applied sciences

MDPI

Article

Open-Set Hyperspectral Land-Cover Classification Using Second-Order Attention and Gaussian Mixture Clustering

Chaoya Hu, Ruixue Xv and Mianshu Chen *

School of Communication Engineering, Jilin University, Changchun 130025, China; huchaoya93@gmail.com (C.H.); rxxu23@163.com (R.X.)
* Correspondence: chenms@jlu.edu.cn

Featured Application

Hyperspectral land-cover mapping with detection of classes absent from the training data.

Abstract

Open-set hyperspectral classification requires accurate recognition of known land-cover classes and detection of samples from unseen classes. We propose GRDOSR, which extends representative-discriminative open-set recognition (RDOSR) by combining second-order attention with Gaussian mixture modeling. Spatial and channel attention capture local feature dependencies, while mixture components define specialized networks that jointly learn reconstruction and classification. Their predictions are combined using mixture posteriors, and a score based on reconstruction error and mixture likelihood is calibrated using known validation data for unknown rejection. Across five hyperspectral benchmarks, GRDOSR improves the mean area under the receiver operating characteristic curve (AUC) by 1.43–11.57 percentage points and known-class overall accuracy (OA) by 0.27–13.57 points relative to RDOSR, with a higher mean open-set classification rate (OSCR) on every dataset. It achieves the highest mean AUC and OSCR on Indian Pines and Pavia University datasets among the compared methods. These results support the use of local context and modeling of known feature distributions for land-cover mapping when the training data cover only part of the class inventory.

Keywords: hyperspectral imaging; open-set recognition; land-cover classification; second-order attention; Gaussian mixture model; multitask learning

Check for updates

Academic Editor: Nathan J. Moore

Received: 7 September 2026
Revised: 15 October 2026
Accepted: 19 October 2026
Published: 22 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Hyperspectral imagery provides detailed spectral information for identifying land-cover materials [1]. In practical mapping, labeled samples may cover only part of the material classes present in a scene. A classifier trained on these samples must recognize familiar materials while detecting previously unseen crop types or urban surfaces. Conventional supervised classifiers assume that all test classes are represented during training and assign each pixel to a known class, even when its material is unfamiliar. Open-set recognition (OSR) extends classification to this setting by distinguishing known classes and rejecting samples from classes that are absent during training [2–4].

Representative-discriminative open-set recognition (RDOSR) couples a supervised spectral embedding with a network that jointly reconstructs features and classifies known samples [5]. Reconstruction complements classification by measuring how closely an input matches the learned representation. However, a single network must represent the

heterogeneous features of all known materials, and its pixelwise input does not explicitly exploit correlations between neighboring pixels or feature channels. Moreover, an unfamiliar spectrum can yield a small reconstruction error if its embedded features resemble those of known materials. Local context and feature distribution modeling therefore offer complementary ways to improve recognition in this setting.

We propose GRDOSR to combine local feature refinement and Gaussian mixture modeling within the RDOSR framework. Second-order spatial and channel attention refines local embeddings by weighting correlations between neighboring positions and feature channels. A Gaussian mixture model (GMM) then groups the refined known features, allowing separate multitask experts to learn different regions of their distribution. At inference, component posterior probabilities weight the predictions of all experts, while reconstruction error and mixture likelihood jointly determine whether a sample should be rejected. This formulation combines local feature refinement, specialized reconstruction and classification, and density-based unknown detection within one framework. The experiments assess the resulting gains in known-class recognition and unknown detection, together with the contributions of attention, expert specialization, and score fusion.

2. Related Work

Classical OSR methods estimate whether a sample is consistent with known classes using probability models, nearest-neighbor distance ratios, or sparse reconstruction [3,6,7]. Zhang and Patel [7] characterized reconstruction error distributions obtained from sparse representations. Deep approaches include activation recalibration in OpenMax [8], generation of unknown examples in Generative OpenMax [9], classwise rejection in deep open classification for text documents [10], and class-specific semantic reconstruction [11].

Joint learning of reconstruction and classification provides a representation for both known-class discrimination and unknown detection. Oza and Patel [12] combined an encoder, decoder, and classifier in a multitask network, while classification–reconstruction learning preserved latent information that is useful for separating unknown samples [13]. C2AE uses class-conditioned decoding [14], and generative–discriminative feature learning evaluated consistency between inputs and their reconstructions [15]. RDOSR applied representative-discriminative learning to satellite land-cover classification [5]. Conditional Gaussian distributions and mixtures of exponential-power distributions modeled the distribution of known features explicitly [16,17], providing a basis for detecting unfamiliar samples that can nevertheless be reconstructed accurately.

Hyperspectral OSR exploits both the structure of spectral features and spatial context. Prototype-based approaches learn feature-consistent representations [18] or feature-invariant prototypes with weighted Pearson distances [19]. IADMRN combines attention, dense connections, and multiscale reconstruction with extreme-value-based rejection [20]; HyperCASR integrates spectral–spatial transformer features with class-specific semantic autoencoders [21]. Other formulations address few-shot recognition through self-supervised multitask learning with adaptive thresholds (SSMTL) [22], use unlabeled mixtures of known and unknown samples in HOpenCls [23], or consider cross-domain recognition through critical-reflective learning [24].

Attention also strengthens feature learning in closed-set hyperspectral classification. Attention-aided CNNs learn spectral and spatial weights [25], and A-SPN combines attention with second-order pooling [26]. Global second-order pooling similarly uses second-order statistics to weight spatial and channel features [27]. GRDOSR uses local correlations to refine the class-logit embedding, preserving a K-dimensional feature vector for component-specific OSR networks.

3. Materials and Methods

3.1. Feature Extraction and Second-Order Attention

Figure 1 presents the GRDOSR architecture. For B spectral bands and I known classes, the feature extractor F has fully connected layers with dimensions $B \to 512 \to 1024 \to 512 \to 32 \to I$. Hidden layers use ReLU activations, and the output consists of linear logits $z(x)$, giving an embedding dimension of $K = I$. For N_k known training spectra, let $z_i = F(x_i) \in \mathbb{R}^I$ and $\mathbf{y}_i$ be the one-hot label of sample i. The embedding is learned by minimizing

$$\mathcal{L}_F = \frac{1}{N_k} \sum_{i=1}^{N_k} \left[-\mathbf{y}_i^{\mathsf{T}} \log \operatorname{softmax}(z_i) + 0.1 \|z_i\|_1 \right]. \tag{1}$$

Here, $x_i \in \mathbb{R}^B$ is the input spectrum after bandwise standardization using known training statistics and scaling by 0.1. Logarithms are natural and apply elementwise to vectors. The notation $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_F$ denotes the ℓ_1, Euclidean, and Frobenius norms, respectively, and superscript T denotes transposition.

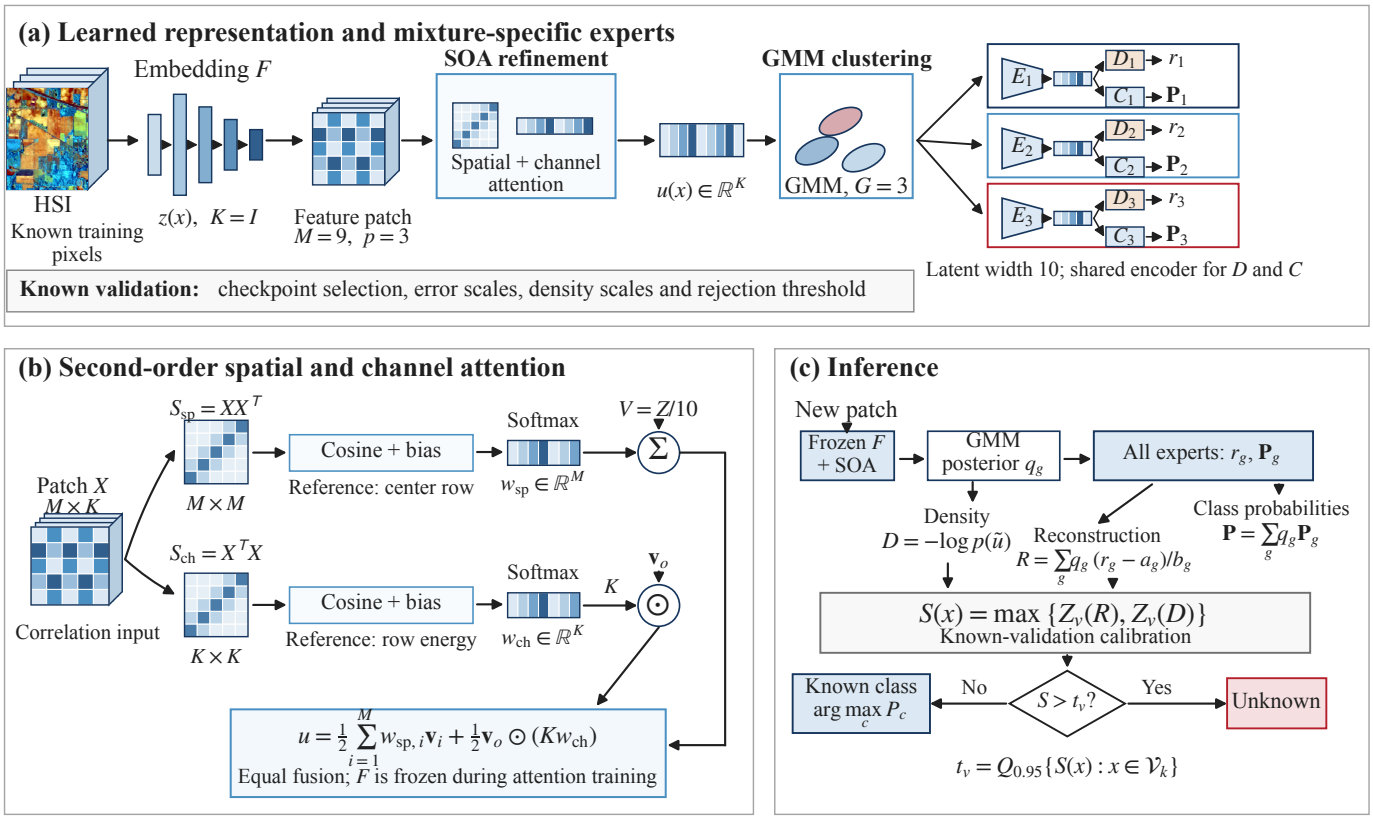


Figure 1. GRDOSR architecture. (**a**) Spectral embedding, feature refinement, and component-specific encoder–decoder–classifier networks. (**b**) Spatial and channel attention. (**c**) Posterior aggregation and rejection based on reconstruction error and mixture density. HSI, SOA, and GMM denote hyperspectral imagery, second-order attention, and a Gaussian mixture model. Colors are schematic; they distinguish feature blocks and model components, and the red output box denotes an unknown prediction.

