LEARNING DISCRETE SEMANTIC INVARIANCE FOR CROSS-SCENE HYPERSPECTRAL IMAGE DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-scene hyperspectral image (HSI) classification requires a model trained on one labeled scene to recognize shared classes in an unseen scene without using target data. Existing domain-generalization methods often synthesize source variations and enforce representation consistency across transformed views. However, a transformation that suppresses scene-specific variation for one material may also alter a narrow spectral cue that distinguishes another class. Uniformly enforcing consistency across these transformations can therefore remove class-relevant information. In this work, we propose a transformation-conditioned discrete self-supervision (TCDS) framework for cross-scene hyperspectral image domain generalization, which learns shared discrete semantics by weighting transformed views according to their semantic preservation. For each source sample, our TCDS constructs plausible virtual scene variations and estimates how much class-relevant evidence each transformation retains. Spectral and spatial-context views are then encoded with separate discrete codebooks whose semantic indices are aligned before forming a preservation-weighted consensus. A complementary residual branch retains class-discriminative information outside the shared semantic codes and contributes to prediction through a preservation-dependent gate. Extensive experiments on three datasets demonstrate the effectiveness and superiority of our proposed TCDS method. The code file is public at https://github.com/Sky-byte-box/TCDS_code.

1 INTRODUCTION

Cross-scene hyperspectral image (HSI) domain generalization studies whether a classifier learned from a labeled source scene can recognize the same land-cover classes in a previously unseen scene without using target data during training. This setting is challenging because hyperspectral observations are sensitive to acquisition conditions. Atmospheric effects, illumination, seasonal differences, sensor response, and spatial resolution can alter spectral signatures and local spatial patterns even when the underlying class semantics are unchanged (Ghamisi et al., 2017; Li et al., 2019). A model fitted to one scene may therefore depend on spectral–spatial regularities that do not persist elsewhere, leading to degraded recognition after transfer. The central goal is to learn representations that retain class-discriminative semantics while reducing dependence on scene-specific variation.

Cross-scene HSI recognition was initially studied mainly through domain adaptation, where unlabeled target samples are available to guide source–target alignment (Zhang et al., 2021; Tang et al., 2022; Zhao et al., 2022; Li et al., 2024). This assumption, however, is unavailable in domain generalization, where the model must be learned only from source observations and transferred directly to an unseen scene. To compensate for the absence of target data, a common strategy is to enlarge the effective source distribution by constructing plausible scene variations. The single-source domain expansion network (SDEnet) introduces spectral–spatial randomization (Zhang et al., 2023), while nonlinear sample generation (NSG) and the spectral–spatial adversarial multidomain synthesis network (S^2AMSNet) generate nonlinear samples or synthetic domains (Wang et al., 2024a; Chen et al., 2024), and the two-stage domain alignment (TSDA) network aligns source-derived domains in two stages (Wang et al., 2024b). Fourier-based spectral–spatial generation offers another way to diversify source-derived scenes while accounting for spatial context (Shi et al., 2025). Frequency-

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

oriented approaches model scene variation from another perspective. The frequency disentanglement and data geometry network (FDGNet) combines frequency disentanglement with data geometry (Qin et al., 2025), whereas the single-source frequency transform (SFT) applies frequency transformations to the source data (Huang et al., 2025). Other studies reduce source-specific shifts directly or simulate structured spectral and acquisition changes that may arise across scenes (Qi et al., 2025; Chen et al., 2026a; Hu et al., 2026). Although implemented differently, these data-centric methods expose the classifier to diverse source-side variations that may occur after cross-scene transfer.

