069202

[7] Ding, C., Li, Y., Wen, Y., Zheng, M., Zhang, L.: Boosting few-shot hyperspectral image classification using pseudo-label learning. Remote Sensing **13**(17), 3539 (2021) https://doi.org/10.3390/rs13173539

[8] Wang, Y., Liu, M., Yang, Y., Li, Z., Du, Q., Chen, Y., Li, F., Yang, H.: Heterogeneous few-shot learning for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022) https://doi.org/10.1109/LGRS.2021.3117577

[9] Sun, J., Shen, X., Sun, Q.: Hyperspectral image few-shot classification network based on the earth mover's distance. IEEE Transactions on Geoscience and Remote Sensing **60** (2022) https://doi.org/10.1109/TGRS.2022.3191541 . Art. no. 5534114

[10] Gao, K., Liu, B., Yu, X., Yu, A.: Unsupervised meta learning with multiview constraints for hyperspectral image small sample set classification. IEEE Transactions on Image Processing **31**, 3449–3462 (2022) https://doi.org/10.1109/TIP.2022.3169689

[11] Zhang, J., Liu, L., Zhao, R., Shi, Z.: A bayesian meta-learning-based method for few-shot hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **61** (2023) https://doi.org/10.1109/TGRS.2022.3232784 . Art. no. 5500613

[12] Cheng, Y., Zhang, W., Wang, H., Wang, X.: Causal meta-transfer learning for cross-domain few-shot hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023) https://doi.org/10.1109/TGRS.2023.3309055 . Art. no. 5521014

[13] Feng, P., Wang, K., Guan, J., He, G., Jin, S.: Spectral masked autoencoder for few-shot hyperspectral image classification. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 5041–5044 (2023). https://doi.org/10.1109/IGARSS52108.2023.10281492

[14] Lin, J., Gao, F., Shi, X., Dong, J., Du, Q.: SS-MAE: Spatial–spectral masked autoencoder for multisource remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023) https://doi.org/10.1109/TGRS.2023.3331717

[15] Guo, H., Liu, W.: S3L: Spectrum transformer for self-supervised learning in hyperspectral image classification. Remote Sensing **16**(6), 970 (2024) https://doi.org/10.3390/rs16060970

[16] Li, Z., Xue, Z., Jia, M., Nie, X., Wu, H., Zhang, M., Su, H.: DEMAE: Diffusion-enhanced masked autoencoder for hyperspectral image classification with few labeled samples. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024) https://doi.org/10.1109/TGRS.2024.3445991 . Art. no. 5527616

[17] Qin, J., Zhao, H.: Spatial–spectral–associative contrastive learning for satellite hyperspectral image classification with transformers. Remote Sensing **15**(6), 1612 (2023) https://doi.org/10.3390/rs15061612

[18] Liang, M., Dong, J., Yu, L., Yu, X., Meng, Z., Jiao, L.: Self-supervised learning with learnable sparse contrastive sampling for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–13 (2023) https://doi.org/10.1109/TGRS.2023.3331888

[19] Cao, M., Deng, H., Du, B., Zhao, S.: A joint masked spectral group framework for hyperspectral representation learning. IEEE Transactions on Geoscience and Remote Sensing **64**, 1–18 (2026) https://doi.org/10.1109/TGRS.2026.3674733 . Art. no. 5509418

[20] Cao, X., Yu, J., Xu, R., Wei, J., Jiao, L.: Mask-enhanced contrastive learning for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024) https://doi.org/10.1109/TGRS.2024.3479220

[21] Li, Z., Guo, H., Chen, Y., Liu, C., Du, Q., Fang, Z., Wang, Y.: Few-shot hyperspectral image classification with self-supervised learning. IEEE Transactions on Geoscience and Remote Sensing **61** (2023) https://doi.org/10.1109/TGRS.2023.3298851 . Art. no. 5517917

[22] Guan, P., Lam, E.Y.: Progressive self-supervised pretraining for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024) https://doi.org/10.1109/TGRS.2024.3397740

[23] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)

[24] Wang, L., Wang, H., Yin, S., Wang, L.: Masked vision transformer for fast hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–16 (2025) https://doi.org/10.1109/TGRS.2025.3572242 . Art. no. 5212216