Second-order attention (SOA) is computed from a $p \times p$ neighborhood, with $p = 3$ and $M = p^2 = 9$ embedded pixels. Their logits form $Z \in \mathbb{R}^{M \times K}$. Channelwise standardization using statistics from known training centers gives A, which is Frobenius-normalized as $X = A / \max(\|A\|_F, \epsilon)$, with $\epsilon = 10^{-8}$. The spatial and channel Gram matrices are

$$S_{\mathrm{sp}} = XX^{\mathsf{T}} \in \mathbb{R}^{M \times M}, \qquad S_{\mathrm{ch}} = X^{\mathsf{T}}X \in \mathbb{R}^{K \times K}. \tag{2}$$

For branch $b \in \{\mathrm{sp}, \mathrm{ch}\}$, let $d_{\mathrm{sp}} = M$ and $d_{\mathrm{ch}} = K$. The ith Gram row, written as a column vector, is denoted by $s_{b,i} \in \mathbb{R}^{d_b}$. Its cosine similarity to reference r_b defines the corresponding attention weight through

$$\rho_{b,i} = \frac{s_{b,i}^{\mathsf{T}} r_b}{\max(\|s_{b,i}\|_2 \|r_b\|_2, \epsilon)}, \qquad w_{b,i} = \frac{\exp(\rho_{b,i} + \beta_{b,i})}{\sum_{j=1}^{d_b} \exp(\rho_{b,j} + \beta_{b,j})}. \tag{3}$$

The indices satisfy $i, j \in \{1, \ldots, d_b\}$, and the trainable bias $\beta_{b,i}$ is initialized to zero. Spatial attention uses the central Gram row as its reference, with $r_{\mathrm{sp}} = s_{\mathrm{sp},o}$, where $o = \lfloor M/2 \rfloor + 1 = 5$. For channel attention, the reference is $(r_{\mathrm{ch}})_j = \sqrt{K^{-1} \sum_{h=1}^{K} (S_{\mathrm{ch}})_{hj}^2 + 10^{-12}}$. Let $V = Z/10$ be the value matrix and $\mathbf{v}_i = V_{i,:}^{\mathsf{T}}$ its rows written as column vectors. The two branches produce the refined feature

$$u(x) = \tfrac{1}{2} \sum_{i=1}^{M} w_{\mathrm{sp},i} \mathbf{v}_i + \tfrac{1}{2} \mathbf{v}_o \odot (K w_{\mathrm{ch}}) \in \mathbb{R}^K. \tag{4}$$

Here, $w_{\mathrm{ch}} = (w_{\mathrm{ch},1}, \ldots, w_{\mathrm{ch},K})^{\mathsf{T}}$ and $\odot$ denotes elementwise multiplication. The normalized features X determine the correlations, while V retains the magnitudes of the logits. Spatial weights sum to one; multiplying the channel weights by K gives a gate with unit mean, so uniform channel attention leaves the central feature unchanged within that branch. A temporary linear classifier trains the attention biases through known-class cross-entropy (CE), with F held fixed.

3.2. Gaussian Mixture Clustering and Multitask Experts

The refined vectors u are standardized using known training statistics to obtain $\tilde{u}$. A GMM represents the resulting feature distribution [28], with density and component posterior probabilities given by

$$p(\tilde{u}) = \sum_{g=1}^{G} \pi_g \mathcal{N}(\tilde{u} \mid \mu_g, \Sigma_g), \qquad q_g(x) = \frac{\pi_g \mathcal{N}(\tilde{u}(x) \mid \mu_g, \Sigma_g)}{p(\tilde{u}(x))}. \tag{5}$$

Here, G is the number of components, with weights $\pi_g \geq 0$ and $\sum_g \pi_g = 1$. The Gaussian density $\mathcal{N}(\cdot \mid \mu_g, \Sigma_g)$ has mean $\mu_g \in \mathbb{R}^K$ and covariance $\Sigma_g \in \mathbb{R}^{K \times K}$. Expectation maximization (EM) fits the mixture to known training features [29]. Assigning each training center to the component with the largest posterior probability defines the subsets for expert learning. These assignments remain fixed while the experts are optimized. Each expert predicts all I known classes, following the mixture-of-experts principle [30].

Expert g comprises an encoder E_g, decoder D_g, and classifier C_g. The encoder follows the RDOSR stick-breaking representation [5]. One branch contains four linear layers with three units each, whose outputs are concatenated with preceding features, followed by ten sigmoid outputs a_j. A second branch contains two such layers followed by a scalar softplus output η. The outputs are constrained to $a_j \in [10^{-7}, 1 - 10^{-7}]$ and $\eta \geq 10^{-6}$. The latent coefficients are $v_j = a_j^{1/\eta}$ and $s_j = v_j \prod_{h=1}^{j-1} (1 - v_h)$ for $j = 1, \ldots, 10$, with the empty product equal to one. A bias-free linear decoder $(10 \to 10 \to K)$ reconstructs u, while a linear softmax classifier $(10 \to I)$ predicts its known class. For N_g training centers assigned to component g, let $u_i = u(x_i)$ and s_{ij} denote coefficient j for sample i, with $\hat{s}_{ij} = s_{ij} / (\sum_{h=1}^{10} s_{ih} + \epsilon)$. The expert minimizes

$$\begin{aligned} \mathcal{L}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} \Big[& 0.5 \|u_i - D_g(E_g(u_i))\|_2^2 - 0.5 \mathbf{y}_i^{\mathsf{T}} \log C_g(E_g(u_i)) \\ & - 0.001 \sum_{j=1}^{10} \hat{s}_{ij} \log(\hat{s}_{ij} + \epsilon) \Big]. \end{aligned} \tag{6}$$

The shared encoder supports reconstruction and known-class discrimination, while the entropy penalty favors sparse normalized latent coefficients.

3.3. Posterior Aggregation and Rejection

For expert g, define the reconstruction error $r_g(x) = \|u(x) - D_g(E_g(u(x)))\|_2^2$ and the known-class probability vector $\mathbf{P}_g(x) = C_g(E_g(u(x))) \in [0,1]^I$. The mean a_g and population standard deviation b_g of reconstruction errors are estimated from known validation samples weighted by q_g. If the total posterior weight is below eight, unweighted statistics from the same validation samples are used, with b_g bounded below by 10^{-6}. Every pixel is processed by all experts, and its component posteriors weight their normalized errors and class probabilities:

$$R(x) = \sum_{g=1}^{G} q_g(x) \frac{r_g(x) - a_g}{b_g}, \qquad \mathbf{P}(x) = \sum_{g=1}^{G} q_g(x) \mathbf{P}_g(x). \tag{7}$$

The mixture density measures how well a feature is supported by the known training distribution. Its negative log likelihood is $D(x) = -\log p(\tilde{u}(x))$. For $h \in \{R, D\}$, let $\mu_{h,v}$ and $\sigma_{h,v}$ be the mean and population standard deviation over known validation centers $\mathcal{V}_k$. Standardizing the two signals gives the rejection score

$$\begin{aligned} Z_v(h(x)) &= \frac{h(x) - \mu_{h,v}}{\max(\sigma_{h,v}, 10^{-6})}, \\ S(x) &= \max\{Z_v(R(x)), Z_v(D(x))\}. \end{aligned} \tag{8}$$

The threshold $t_v = Q_{0.95}(\{S(x) : x \in \mathcal{V}_k\})$ is the empirical 95th percentile of known validation scores. A pixel is rejected as unknown if $S(x) > t_v$; otherwise, it is assigned to $\hat{y}(x) = \arg\max_{c=1,\ldots,I} P_c(x)$, where P_c is the cth entry of $\mathbf{P}$.

4. Experimental Settings

4.1. Data and Spatial Splits

The experiments used five benchmarks from the Hyperspectral Remote Sensing Scenes repository [31]: Indian Pines (IN), Pavia University (PU), Pavia Center (PC), Salinas (SA), and Kennedy Space Center (KSC). IN has 145 × 145 pixels, 200 spectral bands, and 16 classes; PU has 610 × 340 pixels, 103 bands, and nine classes; PC has 1096 × 715 pixels, 102 bands, and nine classes. SA has 512 × 217 pixels, 204 bands, and 16 classes, while KSC has 512 × 614 pixels, 176 bands, and 13 classes. IN and SA depict agricultural areas acquired by AVIRIS, KSC is an AVIRIS wetland scene, and PU and PC are urban ROSIS scenes.

Each class is held out in turn as unknown, giving 63 single-class holdout configurations across the five datasets. Spatial block partitioning separates the training, validation, and test regions to reduce dependence from overlapping neighborhoods [32]. Block side lengths are six pixels for IN; 24 for PU, PC, and SA; and 12 for KSC. The target proportions were 60%, 20%, and 20%. Eroding block boundaries by one pixel prevents overlap between 3 × 3 patch supports. Training and validation windows containing a held-out-class pixel are excluded. The embedding F uses known spectra within training blocks; attention and expert learning use eligible patch centers. Block assignments are balanced by class counts. Input and feature standardization and GMM fitting use only known training data.

The random seeds 3020, 3021, and 3022 defined three repetitions with different spatial partitions and model initializations. Within each repetition, dataset-level metrics averaged the held-out classes equally; reported means and sample standard deviations (SDs) were then computed across the three repetitions. Paired differences and 95% confidence intervals used the corresponding run-level means. Scene maps and example curves used seed 3020, with unknown class $L = 3$ for IN, PU, SA, and KSC and $L = 6$ for PC. Experiments with two unknown classes held out IN $\{3, 7\}$, PU $\{3, 5\}$, PC $\{3, 6\}$, SA $\{3, 7\}$, and KSC $\{3, 7\}$. Together, the single-class and two-unknown-class configurations yielded 204 runs.

4.2. Training and Comparison Methods

The GRDOSR networks were trained using Adam [33], with a constant learning rate of 10^{-3}. The embedding F and each expert used at most 15,000 optimization steps and batches of up to 2048 samples; attention used at most 5000 steps and batches of up to 512. Attention training included a 10^{-4} squared-bias penalty in addition to CE.

The GMM used $G = 3$ and full covariance matrices throughout the main comparison. EM used diagonal regularization of 10^{-6}, three k-means initializations, a tolerance of 10^{-3}, and at most 200 iterations. The solution with the highest lower bound on the training-data log likelihood was retained. Sensitivity experiments compared $G = 1, 2, 3, 4$.

Validation losses were checked after the first optimization step and then every 250 steps for F and attention, and every 500 steps for the experts. The criteria were CE plus logit sparsity for F, CE for attention, and 0.5 reconstruction loss plus 0.5 CE for each expert. Expert validation used posterior-assigned known centers; all known validation centers were used if fewer than five were assigned. Early stopping was allowed after 2000, 1500, and 3000 steps for F, attention, and experts, with patience amounts of eight, six, and six checks, respectively. Improvements needed to exceed 10^{-5}, and the lowest-loss checkpoint among all validation checks was restored.

Experiments were run on an RTX 4090 D under Ubuntu 20.04.3, with NVIDIA driver 580.105.08, NVIDIA TensorFlow 1.15.5+nv21.10, Python 3.8.10, CUDA 11.4, and cuDNN 8.2.4. Softmax, OpenMax, AEC, and AECD inference used FP32 and fixed-size batches; padded entries were excluded from predictions and calibration. OpenAI Codex assisted implementation and manuscript editing; the authors reviewed and verified its outputs.

All methods used the same spatial test masks. Softmax rejects samples according to one minus their largest predicted class probability [34]. OpenMax fits Weibull tails to correctly classified known training activations [8], with a tail size of at most ten and recalibration rank $\min(10, I)$. Its distance measure combines Euclidean distance scaled by $1/200$ with cosine distance. Classes with fewer than two correctly classified training samples receive up to 2000 additional class-balanced training steps, with checkpoints selected by known-validation CE. AEC and AECD jointly reconstruct and classify spectral inputs using affine and stick-breaking encoders, respectively. AEC denotes the autoencoder–classifier control. RDOSR uses the same embedding and expert objectives with $G = 1$ and no attention. RDOSR+GS adds single-Gaussian support using the same maximum-score fusion as GRDOSR, whereas RDOSR+GMM uses three experts and Gaussian support without attention.

SSMTL-A adapts the published few-shot SSMTL network [22] to the spatial sampling protocol. Its patch size was reduced from $p = 9$ to $p = 3$, and training used all eligible known centers. PCA fitted to known training spectra retained 64 whitened channels. Reconstruction pretraining lasted 600 steps, followed by at most 1500 multitask steps at each Adadelta learning rate of 1, 0.1, and 0.01. Known validation data determined checkpoints and rejection thresholds. The same adaptation was used for every single-class and two-unknown-class experiment.