A complementary line of work focuses on representations that retain class-related information across scene changes. The central idea is to reduce sensitivity to scene-dependent spectral–spatial variation while preserving features that remain discriminative for the same land-cover classes under different acquisition conditions. Dong et al. (2024) proposed the spectral–spatial enhancement and causal constraint network (S^2ECNet), which combines spectral–spatial enhancement with causal constraints to reduce the influence of scene-dependent factors. Peng et al. (2024) developed a disentanglement-inspired single-source domain-generalization network to separate transferable information from domain-related variation. Zhao et al. (2025) introduced adversarial decoupling for cross-scene HSI classification, while Chen et al. (2026b) incorporated spectral relations and causal learning to model information that can persist across scenes. Further, Chen et al. (2025) decoupled and aligned modality-shared semantics, while Wang et al. (2026) used a spectrum information bottleneck to disentangle invariant from scene-specific features. Huang et al. (2026) used multilevel alignment to encourage transferable representations, and Xie et al. (2026) exploited joint variable consistency and anchor-structure invariance to preserve structural relations under domain changes.

It should be noted that many source-side domain-generalization methods construct transformed views and encourage consistent class representations across them. However, a transformation that reduces scene-specific variation does not necessarily preserve class evidence for every sample. As illustrated in Fig. 1, adjacent-band mixing may retain a broad spectral pattern while weakening a narrow absorption feature, whereas spatial smoothing may preserve coarse structure while removing boundary or texture cues. In other words, enforcing uniform agreement across transformed views can suppress the information needed to distinguish classes in an unseen scene.

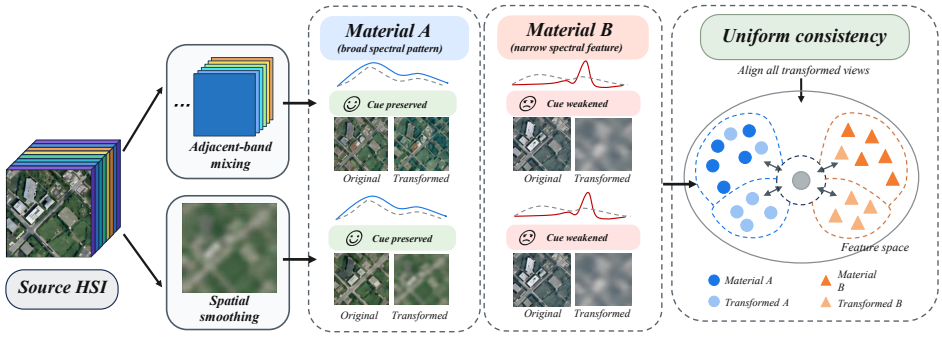


Figure 1: The same transformation may preserve class cues for one material while weakening them for another, making uniform consistency potentially suppress discriminative information.

To remedy such deficiency, we propose the transformation-conditioned discrete self-supervision (TCDS) framework shown in Fig. 2. From source observations, TCDS creates virtual scene variations and estimates their sample-wise semantic preservation through spectral and predictive consistency. Separate spectral and spatial codebooks are aligned at the semantic-index level, and their assignments form a preservation-weighted consensus that favors views retaining more class evidence. A residual branch retains complementary information and contributes to prediction through a preservation-dependent gate. Training constraints regulate semantic agreement, residual separation, code usage, and virtual-scene weighting. **Our main contributions** are summarized as follows: *First*, we introduce a transformation-conditioned formulation for cross-scene HSI domain generalization. The preservation of each source-derived transformation is estimated for each sample from spectral and predictive consistency. *Second*, we align separate spectral and spatial codebooks at the semantic-index level and form a preservation-weighted consensus across transformed views. *Third*, a complementary residual representation retains information outside the shared semantic codes and contributes to