Component and expert-count analyses used $L = 3$ on IN, PU, SA, and KSC and $L = 6$ on PC, repeated with the same three seeds. Component analyses compared spatial, channel, and dual attentions at $G \in \{1, 3\}$. These variants shared the feature checkpoint, value representation, loss, spatial split, and joint rejection rule. Mean pooling and uniform SOA distinguished local averaging from learned weighting. Random training groups matched the GMM component sizes and retained the fitted posterior weights, density model, and expert validation cohorts, so only the training membership changed.

Models were trained on known blocks within each scene under the same inference and calibration rules. SA and KSC were excluded from the method selection.

Source code and reproducibility resources are provided in the Supplementary Material, `GRDOSR_method_reproducibility`.

4.3. Evaluation

Unknown samples were the positive class for detection. ROC-AUC measures how well rejection scores rank unknown samples above known samples [35]. The true-positive rate (TPR), false-positive rate (FPR), and area under the ROC curve are

$$\mathrm{TPR}(t) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FPR}(t) = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \quad \mathrm{AUC} = \int_0^1 \mathrm{TPR}\, \mathrm{dFPR}. \tag{9}$$

With $S(x) > t$ denoting rejection, TP and FN are the numbers of rejected and accepted unknown pixels, and FP and TN are the numbers of rejected and accepted known pixels. Known-class overall accuracy (OA), average per-class accuracy (AA), and macro-F1 are computed before rejection. F1 is the harmonic mean of precision and recall. AA averages classwise accuracy over known classes present in the test set; macro-F1 averages over the full known-class inventory, assigning zero to undefined scores. At the operating threshold t_v, known-class OA after rejection is the fraction of known test pixels that are both correctly classified and accepted. Open-set macro-F1 includes the known classes and one pooled unknown class, with unknown precision, recall, and F1 reported separately. Detection ranking is also assessed by AUPR, computed as the average precision, and FPR at 95% unknown recall (FPR95). Precision–recall measures complement ROC-AUC under class imbalance [36]. Following Dhamija et al. [37], OSCR measures correct known-class recognition jointly with unknown rejection:

$$\begin{aligned} \mathrm{CCR}(t) &= \frac{1}{n_k} \sum_{x \in \mathcal{I}_k} \mathbf{1}[\hat{y}(x) = y(x), S(x) \le t], \\ \mathrm{UAR}(t) &= \frac{1}{n_u} \sum_{x \in \mathcal{I}_u} \mathbf{1}[S(x) \le t], \qquad \mathrm{OSCR} = \int_0^1 \mathrm{CCR}\, \mathrm{dUAR}. \end{aligned} \tag{10}$$

Here, $\mathcal{I}_k$ and $\mathcal{I}_u$ are the known and unknown test sets, with sizes n_k and n_u, and $\mathbf{1}[\cdot]$ is the indicator function. The labels $y(x)$ and $\hat{y}(x)$ are the true and predicted classes before rejection. CCR denotes the correct classification rate and UAR the unknown acceptance rate. ROC-AUC and OSCR are computed by trapezoidal integration, grouping tied scores together.

5. Results and Discussion

5.1. Detection and Known-Class Classification

Table 1 summarizes the detection, known-class recognition, and joint performance on IN, PU, and PC. GRDOSR has the highest mean AUC and OSCR among the compared methods on IN and PU. On IN, the known-class OA increases from 71.38% for RDOSR to 84.95%, and the OSCR increases from 54.18% to 66.85%. On PC, RDOSR+GS has a higher mean AUC (94.08% versus 92.99%) and OSCR (93.49% versus 92.62%).

The class-balanced results in Table 2 distinguish recognition gains from the influence of class frequencies. GRDOSR achieves known-class macro-F1s of 66.77%, 95.91%, and 97.78% on IN, PU, and PC, exceeding RDOSR on each dataset. It also improves for Softmax on PU and PC, while SSMTL-A has the strongest AA and macro-F1 across the three datasets.

Table 1. Detection, known-class classification, and joint performance across all single-class holdouts. Each value is a percentage reported as the mean ± sample standard deviation (SD) over three runs. AUC is the area under the receiver operating characteristic (ROC) curve, known OA is the overall accuracy before rejection, and OSCR is the open-set classification rate. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best means, respectively. Upward arrows indicate that higher values are better.

	Indian Pines (IN)			Pavia University (PU)			Pavia Center (PC)		
Method	**AUC ↑**	**Known OA ↑**	**OSCR ↑**	**AUC ↑**	**Known OA ↑**	**OSCR ↑**	**AUC ↑**	**Known OA ↑**	**OSCR ↑**
Softmax	66.02 ± 1.79	85.16 ± 1.14	61.02 ± 1.51	53.69 ± 5.14	94.93 ± 0.66	52.92 ± 4.95	87.36 ± 7.19	98.85 ± 0.06	86.99 ± 7.07
OpenMax	66.46 ± 2.37	83.73 ± 1.31	57.31 ± 2.15	52.08 ± 11.88	94.69 ± 0.74	49.65 ± 11.44	79.03 ± 8.57	98.78 ± 0.08	78.48 ± 8.40
AEC	65.16 ± 1.55	72.63 ± 2.93	47.61 ± 1.62	59.63 ± 1.86	88.20 ± 0.79	52.84 ± 2.20	74.87 ± 0.67	98.23 ± 0.36	73.87 ± 0.85
AECD	67.36 ± 0.60	67.03 ± 2.92	46.96 ± 1.69	73.04 ± 4.87	90.22 ± 1.05	66.15 ± 4.93	92.05 ± 1.78	97.90 ± 0.40	90.38 ± 1.47
RDOSR	65.12 ± 2.50	71.38 ± 3.74	54.18 ± 0.89	69.23 ± 2.41	95.23 ± 0.32	67.21 ± 2.16	91.56 ± 2.92	99.06 ± 0.08	91.01 ± 2.95
RDOSR+GS	67.26 ± 1.78	71.38 ± 3.74	54.55 ± 1.46	66.32 ± 5.28	95.23 ± 0.32	64.15 ± 4.96	**94.08** ± 0.32	99.06 ± 0.08	**93.49** ± 0.36
RDOSR+GMM	69.99 ± 2.02	76.72 ± 1.53	61.03 ± 1.90	73.67 ± 1.49	95.29 ± 0.36	71.42 ± 1.21	92.24 ± 2.11	99.04 ± 0.08	91.63 ± 2.03
SSMTL-A	72.71 ± 1.03	**87.11** ± 2.27	64.23 ± 1.96	63.46 ± 2.78	**97.33** ± 0.21	61.93 ± 2.73	76.03 ± 1.60	**99.49** ± 0.16	75.83 ± 1.57
GRDOSR	**72.85** ± 3.52	84.95 ± 1.46	**66.85** ± 3.37	**76.12** ± 3.32	96.63 ± 0.37	**74.76** ± 3.18	92.99 ± 1.53	99.33 ± 0.04	92.62 ± 1.47

Table 2. Known-class and detection-ranking metrics across all single-class holdouts. Values are percentages (mean ± sample SD over three runs). AA is the average per-class accuracy, macro-F1 is the classwise macro-average of F1, AUPR is computed as the average precision, and FPR95 is the false-positive rate at 95% unknown recall. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best means, respectively. Upward and downward arrows indicate that higher and lower values are better, respectively.

		Known-Class Classification		Unknown-Class Detection	
Scene	**Method**	**Average Per-Class Accuracy (AA) ↑**	**Known-Class Macro-F1 ↑**	**AUPR ↑**	**FPR95 ↓**
IN	Softmax	78.53 ± 2.94	77.76 ± 4.90	12.29 ± 0.86	67.87 ± 1.57
	OpenMax	72.48 ± 4.22	72.00 ± 6.40	**19.96** ± 1.82	58.78 ± 4.04
	AEC	62.69 ± 1.14	62.20 ± 1.00	15.79 ± 1.34	70.61 ± 3.69
	AECD	48.45 ± 0.81	46.49 ± 0.60	16.46 ± 2.35	64.29 ± 0.79
	RDOSR	52.61 ± 0.53	47.75 ± 1.66	14.85 ± 2.25	63.17 ± 2.78
	RDOSR+GS	52.61 ± 0.53	47.75 ± 1.66	14.73 ± 0.48	60.83 ± 1.94
	RDOSR+GMM	58.35 ± 2.66	54.28 ± 2.32	14.66 ± 2.14	59.26 ± 2.76
	SSMTL-A	**79.31** ± 0.81	**79.81** ± 0.05	16.21 ± 1.25	**46.27** ± 1.09
	GRDOSR	70.42 ± 0.64	66.77 ± 1.56	16.11 ± 1.10	53.55 ± 6.73
PU	Softmax	93.70 ± 0.67	93.80 ± 0.82	15.99 ± 3.58	83.33 ± 6.07
	OpenMax	92.98 ± 0.84	93.40 ± 0.99	25.35 ± 5.74	73.77 ± 14.20
	AEC	85.78 ± 0.68	86.26 ± 1.11	28.46 ± 2.75	77.11 ± 1.66
	AECD	88.25 ± 0.25	88.09 ± 0.72	**41.79** ± 4.93	56.51 ± 2.92
	RDOSR	93.98 ± 0.60	93.82 ± 0.54	27.43 ± 2.11	62.40 ± 1.69
	RDOSR+GS	93.98 ± 0.60	93.82 ± 0.54	27.56 ± 1.45	58.32 ± 5.64
	RDOSR+GMM	94.29 ± 0.47	94.18 ± 0.64	34.50 ± 3.23	52.91 ± 5.53
	SSMTL-A	**96.44** ± 0.43	**96.51** ± 0.50	30.75 ± 1.91	67.36 ± 2.08
	GRDOSR	96.26 ± 0.34	95.91 ± 0.61	36.71 ± 3.44	**48.98** ± 1.73
PC	Softmax	96.30 ± 0.58	96.24 ± 0.53	41.74 ± 8.95	30.66 ± 12.84
	OpenMax	96.25 ± 0.69	96.10 ± 0.64	41.79 ± 13.78	34.65 ± 10.36
	AEC	93.95 ± 1.59	93.94 ± 1.48	36.90 ± 4.07	56.04 ± 1.13
	AECD	92.50 ± 1.69	92.43 ± 1.55	**49.94** ± 0.66	22.30 ± 2.01
	RDOSR	96.91 ± 0.55	96.93 ± 0.34	45.53 ± 5.59	20.03 ± 3.28
	RDOSR+GS	96.91 ± 0.55	96.93 ± 0.34	45.68 ± 3.49	13.96 ± 1.23
	RDOSR+GMM	96.93 ± 0.65	96.91 ± 0.42	42.36 ± 5.26	17.09 ± 5.14
	SSMTL-A	**98.20** ± 0.76	**98.36** ± 0.50	33.73 ± 1.24	43.33 ± 1.46
	GRDOSR	97.70 ± 0.50	97.78 ± 0.31	48.44 ± 5.41	**13.93** ± 1.52

At the validation-calibrated threshold (Table 3), GRDOSR has a higher open-set macro-F1 than RDOSR on all three datasets. The unknown-class F1 increases from 19.93% to 30.87% on PU, but decreases from 10.37% to 9.18% on IN and from 46.83% to 44.30% on PC. The higher open-set macro-F1s on IN and PC therefore reflect improved known-class recognition alongside lower unknown-class F1s.