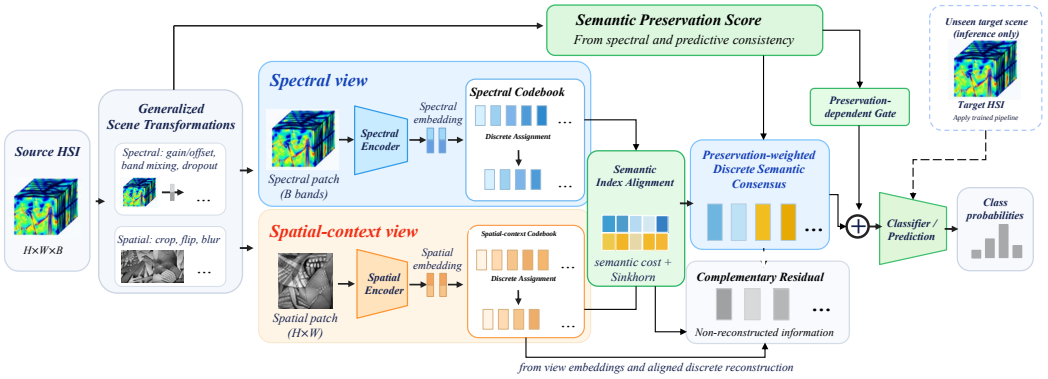


Figure 2: Overall framework of TCDS for cross-scene HSI domain generalization.

prediction through a preservation-dependent gate. *Fourth*, extensive experimental results on three datasets have been provided to demonstrate the effectiveness and superiority of TCDS.

2 Related Work

2.1 Self-supervised invariance

Self-supervised learning seeks invariance across transformed views of the same observation. The simple framework for contrastive learning of visual representations (SimCLR), bootstrap your own latent (BYOL), and variance–invariance–covariance regularization (VICReg) use different objectives, but each assumes that the transformations preserve semantic content (Chen et al., 2020; Grill et al., 2020; Bardes et al., 2022). Invariant risk minimization (IRM) and variance risk extrapolation (VREx) address invariance across training environments by seeking a shared optimal classifier and reducing risk variation, respectively (Arjovsky et al., 2019; Krueger et al., 2021). In single-source cross-scene HSI domain generalization, multiple observed training environments are unavailable. Source-derived transformations provide virtual scene variations, but their contribution should reflect how well they retain class-relevant spectral and spatial cues.

2.2 Discrete representation

Discrete representation maps continuous features to a finite set of reusable codes. Vector-quantized models such as the vector-quantized variational autoencoder (VQ-VAE) and VQ-VAE-2 jointly learn an encoder and a codebook, allowing samples with different continuous features to share a common discrete index (van den Oord et al., 2017; Razavi et al., 2019). Such indices provide a compact description of recurring feature patterns and are well suited to comparing semantic assignments when continuous representations shift across scenes. For cross-scene HSI, this property is useful only if independently learned spectral and spatial codebooks can be expressed in a common semantic index system.

3 Methodology

3.1 Problem formulation

Cross-scene HSI classification requires a model learned from a labeled source scene to remain discriminative under previously unseen scene variations. Since no target observations are available during training, the learning problem is formulated from the source domain, while the target risk serves only as the generalization objective. Let $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{N_s}$ denote the labeled source scene, where N_s is the number of labeled source samples, $x_i \in \mathbb{R}^{H\times W\times B}$ is an HSI patch, and $y_i \in \{1, \ldots, C\}$ is its land-cover label. Here $H \times W$ is the patch size, B is the number of spectral bands, and C is the number of classes. The unseen target distribution is denoted by $\mathcal{D}_t$, with target risk

$$R_t(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\ell(f_\theta(x), y), \tag{1}$$

where θ represents the learnable model parameters, f_θ is the classifier, $\ell(\cdot,\cdot)$ is the classification loss, and $\mathbb{E}$ denotes expectation over $\mathcal{D}_t$. Accordingly, the model is learned from the labeled source observations and the source-derived scene variations introduced below. For each HSI patch, the spectral view $x_i^\lambda = \mathcal{S}(x_i)$ retains the central and locally pooled spectra, while the spatial-context view $x_i^s = \mathcal{P}(x_i)$ describes neighborhood structure after a low-dimensional spectral projection. The deterministic operators $\mathcal{S}$ and $\mathcal{P}$ are applied to the same HSI cube, so the two views remain synchronized.

3.2 Generalized scene transformations

The two views describe complementary properties of the source scene but do not expose the model to changes in acquisition conditions. As illustrated in Fig. 2, we introduce such variation without target data by constructing virtual scenes from source observations using a transformation bank $\mathcal{T} = \{\tau_m\}_{m=0}^{M}$ containing $M+1$ transformations, where m indexes the transformation and τ_0 is the identity. The non-identity transformations are selected from spectral-response smoothing, adjacent-band mixing, band masking, random spectral convolution, and spatial-context changes according to the task-specific bank described in Sec. 4.1. Spectral and spatial augmentation has also been used to diversify source-derived scenes in cross-scene HSI classification (Yang et al., 2025). Because these transformations can retain class evidence to different degrees, their contributions are determined from spectral and predictive consistency. The spectral discrepancy is measured using the spectral angle mapper (SAM) as $d_{im}^\lambda = \mathrm{SAM}(\bar{x}_i, \overline{\tau_m(x_i)})/\pi$, where $\bar{x}_i$ and $\overline{\tau_m(x_i)}$ are the spatially averaged spectral vectors. The predictive discrepancy uses the Jensen–Shannon divergence (JSD) as $d_{im}^p = \mathrm{JSD}(p_0(\cdot \mid x_i), p_0(\cdot \mid \tau_m(x_i)))$, where p_0 is the warm-start source classifier. These two discrepancies define the semantic preservation score by

$$a_{im} = \exp\left[-\frac{d_{im}^\lambda + d_{im}^p}{\rho}\right] \in (0,1]. \tag{2}$$

where $\rho > 0$ controls the smoothness of the score. A larger a_{im} assigns greater influence to transformations that retain the source class evidence.

3.3 View-specific discrete semantic encoding

The preservation score determines how strongly a transformed sample affects invariant learning, while the spectral and spatial observations retain separate representations. For view $v \in \{\lambda, s\}$, the transformed observation is encoded as $z_i^{v,m} = g_v(\mathcal{V}_v(\tau_m(x_i)))$, where g_v is the view-specific encoder, $\mathcal{V}_\lambda = \mathcal{S}$, and $\mathcal{V}_s = \mathcal{P}$. Discrete representations have been used to establish shared coordinates across heterogeneous data in cross-modal adaptation (Liang et al., 2025). Here, the HSI spectral and spatial views retain separate codebooks whose semantic indices are subsequently aligned. Each view has an independent codebook $E^v = [e_1^v, \ldots, e_K^v]^\top \in \mathbb{R}^{K\times d}$, where d is the code embedding dimension. The assignment to code k is

$$q_{ik}^{v,m} = \frac{\exp\left(-\|z_i^{v,m} - e_k^v\|_2^2/T_q\right)}{\sum_{j=1}^{K}\exp\left(-\|z_i^{v,m} - e_j^v\|_2^2/T_q\right)}, \tag{3}$$

where $q_{ik}^{v,m}$ is the assignment probability, K is the codebook size, and $T_q > 0$ is the assignment temperature. The vector $q_i^{v,m} \in \mathbb{R}^K$ gives the discrete semantic coordinates of view v.

Since the two codebooks are learned independently, equal indices do not necessarily represent the same class semantics. Their correspondence is established from the class evidence accumulated by each code. Let $\mu_k^v \in \mathbb{R}^C$ denote the source-label distribution associated with code k in view v. We measure the cross-view semantic cost by $\Gamma_{kj} = 1 - \mu_k^{\lambda\top}\mu_j^s/(\|\mu_k^\lambda\|_2\|\mu_j^s\|_2)$ and obtain the index alignment as follows:

$$A = \mathrm{Sinkhorn}(-\Gamma/T_a)\,. \tag{4}$$

where $\Gamma \in \mathbb{R}^{K\times K}$ is the semantic-cost matrix, $A \in \mathbb{R}_+^{K\times K}$ is the resulting alignment matrix, $T_a > 0$ is the alignment temperature, and $\mathrm{Sinkhorn}(\cdot)$ denotes Sinkhorn normalization. The normalization makes A approximately doubly stochastic, satisfying $A\mathbf{1} \approx \mathbf{1}$ and $A^\top\mathbf{1} \approx \mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^K$ is the all-ones vector. The spatial assignment is then transported into the spectral index space by $\tilde{q}_i^{s,m} = Aq_i^{s,m}$, while $\tilde{q}_i^{\lambda,m} = q_i^{\lambda,m}$. The aligned assignments thus share a common index system while the two codebooks retain their view-specific geometry.