Table 3. Classification and detection at the known-validation operating threshold. Each value is a percentage (mean ± sample SD over three runs). OA denotes the known-class overall accuracy after rejection; the open-set macro-F1 treats all unknown samples as one class. Precision, recall, and F1 refer to the unknown class. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best means, respectively. Upward and downward arrows indicate that higher and lower values are better, respectively.

		Classification with Rejection		Unknown-Class Detection			
Scene	**Method**	**Known OA After Rejection ↑**	**Open-Set Macro-F1 ↑**	**Precision ↑**	**Recall ↑**	**F1 ↑**	**Known False Rejection ↓**
IN	Softmax	82.44 ± 1.42	69.46 ± 3.35	11.56 ± 0.92	11.89 ± 0.49	10.32 ± 0.42	6.29 ± 0.82
	OpenMax	79.14 ± 1.08	57.19 ± 1.05	11.84 ± 1.11	**25.19** ± 4.25	**14.50** ± 1.38	6.90 ± 2.41
	AEC	69.29 ± 2.42	54.89 ± 0.19	11.43 ± 1.64	12.78 ± 0.84	10.31 ± 1.98	**4.84** ± 0.72
	AECD	63.95 ± 2.40	41.36 ± 0.86	**13.37** ± 6.45	16.36 ± 5.29	12.46 ± 5.95	5.59 ± 1.01
	RDOSR	70.34 ± 3.81	43.27 ± 2.45	10.40 ± 2.35	13.34 ± 4.76	10.37 ± 2.66	5.99 ± 0.46
	RDOSR+GS	70.02 ± 4.05	41.72 ± 3.28	10.64 ± 2.02	18.16 ± 3.11	11.87 ± 1.87	6.24 ± 0.45
	RDOSR+GMM	74.88 ± 1.05	50.04 ± 1.31	12.89 ± 2.80	16.83 ± 4.01	12.33 ± 2.64	6.21 ± 1.33
	SSMTL-A	**82.89** ± 1.76	**70.88** ± 1.16	11.70 ± 2.10	13.33 ± 0.67	10.78 ± 1.90	5.12 ± 0.20
	GRDOSR	82.01 ± 2.09	59.43 ± 1.52	9.40 ± 3.32	13.52 ± 0.91	9.18 ± 2.57	5.91 ± 0.96
PU	Softmax	91.94 ± 0.89	79.89 ± 0.80	17.03 ± 5.92	8.62 ± 2.60	9.37 ± 2.12	5.13 ± 0.62
	OpenMax	91.04 ± 0.93	79.59 ± 0.78	22.32 ± 5.46	20.75 ± 11.85	16.84 ± 5.17	4.51 ± 1.20
	AEC	84.21 ± 0.59	72.59 ± 1.58	17.81 ± 0.32	20.86 ± 1.55	14.44 ± 1.95	4.79 ± 0.48
	AECD	85.76 ± 1.15	74.12 ± 1.17	24.83 ± 5.59	**35.77** ± 3.29	23.52 ± 4.92	5.28 ± 0.48
	RDOSR	91.59 ± 0.45	76.25 ± 2.50	26.41 ± 1.09	25.30 ± 6.51	19.93 ± 4.37	5.07 ± 0.32
	RDOSR+GS	90.96 ± 0.68	70.72 ± 1.92	23.15 ± 0.84	21.12 ± 1.67	17.17 ± 2.98	5.05 ± 0.64
	RDOSR+GMM	91.95 ± 0.69	81.78 ± 2.21	31.88 ± 3.95	30.16 ± 8.11	25.03 ± 5.87	4.89 ± 0.98
	SSMTL-A	93.03 ± 1.05	76.07 ± 1.99	12.95 ± 2.93	20.03 ± 0.50	13.37 ± 2.17	**4.45** ± 1.01
	GRDOSR	**93.20** ± 0.91	**84.96** ± 1.64	**35.81** ± 5.97	35.75 ± 5.45	**30.87** ± 6.88	5.05 ± 1.26
PC	Softmax	95.13 ± 0.49	79.60 ± 1.81	36.02 ± 10.22	52.08 ± 11.89	39.27 ± 11.49	4.66 ± 0.48
	OpenMax	94.67 ± 0.68	76.53 ± 2.80	29.11 ± 8.33	42.50 ± 15.10	29.98 ± 12.20	4.70 ± 0.58
	AEC	93.60 ± 0.93	77.60 ± 0.99	28.56 ± 2.77	35.50 ± 3.25	27.90 ± 2.60	4.94 ± 1.26
	AECD	93.51 ± 0.50	80.19 ± 0.55	**41.20** ± 3.36	61.51 ± 3.63	**46.99** ± 3.64	4.81 ± 0.95
	RDOSR	94.41 ± 0.84	76.81 ± 1.83	40.89 ± 4.97	**63.47** ± 7.76	46.83 ± 6.26	5.23 ± 0.78
	RDOSR+GS	94.04 ± 0.40	69.71 ± 0.71	38.18 ± 0.21	57.59 ± 3.68	41.89 ± 1.04	5.46 ± 0.40
	RDOSR+GMM	94.48 ± 0.62	82.22 ± 0.78	35.89 ± 2.26	50.58 ± 12.66	38.71 ± 4.98	5.00 ± 0.60
	SSMTL-A	**95.16** ± 1.17	82.56 ± 1.28	27.79 ± 2.78	38.08 ± 2.19	30.14 ± 1.25	**4.48** ± 1.31
	GRDOSR	95.09 ± 0.50	**84.94** ± 0.44	40.30 ± 1.42	57.47 ± 7.72	44.30 ± 3.50	4.69 ± 0.50

Table 4 gives paired differences relative to RDOSR and RDOSR+GS. The PC AUC difference relative to RDOSR is 1.43 percentage points (95% CI: [−5.63, +8.49]). Figures 2–4 show example classification maps, and Figure 5 gives the corresponding ROC and OSCR curves.

Table 4. Paired GRDOSR differences relative to RDOSR and RDOSR+GS. Values are in percentage points; each cell reports the mean difference (GRDOSR minus reference) and a descriptive 95% paired confidence interval ($n = 3$, $df = 2$).

		Mean Difference and 95% Confidence Interval		
Scene	**Reference Method**	**AUC**	**Known OA**	**OSCR**
IN	RDOSR	+7.73 [−3.77, +19.23]	+13.57 [+7.90, +19.25]	+12.68 [+5.39, +19.96]
	RDOSR+GS	+5.59 [−3.30, +14.48]	+13.57 [+7.90, +19.25]	+12.30 [+7.55, +17.06]

Table 4. *Cont.*

Scene	Reference Method	Mean Difference and 95% Confidence Interval		
		AUC	Known OA	OSCR
PU	RDOSR	+6.88 [+4.19, +9.57]	+1.40 [+0.33, +2.47]	+7.54 [+4.64, +10.45]
	RDOSR+GS	+9.79 [+3.76, +15.83]	+1.40 [+0.33, +2.47]	+10.60 [+4.99, +16.22]
PC	RDOSR	+1.43 [−5.63, +8.49]	+0.27 [+0.14, +0.40]	+1.60 [−5.45, +8.65]
	RDOSR+GS	−1.09 [−4.76, +2.58]	+0.27 [+0.14, +0.40]	−0.87 [−4.49, +2.75]
SA	RDOSR	+11.57 [+9.63, +13.50]	+1.72 [+0.57, +2.88]	+12.93 [+11.56, +14.31]
	RDOSR+GS	+8.19 [−0.41, +16.79]	+1.72 [+0.57, +2.88]	+10.33 [+3.16, +17.51]
KSC	RDOSR	+9.00 [−13.70, +31.70]	+6.35 [+0.29, +12.41]	+10.88 [−4.69, +26.44]
	RDOSR+GS	+2.12 [−15.28, +19.52]	+6.35 [+0.29, +12.41]	+5.06 [−5.19, +15.32]

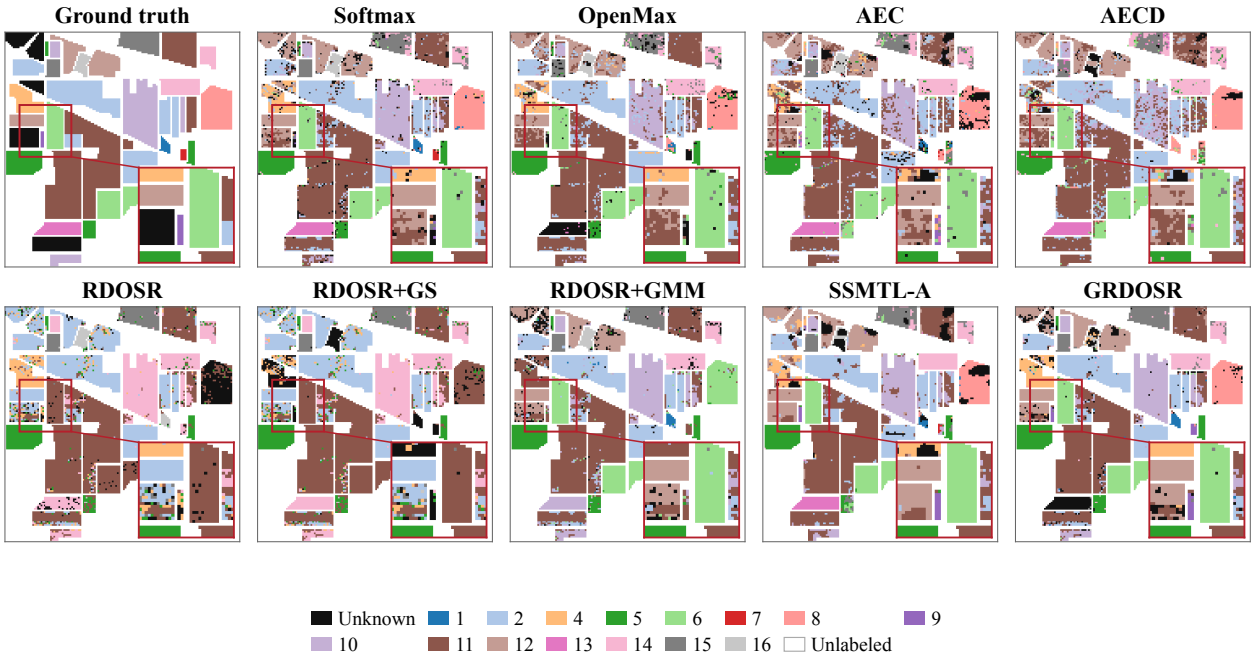


Figure 2. Open-set classification maps on Indian Pines ($L = 3$, seed 3020). Red boxes and insets show the same region for each method; white indicates unannotated pixels.

5.2. Evaluation on Salinas and KSC

Table 5 evaluates the methods on SA and KSC, which were excluded from the method selection. GRDOSR improves the known-class OA over RDOSR from 92.48% to 94.20% on SA and from 78.04% to 84.39% on KSC.

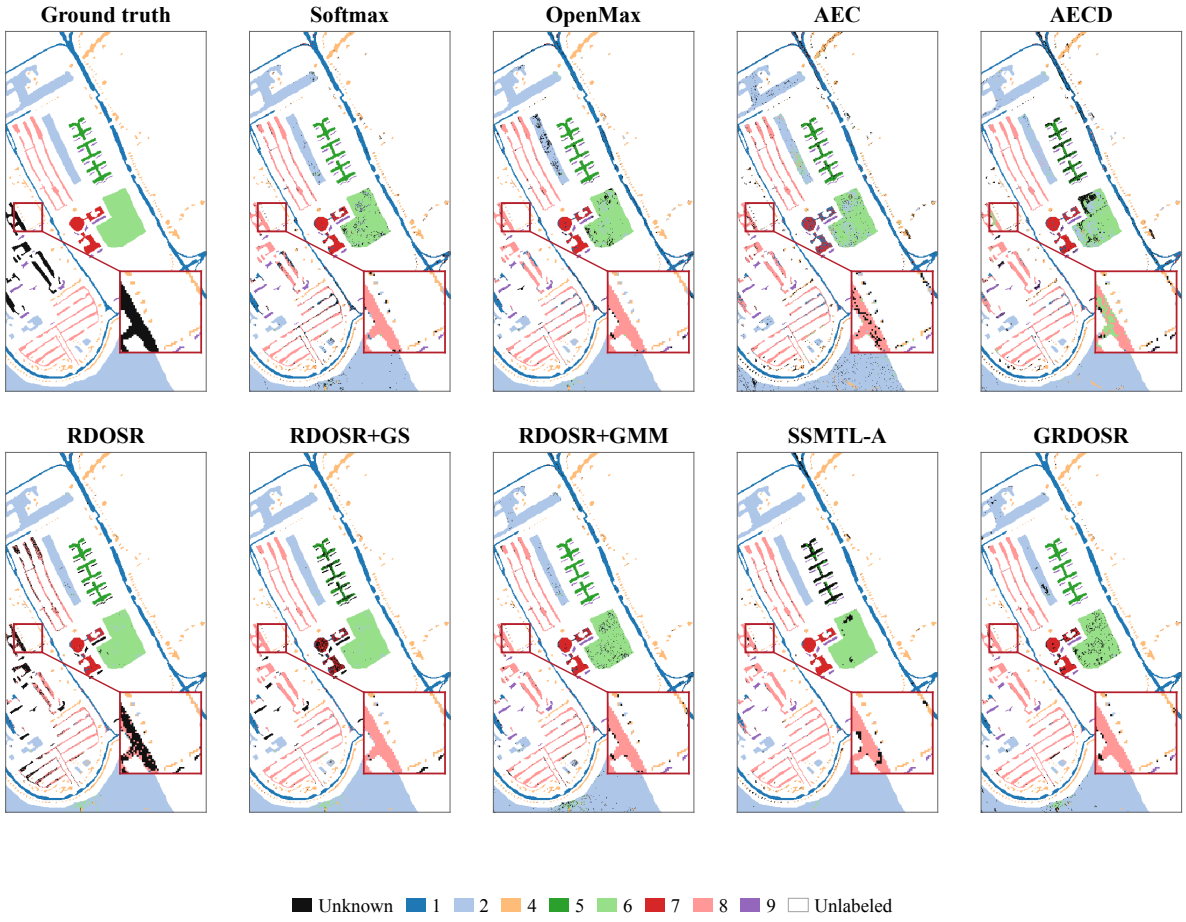


Figure 3. Open-set classification maps on Pavia University ($L = 3$, seed 3020). Red boxes and insets show the same region for each method; white indicates unannotated pixels.

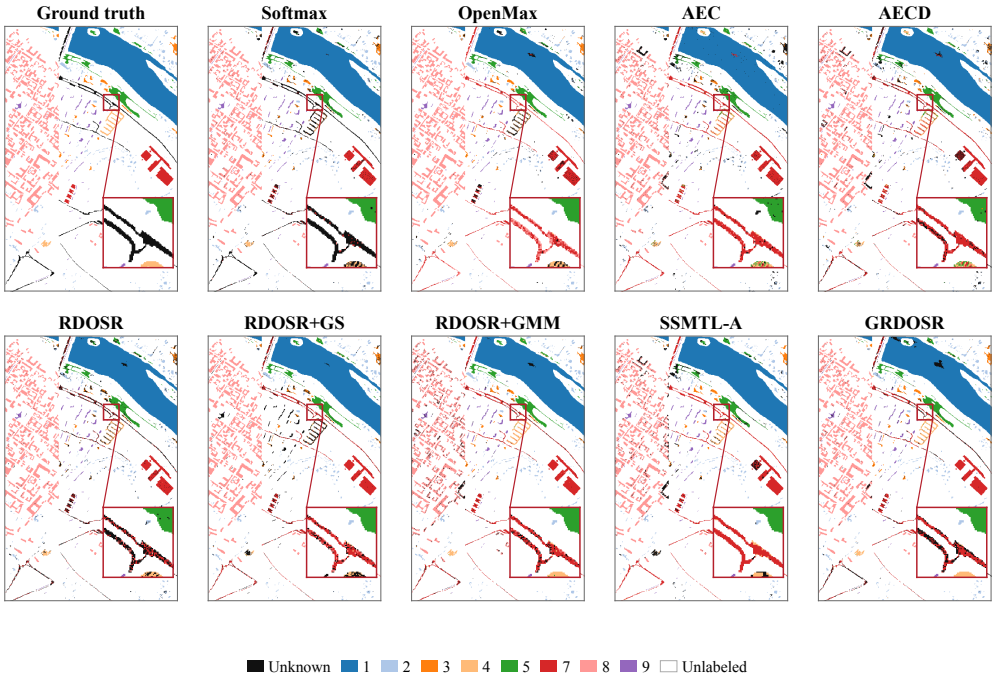


Figure 4. Open-set classification maps on Pavia Center ($L = 6$, seed 3020). Red boxes and insets show the same region for each method; white indicates unannotated pixels.

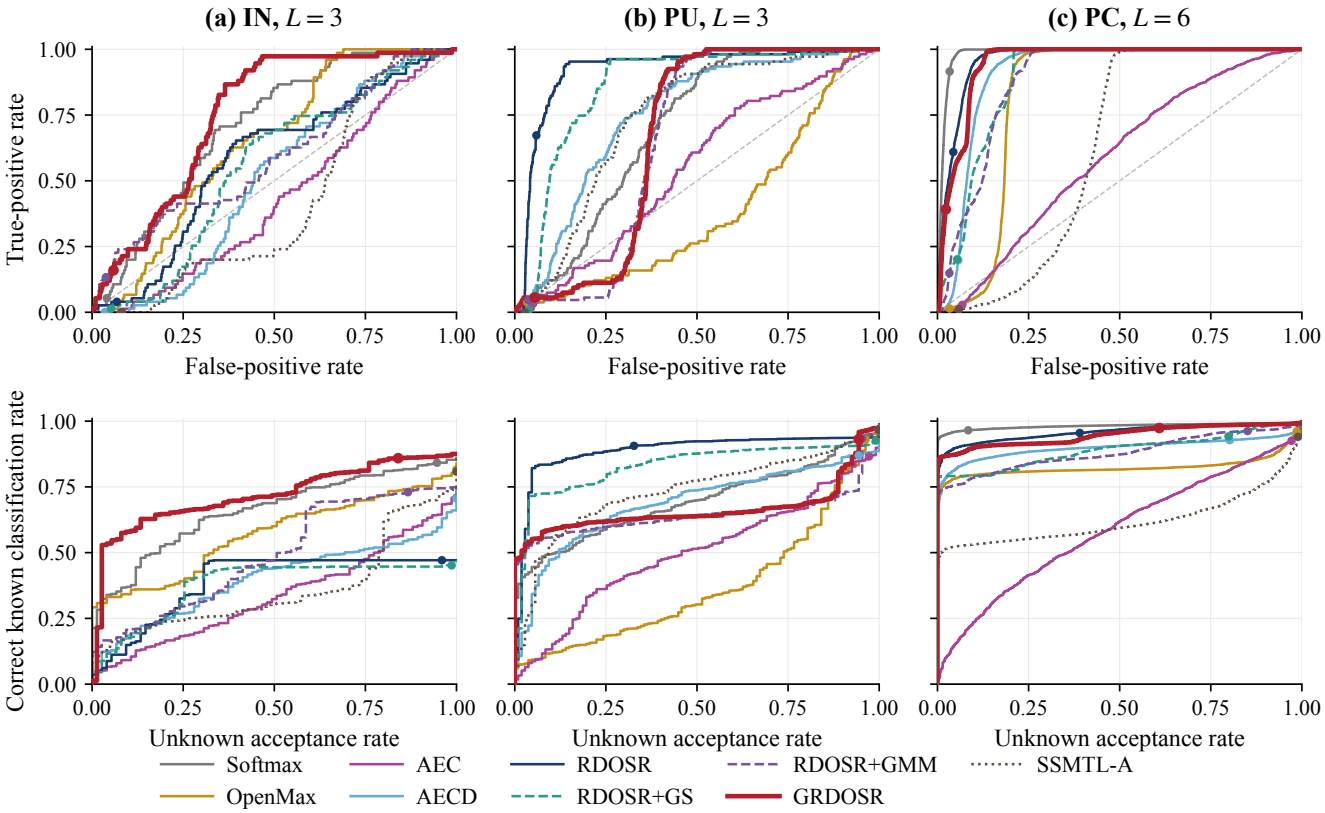


Figure 5. ROC and OSCR curves for the examples in Figures 2–4 (seed 3020). Markers indicate the thresholds estimated from known validation samples.

Table 5. Performances over all single-class holdouts on Salinas (SA) and Kennedy Space Center (KSC). Each value is a percentage (mean ± sample SD over three runs). AUC, OA, AA, and OSCR denote the ROC area, known-class overall accuracy, average per-class accuracy, and open-set classification rate. Macro-F1 is the classwise macro-average; open-set macro-F1 includes one pooled unknown class. AUPR is computed as the average precision; FPR95 is FPR at 95% unknown recall. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best means, respectively. Upward and downward arrows indicate that higher and lower values are better, respectively.

A. Classification and joint performance

Scene	Method	AUC ↑	Known OA Before Rejection ↑	AA ↑	Known Macro-F1 ↑	OSCR ↑	Open-Set Macro-F1 ↑
SA	Softmax	60.95 ± 6.16	92.68 ± 2.23	95.64 ± 1.38	93.95 ± 3.26	59.20 ± 6.53	86.15 ± 2.98
	OpenMax	71.07 ± 9.64	92.11 ± 2.22	94.97 ± 1.53	93.35 ± 3.00	64.91 ± 7.97	82.91 ± 2.35
	AEC	74.00 ± 2.22	88.57 ± 1.12	92.50 ± 0.51	90.69 ± 1.79	64.98 ± 1.37	80.90 ± 1.80
	AECD	82.72 ± 1.93	90.16 ± 2.69	93.12 ± 1.68	90.97 ± 3.39	74.43 ± 3.75	83.88 ± 3.63
	RDOSR	62.40 ± 1.26	92.48 ± 1.75	94.30 ± 2.29	92.39 ± 2.24	57.26 ± 2.41	75.41 ± 4.77
	RDOSR+GS	65.78 ± 1.64	92.48 ± 1.75	94.30 ± 2.29	92.39 ± 2.24	59.86 ± 0.30	75.78 ± 5.28
	RDOSR+GMM	71.48 ± 3.41	92.56 ± 1.72	94.06 ± 1.39	92.06 ± 3.07	66.48 ± 2.21	79.67 ± 4.18
	SSMTL-A	**85.14** ± 1.73	**94.97** ± 1.86	**97.30** ± 0.95	**96.75** ± 1.49	**80.83** ± 2.13	**89.59** ± 1.64
	GRDOSR	73.96 ± 2.00	94.20 ± 1.65	95.50 ± 1.84	93.88 ± 3.48	70.19 ± 2.90	79.36 ± 7.81
KSC	Softmax	66.30 ± 7.45	63.88 ± 22.65	56.74 ± 25.92	51.43 ± 29.36	51.03 ± 17.58	46.37 ± 26.22
	OpenMax	**74.11** ± 6.65	**85.62** ± 2.85	**79.67** ± 4.68	**76.33** ± 6.07	**66.01** ± 6.83	**68.19** ± 4.49
	AEC	63.45 ± 1.26	80.01 ± 4.13	72.51 ± 3.68	68.97 ± 4.85	50.11 ± 1.38	59.95 ± 3.65
	AECD	66.75 ± 4.57	74.05 ± 9.53	62.03 ± 11.85	56.06 ± 12.99	50.35 ± 7.88	50.02 ± 11.80
	RDOSR	63.60 ± 6.38	78.04 ± 2.77	70.36 ± 3.13	64.86 ± 4.14	54.91 ± 6.38	56.48 ± 2.31
	RDOSR+GS	70.48 ± 1.01	78.04 ± 2.77	70.36 ± 3.13	64.86 ± 4.14	60.73 ± 2.92	57.12 ± 3.13
	RDOSR+GMM	71.88 ± 0.92	82.01 ± 0.98	74.32 ± 1.60	69.40 ± 1.90	63.37 ± 1.12	61.80 ± 2.28
	SSMTL-A	71.69 ± 3.02	70.39 ± 13.17	63.33 ± 18.23	58.52 ± 18.36	48.52 ± 9.56	52.43 ± 16.59
	GRDOSR	72.60 ± 6.95	84.39 ± 1.78	78.43 ± 4.71	74.64 ± 4.87	65.79 ± 6.87	64.25 ± 3.14

Table 5. *Cont.*

B. Operating-Point and Detection Metrics

Scene	Method	Known OA After Rejection ↑	Unknown Precision ↑	Unknown Recall ↑	Unknown F1 ↑	AUPR ↑	FPR95 ↓
SA	Softmax	89.93 ± 2.31	10.04 ± 3.96	10.40 ± 3.80	8.88 ± 3.92	12.61 ± 3.47	59.25 ± 5.40
	OpenMax	87.47 ± 1.56	11.19 ± 0.82	20.14 ± 2.48	13.43 ± 0.93	22.40 ± 2.99	44.90 ± 9.94
	AEC	84.45 ± 2.06	14.83 ± 2.87	26.50 ± 5.01	15.71 ± 2.81	22.96 ± 0.99	50.94 ± 0.65
	AECD	85.98 ± 2.93	25.23 ± 4.82	46.42 ± 4.58	29.39 ± 6.68	35.24 ± 1.26	35.55 ± 2.86
	RDOSR	86.39 ± 3.22	4.82 ± 1.38	6.90 ± 3.75	5.09 ± 1.46	13.50 ± 0.55	49.75 ± 2.86
	RDOSR+GS	86.47 ± 2.70	9.64 ± 2.86	18.20 ± 4.99	11.88 ± 3.37	18.64 ± 4.40	49.96 ± 1.00
	RDOSR+GMM	87.89 ± 1.96	11.17 ± 1.53	18.64 ± 1.82	12.12 ± 1.93	20.77 ± 2.54	40.36 ± 5.99
	SSMTL-A	**90.29** ± 2.82	**28.66** ± 5.03	**56.85** ± 4.35	**34.16** ± 4.04	**38.64** ± 7.89	**29.87** ± 4.31
	GRDOSR	88.75 ± 3.14	11.97 ± 3.11	20.03 ± 4.30	13.46 ± 3.85	22.51 ± 4.58	36.37 ± 1.30
KSC	Softmax	62.25 ± 21.76	17.39 ± 4.52	16.63 ± 8.27	15.65 ± 6.36	20.40 ± 4.05	55.07 ± 6.89
	OpenMax	**82.22** ± 4.14	**24.73** ± 3.49	26.10 ± 9.91	22.79 ± 5.21	30.50 ± 5.19	48.58 ± 13.85
	AEC	73.70 ± 1.31	13.89 ± 3.84	17.73 ± 5.73	14.51 ± 4.17	22.40 ± 1.94	58.14 ± 2.36
	AECD	68.71 ± 8.83	19.23 ± 7.09	23.75 ± 7.07	20.09 ± 6.42	26.02 ± 4.55	59.07 ± 5.82
	RDOSR	75.60 ± 2.16	14.65 ± 5.23	11.90 ± 2.85	11.50 ± 3.63	20.20 ± 5.62	58.10 ± 6.32
	RDOSR+GS	75.64 ± 2.08	19.58 ± 0.31	16.70 ± 5.26	15.87 ± 2.70	24.40 ± 1.65	56.92 ± 3.68
	RDOSR+GMM	78.06 ± 1.19	18.11 ± 1.55	20.57 ± 8.76	17.39 ± 4.96	25.98 ± 2.08	49.81 ± 1.53
	SSMTL-A	65.79 ± 15.10	24.07 ± 2.22	**26.43** ± 12.12	**22.89** ± 8.12	**32.21** ± 2.35	46.59 ± 1.67
	GRDOSR	79.29 ± 2.32	14.13 ± 2.55	16.08 ± 5.98	14.14 ± 3.81	23.48 ± 0.67	**43.43** ± 8.88

The mean AUC increases from 62.40% to 73.96% on SA and from 63.60% to 72.60% on KSC. SSMTL-A achieves the highest mean AUC and OSCR on SA, and OpenMax on KSC. At the calibrated threshold, the GRDOSR unknown-class F1 is 13.46% on SA and 14.14% on KSC.

5.3. Component and Sensitivity Analyses

Tables 6 and 7 display the contributions of attention and expert grouping. The IN, PU, and PC results provide a comparison between attention branches, local averaging, uniform weighting, and random training groups under the settings described in Section 4.2.

Table 6. Ablation results for IN and PU ($L = 3$) and PC ($L = 6$). Each value is a percentage (mean ± sample SD over three runs). SOA denotes the second-order attention; uniform SOA uses fixed, uniform spatial and channel weights. G is the expert count, AUC denotes the ROC area, and known OA denotes the accuracy before rejection. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best means, respectively. Upward arrows indicate that higher values are better.

	IN		PU		PC	
Configuration	**AUC ↑**	**Known OA ↑**	**AUC ↑**	**Known OA ↑**	**AUC ↑**	**Known OA ↑**
RDOSR+GS, $G = 1$	49.90 ± 4.81	66.15 ± 16.64	78.92 ± 8.28	96.41 ± 0.51	91.57 ± 2.72	99.14 ± 0.16
RDOSR+GMM, $G = 3$	62.33 ± 8.16	76.80 ± 2.63	72.60 ± 9.12	96.53 ± 0.48	89.29 ± 1.50	99.16 ± 0.15
Spatial SOA, $G = 1$	51.47 ± 4.52	70.41 ± 20.46	**80.84** ± 5.95	97.26 ± 1.15	93.16 ± 1.01	99.54 ± 0.13
Spatial SOA, $G = 3$	58.71 ± 11.13	80.79 ± 5.17	63.47 ± 19.37	97.36 ± 1.01	92.79 ± 2.25	99.49 ± 0.12
Channel SOA, $G = 1$	67.31 ± 8.24	87.95 ± 1.11	74.58 ± 7.38	97.21 ± 0.61	90.88 ± 2.87	99.28 ± 0.11
Channel SOA, $G = 3$	67.03 ± 5.32	84.46 ± 5.32	60.94 ± 28.31	97.20 ± 0.82	93.04 ± 1.26	99.26 ± 0.14
Dual SOA, $G = 1$	64.82 ± 16.18	88.30 ± 2.38	75.59 ± 6.43	97.66 ± 0.75	92.89 ± 0.72	99.43 ± 0.12
Dual SOA, $G = 3$	**70.76** ± 18.47	85.78 ± 3.58	64.85 ± 11.09	97.72 ± 0.63	93.20 ± 2.73	99.38 ± 0.20
Mean pooling, $G = 1$	62.65 ± 4.35	**91.52** ± 1.54	79.59 ± 4.20	98.97 ± 0.52	92.36 ± 1.58	99.02 ± 1.15
Mean pooling, $G = 3$	70.68 ± 5.17	86.62 ± 5.28	65.83 ± 24.46	**98.98** ± 0.31	90.82 ± 3.05	**99.57** ± 0.07
Uniform SOA, $G = 1$	50.12 ± 11.73	83.56 ± 6.41	80.40 ± 5.24	97.70 ± 0.52	92.69 ± 1.17	99.44 ± 0.13
Uniform SOA, $G = 3$	68.34 ± 8.60	83.44 ± 1.26	58.12 ± 13.11	97.72 ± 0.44	92.31 ± 1.48	99.39 ± 0.11
Random training groups Fixed posterior, $G = 3$	62.95 ± 9.70	82.41 ± 6.04	68.97 ± 10.36	97.52 ± 0.58	**94.11** ± 1.92	99.39 ± 0.17

At $G = 3$, learned dual attention increases the mean AUC over the uniform SOA by 2.42, 6.73, and 0.90 percentage points on IN, PU, and PC. Learned weighting thus improves detection relative to uniform attention in these examples, while mean pooling gives a higher known-class OA of 86.62%, 98.98%, and 99.57%, respectively.

The SA and KSC controls (Table 7) show AUC gains of 9.97 and 6.82 percentage points over the uniform SOA at $G = 3$. The dual-attention model has a known-class OAs of 94.00% and 89.10%, compared with 95.16% and 87.21% for mean pooling. These results separate the effects of learned weighting on detection and classification.

Table 7. Ablation results for SA and KSC ($L = 3$). Each value is a percentage (mean ± sample SD over three runs). SOA denotes second-order attention; uniform SOA uses fixed, uniform spatial and channel weights. G is the expert count, AUC denotes the ROC area, and known OA denotes the accuracy before rejection. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best means, respectively. Upward arrows indicate that higher values are better.

	SA		KSC	
Configuration	**AUC ↑**	**Known OA ↑**	**AUC ↑**	**Known OA ↑**
RDOSR+GS, $G = 1$	60.72 ± 8.54	92.68 ± 2.83	76.87 ± 12.73	83.99 ± 4.10
RDOSR+GMM, $G = 3$	59.46 ± 11.65	92.77 ± 3.45	72.19 ± 17.77	86.03 ± 7.79
Spatial SOA, $G = 1$	61.99 ± 2.11	93.34 ± 2.87	71.90 ± 9.04	84.38 ± 4.41
Spatial SOA, $G = 3$	70.76 ± 11.42	93.32 ± 3.04	71.81 ± 6.05	86.13 ± 7.83
Channel SOA, $G = 1$	61.11 ± 7.51	93.89 ± 3.00	71.55 ± 13.26	88.53 ± 2.10
Channel SOA, $G = 3$	76.37 ± 3.20	93.81 ± 2.79	**80.33** ± 16.46	**89.93** ± 3.78
Dual SOA, $G = 1$	66.72 ± 12.55	94.00 ± 3.10	70.01 ± 18.95	86.12 ± 6.81
Dual SOA, $G = 3$	75.53 ± 9.90	94.00 ± 2.98	74.91 ± 12.58	89.10 ± 3.05
Mean pooling, $G = 1$	64.12 ± 3.18	94.16 ± 4.43	72.13 ± 9.06	77.96 ± 18.18
Mean pooling, $G = 3$	**86.48** ± 6.01	**95.16** ± 3.31	71.47 ± 22.67	87.21 ± 2.07
Uniform SOA, $G = 1$	54.37 ± 15.06	93.86 ± 2.88	62.47 ± 17.54	85.38 ± 4.48
Uniform SOA, $G = 3$	65.56 ± 32.58	93.73 ± 2.62	68.09 ± 15.29	83.02 ± 3.99
Random training groups Fixed posterior, $G = 3$	71.45 ± 11.61	94.03 ± 3.29	64.22 ± 13.61	84.42 ± 1.86

The effect of expert count varies by dataset (Table 8). The highest AUC in the illustrated conditions occurs at $G = 4$ on IN, $G = 1$ on PU, $G = 2$ on PC, and $G = 3$ on SA and KSC. The main comparisons retain $G = 3$ on every dataset.

Table 8. Sensitivity to expert count with $L = 3$ ($L = 6$ for PC). Each value is a percentage (mean ± sample SD over three runs). G is the expert count, AUC is the ROC area, known OA is the accuracy before rejection, and OSCR is the open-set classification rate. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best means, respectively. Upward arrows indicate that higher values are better.

Scene	**Configuration**	**AUC ↑**	**Known OA ↑**	**OSCR ↑**
IN	$G = 1$	64.82 ± 16.18	**88.30** ± 2.38	60.01 ± 14.07
	$G = 2$	71.62 ± 13.16	88.07 ± 0.77	67.19 ± 11.58
	$G = 3$	70.76 ± 18.47	85.78 ± 3.58	66.23 ± 16.64
	$G = 4$	**74.09** ± 12.40	86.14 ± 2.58	**68.05** ± 11.37
PU	$G = 1$	**75.59** ± 6.43	97.66 ± 0.75	**75.28** ± 6.19
	$G = 2$	73.67 ± 1.18	97.68 ± 0.63	73.54 ± 1.16
	$G = 3$	64.85 ± 11.09	**97.72** ± 0.63	64.72 ± 11.02
	$G = 4$	61.39 ± 14.28	97.58 ± 1.04	61.26 ± 14.23

Table 8. *Cont.*

Scene	Configuration	AUC ↑	Known OA ↑	OSCR ↑
PC	$G = 1$	92.89 ± 0.72	**99.43** ± 0.12	92.68 ± 0.71
	$G = 2$	**95.86** ± 0.94	99.41 ± 0.13	**95.51** ± 0.89
	$G = 3$	93.20 ± 2.73	99.38 ± 0.20	92.91 ± 2.69
	$G = 4$	93.13 ± 5.32	99.40 ± 0.13	92.86 ± 5.25
SA	$G = 1$	66.72 ± 12.55	**94.00** ± 3.10	61.83 ± 8.86
	$G = 2$	75.42 ± 18.79	90.49 ± 4.93	69.69 ± 13.27
	$G = 3$	**75.53** ± 9.90	**94.00** ± 2.98	**70.84** ± 6.41
	$G = 4$	72.46 ± 31.11	93.99 ± 2.55	68.26 ± 27.76
KSC	$G = 1$	70.01 ± 18.95	86.12 ± 6.81	67.66 ± 18.96
	$G = 2$	66.91 ± 7.86	87.41 ± 4.88	63.31 ± 7.60
	$G = 3$	**74.91** ± 12.58	**89.10** ± 3.05	**71.29** ± 10.60
	$G = 4$	71.07 ± 20.29	87.36 ± 5.51	64.96 ± 12.12

With two classes held out (Table 9), GRDOSR has higher mean AUCs than RDOSR on IN, PC, SA, and KSC. The largest increases occur on SA (68.67% to 91.47%) and KSC (58.80% to 83.51%). On PU, AUC decreases from 91.33% to 87.26%, indicating that the gain depends on the composition of the unknown classes.

Table 9. Evaluation with two unknown classes on each scene. Each value is a percentage (mean ± sample SD over three runs). AUC is the ROC area, known OA is the accuracy before rejection, and OSCR is the open-set classification rate. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best means, respectively. Upward arrows indicate that higher values are better.

Scene	Method	AUC ↑	Known OA ↑	OSCR ↑
IN	RDOSR	57.00 ± 4.74	81.54 ± 9.65	51.77 ± 3.64
	GRDOSR	**71.30** ± 11.02	**81.94** ± 11.45	**64.83** ± 13.26
PU	RDOSR	**91.33** ± 5.72	95.63 ± 1.92	**89.02** ± 5.21
	GRDOSR	87.26 ± 12.50	**97.78** ± 0.22	86.01 ± 11.82
PC	RDOSR	87.49 ± 4.66	99.72 ± 0.15	87.41 ± 4.63
	GRDOSR	**88.51** ± 3.91	**99.82** ± 0.09	**88.47** ± 3.92
SA	RDOSR	68.67 ± 27.46	90.07 ± 4.98	61.74 ± 21.50
	GRDOSR	**91.47** ± 7.63	**92.81** ± 5.28	**85.40** ± 7.44
KSC	RDOSR	58.80 ± 26.16	75.08 ± 1.09	53.96 ± 23.18
	GRDOSR	**83.51** ± 3.30	**82.76** ± 10.04	**70.34** ± 9.20

5.4. Scene Inference and Expert Aggregation

Whole-scene inference uses the spectral cube and fitted model and calibration parameters. It reproduces all 43,424 saved test-center predictions (Table 10).

Table 10. Whole-scene inference from the spectral cube and saved model parameters. Mismatches are differences from the saved predictions at the same test centers.

	Label-Free Scene Inference		
Scene	Scene Pixels	Test Centers	Mismatches
IN	21,025	941	0
PU	207,400	6640	0
PC	783,640	26,146	0
SA	111,104	8991	0
KSC	314,368	706	0

Table 11 compares expert aggregation using the same fitted networks. The last two columns summarize the component posteriors, with $q_{\max}(x) = \max_g q_g(x)$ and normalized entropy $H_q(x) = -\sum_g q_g(x) \log q_g(x) / \log G$, where $0 \log 0 = 0$.

Table 11. Expert aggregation using the same fitted networks ($L = 3$ on IN, PU, SA, and KSC; $L = 6$ on PC). Performance is reported as a percentage (mean ± sample SD over three runs). The first three rules use reconstruction scores; GRDOSR uses maximum fusion with density. The last two columns give the fraction of test posteriors with $q_{\max} \geq 0.99$ and mean normalized entropy, averaged over the same runs. **Red bold**, dark blue, and *light blue* indicate the best, second-best, and third-best performance means. Upward arrows indicate that higher values are better.

		Open-Set Performance			GMM Posteriors	
Scene	**Aggregation Rule**	**AUC ↑**	**Known OA ↑**	**OSCR ↑**	$q_{\max} \geq 0.99$ (%)	**Mean H_q**
IN	Hard GMM routing	69.28 ± 16.65	**85.78** ± 3.58	64.72 ± 14.40	100.000	2.18×10^{-8}
	Equal expert averaging	28.89 ± 1.23	54.92 ± 15.39	13.75 ± 2.39		
	Posterior aggregation	69.28 ± 16.65	**85.78** ± 3.58	64.72 ± 14.40		
	GRDOSR	**70.76** ± 18.47	**85.78** ± 3.58	**66.23** ± 16.64		
PU	Hard GMM routing	52.88 ± 7.07	**97.72** ± 0.63	52.75 ± 7.11	99.925	1.86×10^{-4}
	Equal expert averaging	57.95 ± 37.09	66.81 ± 25.18	39.11 ± 25.17		
	Posterior aggregation	52.88 ± 7.07	**97.72** ± 0.63	52.74 ± 7.11		
	GRDOSR	**64.85** ± 11.09	**97.72** ± 0.63	**64.72** ± 11.02		
PC	Hard GMM routing	75.91 ± 5.50	**99.38** ± 0.20	75.64 ± 5.42	99.882	3.94×10^{-4}
	Equal expert averaging	36.63 ± 23.59	73.50 ± 30.72	30.98 ± 24.49		
	Posterior aggregation	75.91 ± 5.50	**99.38** ± 0.20	75.63 ± 5.43		
	GRDOSR	**93.20** ± 2.73	**99.38** ± 0.20	**92.91** ± 2.69		
SA	Hard GMM routing	65.90 ± 8.04	**94.00** ± 2.98	62.67 ± 4.79	99.997	8.10×10^{-6}
	Equal expert averaging	42.08 ± 31.80	51.16 ± 20.43	25.70 ± 19.32		
	Posterior aggregation	65.90 ± 8.04	**94.00** ± 2.98	62.67 ± 4.79		
	GRDOSR	**75.53** ± 9.90	**94.00** ± 2.98	**70.84** ± 6.41		
KSC	Hard GMM routing	70.84 ± 11.43	**89.10** ± 3.05	68.08 ± 11.45	100.000	$<10^{-12}$
	Equal expert averaging	27.53 ± 17.51	41.04 ± 21.77	10.59 ± 4.54		
	Posterior aggregation	70.84 ± 11.43	**89.10** ± 3.05	68.08 ± 11.45		
	GRDOSR	**74.91** ± 12.58	**89.10** ± 3.05	**71.29** ± 10.60		

Concentrated component posteriors make weighted aggregation approach hard routing. Equal averaging reduces the known-class accuracy, supporting weights that depend on the input.

5.5. Reconstruction and Distributional Support

Figure 6 summarizes the gains of GRDOSR over RDOSR in the mean AUC, known-class OA, and OSCR across the five datasets. Figure 7 provides the classwise view: IN has more confusion between known classes than PU and PC. Component-specific experts learn from different class compositions and regions of the feature distribution [28]; posterior weighting combines these specialized predictions [30].

Reconstruction error measures consistency with the expert mappings, while mixture likelihood measures support under the known feature distribution [16,17]. The maximum rule in Equation (8) allows either signal to trigger rejection. Reconstruction can complement density when an unknown feature has plausible mixture likelihood but is poorly reconstructed by the associated experts.

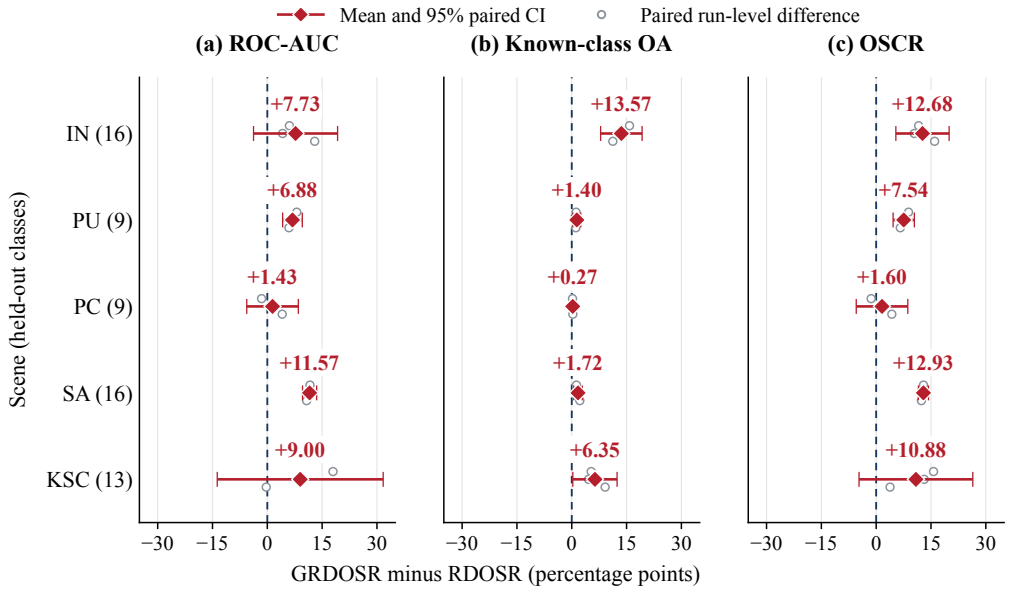


Figure 6. GRDOSR gains over RDOSR in the ROC-AUC, known-class OA before rejection, and OSCR. Red diamonds denote mean paired differences, red bars the descriptive 95% paired confidence intervals, and gray circles the individual run differences ($n = 3$, $df = 2$). Each run averages all single-class holdouts; parentheses give their number. Values are percentage points, and dashed lines mark zero differences.

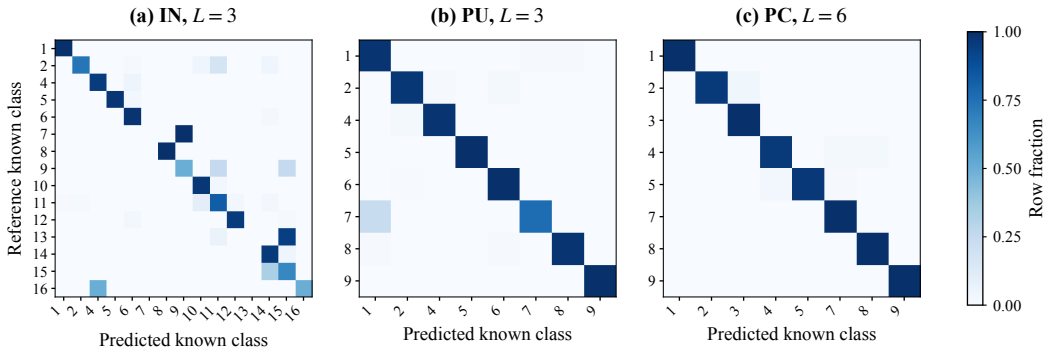


Figure 7. Row-normalized known-class confusion matrices for GRDOSR before rejection, using the examples in Figures 2–4 (seed 3020). Rows denote true classes and columns predicted classes; entries are fractions.

Across all single-class holdouts, density-only scoring gives a higher mean AUC than maximum fusion on IN (75.38% versus 72.85%), PC (95.26% versus 92.99%), SA (80.15% versus 73.96%), and KSC (77.92% versus 72.60%). On PU, maximum fusion improves the mean AUC from 75.02% to 76.12%. The benefit also varies with the held-out class: fusion increases the mean AUC over density alone by 11.81 percentage points for PU with $L = 6$ and by 15.94 points for KSC with $L = 13$. When reconstruction provides weaker separation, taking the maximum can raise known-sample scores and reduce the separation provided by density alone.

Figure 8 illustrates recognition on SA and KSC, complementing their dataset-level results in Table 5.

The experiments evaluate spatial generalization within each scene, and three repetitions give limited precision for comparative estimates. Rare known classes have no eligible validation centers in 38 of 51 IN experiments and 27 of 42 KSC experiments, reducing the evidence available for calibration. Future work should examine geographic transfer, more diverse combinations of unknown classes, and the computational cost of multiple experts.

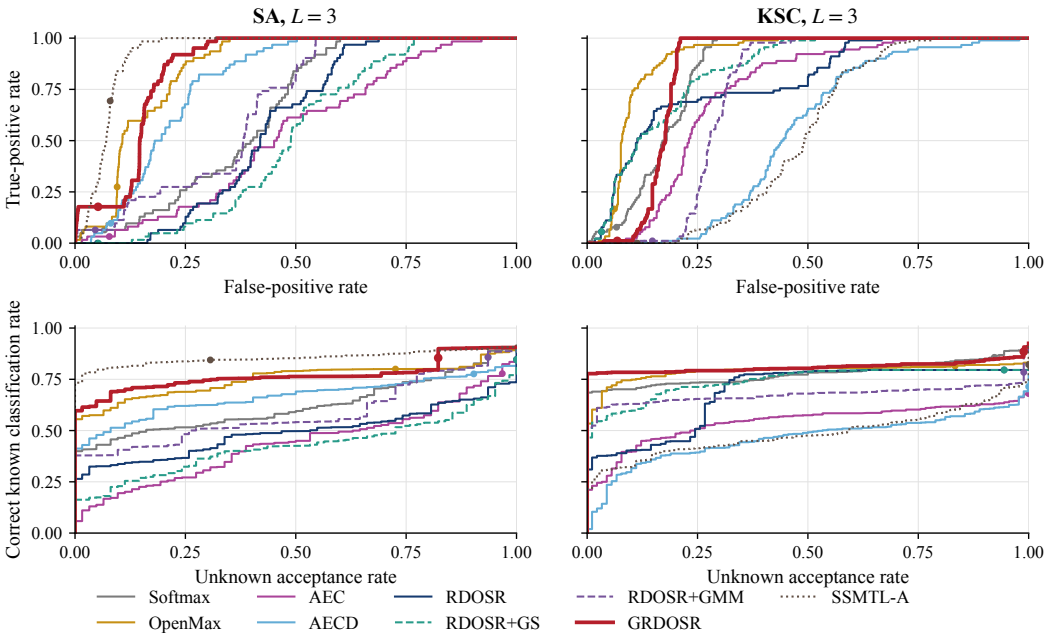


Figure 8. ROC and OSCR curves on SA and KSC ($L = 3$, seed 3020). Markers indicate the thresholds estimated from known validation samples.

6. Conclusions

GRDOSR combines second-order attention and Gaussian mixture modeling for hyperspectral open-set classification. Across five benchmarks, it improves the mean AUC, known-class OA, and OSCR over RDOSR, with the highest mean AUC and OSCR among the compared methods on IN and PU. Posterior concentration explains the agreement between soft and hard routing, while the effect of score fusion depends on the unknown-class composition. The framework supports land-cover mapping when some classes are absent from the training data.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app162110468/s1, `GRDOSR_method_reproducibility`: source code and reproducibility resources.

Author Contributions: Methodology, C.H. and R.X.; software, C.H.; investigation, C.H.; writing—original draft preparation, C.H.; writing—review and editing, R.X. and M.C.; supervision, M.C.; funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Key Research and Development Project of the Jilin Provincial Science and Technology Development Plan, grant number 20210201079GX.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The benchmark data are available from the Remote Sensing Laboratory, University of Tehran, at https://rslab.ut.ac.ir/data (accessed on 18 October 2026).

Acknowledgments: OpenAI Codex (OpenAI, `gpt-6.1-sol`) was used for auxiliary technical and editorial support. The authors reviewed the outputs and take full responsibility for the content.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in study design, data analysis, manuscript preparation, or the decision to publish.

References

1. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep learning for hyperspectral image classification: An overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
2. Scheirer, W.J.; de Rezende Rocha, A.; Sapkota, A.; Boult, T.E. Toward open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1757–1772. [CrossRef] [PubMed]
3. Scheirer, W.J.; Jain, L.P.; Boult, T.E. Probability models for open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2317–2324. [CrossRef] [PubMed]
4. Al Rahhal, M.M.; Bazi, Y.; Al-Dayil, R.; Alwadei, B.M.; Ammour, N.; Alajlan, N. Energy-based learning for open-set classification in remote sensing imagery. *Int. J. Remote Sens.* **2022**, *43*, 6027–6037. [CrossRef]
5. Kaviani Baghbaderani, R.; Qu, Y.; Qi, H.; Stutts, C. Representative-discriminative learning for open-set land cover classification of satellite imagery. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12375, pp. 1–17. [CrossRef]
6. Mendes Júnior, P.R.; de Souza, R.M.; Werneck, R.d.O.; Stein, B.V.; Pazinato, D.V.; de Almeida, W.R.; Penatti, O.A.B.; Torres, R.d.S.; Rocha, A. Nearest neighbors distance ratio open-set classifier. *Mach. Learn.* **2017**, *106*, 359–386. [CrossRef]
7. Zhang, H.; Patel, V.M. Sparse representation-based open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1690–1696. [CrossRef] [PubMed]
8. Bendale, A.; Boult, T.E. Towards open set deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1563–1572. [CrossRef]
9. Ge, Z.; Demyanov, S.; Chen, Z.; Garnavi, R. Generative OpenMax for multi-class open set classification. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 4–7 September 2017; pp. 42.1–42.12. [CrossRef]
10. Shu, L.; Xu, H.; Liu, B. DOC: Deep open classification of text documents. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2911–2916. [CrossRef]
11. Huang, H.; Wang, Y.; Hu, Q.; Cheng, M.M. Class-specific semantic reconstruction for open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4214–4228. [CrossRef] [PubMed]
12. Oza, P.; Patel, V.M. Deep CNN-based multi-task learning for open-set recognition. *arXiv* **2019**, arXiv:1903.03161. [CrossRef]
13. Yoshihashi, R.; Shao, W.; Kawakami, R.; You, S.; Iida, M.; Naemura, T. Classification-reconstruction learning for open-set recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4011–4020. [CrossRef]
14. Oza, P.; Patel, V.M. C2AE: Class conditioned auto-encoder for open-set recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2302–2311. [CrossRef]
15. Perera, P.; Morariu, V.I.; Jain, R.; Manjunatha, V.; Wigington, C.; Ordonez, V.; Patel, V.M. Generative-discriminative feature representations for open-set recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11811–11820. [CrossRef]
16. Sun, X.; Yang, Z.; Zhang, C.; Ling, K.V.; Peng, G. Conditional Gaussian distribution learning for open set recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13477–13486. [CrossRef]
17. Sun, J.; Wang, H.; Dong, Q. MoEP-AE: Autoencoding mixtures of exponential power distributions for open-set recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 312–325. [CrossRef]
18. Xie, Z.; Duan, P.; Liu, W.; Kang, X.; Wei, X.; Li, S. Feature consistency-based prototype network for open-set hyperspectral image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 9286–9296. [CrossRef] [PubMed]
19. Du, Y.; Li, X.; Shi, L.; Li, F.; Xu, T. A prototype network for hyperspectral image open-set classification based on feature invariance and weighted Pearson distance measurement. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef]
20. Zhou, H.; Wu, H.; Wang, A.; Iwahori, Y.; Yu, X. Incorporating attention mechanism, dense connection blocks, and multi-scale reconstruction networks for open-set hyperspectral image classification. *Remote Sens.* **2023**, *15*, 4535. [CrossRef]
21. Xi, B.; Zhang, W.; Li, J.; Song, R.; Li, Y. HyperCASR: Spectral-spatial open-set recognition with category-aware semantic reconstruction for hyperspectral imagery. *IEEE Trans. Image Process.* **2025**, *34*, 7642–7655. [CrossRef] [PubMed]
22. Mu, C.; Liu, Y.; Yan, X.; Ali, A.; Liu, Y. Few-shot open-set hyperspectral image classification with adaptive threshold using self-supervised multitask learning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5526618. [CrossRef]
23. Zhao, H.; Wang, X.; Zheng, Z.; Li, J.; Zhong, Y. HOpenCls: Training hyperspectral image open-set classifiers in their living environments. *arXiv* **2025**, arXiv:2502.15163. [CrossRef]
24. Wang, H.; Qiao, Z.; Zhang, W.; Liu, X. Open set cross-domain hyperspectral image classification based on critical reflective learning network. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5521114. [CrossRef]
25. Hang, R.; Li, Z.; Liu, Q.; Ghamisi, P.; Bhattacharyya, S.S. Hyperspectral image classification with attention-aided CNNs. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 2281–2293. [CrossRef]

26. Xue, Z.; Zhang, M.; Liu, Y.; Du, P. Attention-based second-order pooling network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9600–9615. [CrossRef]
27. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3019–3028. [CrossRef]
28. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
29. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–22. [CrossRef]
30. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
31. Grupo de Inteligencia Computacional. Hyperspectral Remote Sensing Scenes. Available online: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 7 October 2026).
32. Liang, J.; Zhou, J.; Qian, Y.; Wen, L.; Bai, X.; Gao, Y. On the sampling strategy for evaluation of spectral-spatial methods in hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 862–880. [CrossRef]
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
34. Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
35. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
36. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
37. Dhamija, A.R.; Günther, M.; Boult, T.E. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.