3.4 DISCRETE SEMANTIC INVARIANCE ACROSS SCENE TRANSFORMATIONS

Once the code indices are aligned, assignments from different virtual scenes can be compared in the same semantic coordinate system. Their contributions should nevertheless depend on how well each transformation preserves class evidence. The aligned assignments are therefore integrated into a preservation-weighted geometric consensus as follows:

$$c_i = \mathrm{softmax}\left(\frac{1}{2T_c \sum_{m=0}^{M} a_{im}} \left[\sum_{m=0}^{M} a_{im} \sum_{v\in\{\lambda,s\}} \log\left(\tilde{q}_{ik}^{v,m} + \epsilon\right) \right]_{k=1}^{K} \right). \tag{5}$$

where $c_i \in \mathbb{R}_+^K$ satisfies $\mathbf{1}^\top c_i = 1$, $T_c > 0$ is the consensus temperature, and $\epsilon > 0$ avoids the logarithm of zero. The factor $2\sum_{m=0}^{M} a_{im}$ normalizes the contributions from the two views and all transformed instances. The resulting c_i emphasizes codes supported across views and transformations that retain class-relevant information. A purely shared representation may discard local information that remains useful for discrimination. The component not reconstructed by the aligned discrete semantics is retained as:

$$r_i^{v,m} = z_i^{v,m} - E^{v\top} \tilde{q}_i^{v,m}, \tag{6}$$

where $r_i^{v,m} \in \mathbb{R}^d$ is the complementary residual. The consensus and residual are combined through a preservation-dependent gate:

$$f_\theta(x_i) = W_c c_i + \gamma_r \beta_i W_r \bar{r}_i, \qquad \beta_i = 1 - \frac{1}{M+1} \sum_{m=0}^{M} a_{im}, \tag{7}$$

where $f_\theta(x_i) \in \mathbb{R}^C$ gives the logits, $W_c \in \mathbb{R}^{C\times K}$ and $W_r \in \mathbb{R}^{C\times d}$ are learnable projections, $\bar{r}_i = \frac{1}{2}(r_i^{\lambda,0} + r_i^{s,0})$, $\beta_i \in [0,1)$ is the preservation-dependent gate, and $\gamma_r > 0$ controls the residual contribution. Accordingly, prediction relies more on the shared discrete semantics at high preservation, while the residual contribution increases when the transformed views retain less class evidence.

3.5 CONSTRAINED LEARNING OF DISCRETE SEMANTIC INVARIANCE

The consensus–residual decomposition must remain meaningful across heterogeneous virtual scenes. Constrained-risk analysis has also been explored for HSI domain generalization (Gao et al., 2026). Here, we allow the virtual-scene weights to vary within a Kullback–Leibler (KL) neighborhood of the uniform distribution instead of averaging their losses uniformly. Let $\pi \in \Delta^{M+1}$ denote the scene weights and let π_0 be uniform. Using cross-entropy (CE) for classification, the resulting robust source risk is

$$\mathcal{R}_{\mathrm{rob}}(\theta) = \max_{\pi\in\Delta^{M+1}:\mathrm{KL}(\pi\|\pi_0)\le\delta} \sum_{m=0}^{M} \pi_m \mathbb{E}_{(x,y)\sim\mathcal{D}_s} \mathrm{CE}\big(f_\theta(\tau_m x), y\big), \tag{8}$$

where π_m is the weight of virtual scene m, Δ^{M+1} is the probability simplex, KL denotes the KL divergence, $\delta > 0$ controls the reweighting range, and CE denotes the CE loss. The maximization assigns more weight to virtual scenes with larger classification risk. Further, robust reweighting alone does not guarantee that the learned discrete space remains semantically coherent. We therefore constrain the preservation-weighted deviation from the consensus as:

$$g_{\mathrm{inv}}(\theta) = \max_{v,m} \mathbb{E}_i[a_{im}\,\mathrm{JSD}(\tilde{q}_i^{v,m}, c_i)] - \varepsilon_{\mathrm{inv}} \le 0, \tag{9}$$

where $\mathbb{E}_i$ averages over source samples and $\varepsilon_{\mathrm{inv}} \ge 0$ is the agreement budget. Residual separation is imposed by requiring the maximum dimension-normalized overlap $\max_{v,m} \mathbb{E}_i[d^{-1}|\langle r_i^{v,m}, E^{v\top} c_i\rangle|]$ not to exceed $\varepsilon_{\mathrm{orth}}$, which defines $g_{\mathrm{orth}}(\theta) \le 0$. With $\bar{c} = N_s^{-1}\sum_i c_i$ and $H(\bar{c}) = -\sum_k \bar{c}_k \log \bar{c}_k$, code usage is further regulated by $H(\bar{c}) \ge \log K - \eta$, defining $g_{\mathrm{bal}}(\theta) \le 0$, where $0 \le \eta < \log K$. These conditions preserve semantic agreement while preventing the residual from duplicating the shared representation or the codebook from collapsing to a small subset of indices.

The model is optimized over this feasible set with nonnegative Lagrange multipliers. Writing $(g_1, g_2, g_3) = (g_{\text{inv}}, g_{\text{orth}}, g_{\text{bal}})$ and $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ gives

$$\begin{aligned} &\min_{\theta} \ \mathcal{R}_{\text{rob}}(\theta) \quad \text{s.t.} \quad g_j(\theta) \le 0, \ j = 1, 2, 3, \\ &\mathcal{L}(\theta, \alpha) = \mathcal{R}_{\text{rob}}(\theta) + \sum_{j=1}^{3} \alpha_j g_j(\theta), \qquad \alpha_j \ge 0. \end{aligned} \tag{10}$$

Optimization alternates between gradient descent on the model parameters and projected ascent on the nonnegative dual multipliers, so constraint violations are corrected during training rather than absorbed into a fixed weighted loss. The scene weights are updated within the same robust objective. With $\hat{R}_m$ denoting the empirical risk of scene m and $\pi_{0m} = 1/(M+1)$, their closed-form reweighting is

$$\pi_m \propto \pi_{0m} \exp\left(\frac{\hat{R}_m}{\nu}\right), \quad \text{with } \nu > 0 \text{ chosen so that } \mathrm{KL}(\pi \| \pi_0) \le \delta. \tag{11}$$

Algorithm 1 summarizes the resulting source-only training procedure.

Algorithm 1 Training TCDS on one labeled source scene

Require: Source patches $\mathcal{D}_s$, transformation bank $\mathcal{T}$, codebook size K, robust radius δ, constraint budgets $\varepsilon_{\text{inv}}, \varepsilon_{\text{orth}}, \eta$.

1: Warm-start p_0 on $\mathcal{D}_s$ and initialize $g_\lambda, g_s, E^\lambda, E^s, W_c, W_r$, $\pi = \pi_0$, and nonnegative dual variables α.
2: **for** each mini-batch **do**
3: Apply $\mathcal{T}$ to the source samples and compute a_{im} for $m = 0, \ldots, M$ by Eq. (2).
4: Encode the spectral and spatial-context views and obtain the discrete assignments by Eq. (3).
5: Estimate the cross-view semantic cost, obtain A by Eq. (4), and transport the spatial assignments into the spectral index space.
6: Form the preservation-weighted consensus c_i and complementary residuals $r_i^{v,m}$ by Eq. (5) and Eq. (6).
7: Compute the gate β_i and prediction logits by Eq. (7).
8: Evaluate the scene-wise risks $\hat{R}_m$, update π by Eq. (11), and compute $\mathcal{R}_{\text{rob}}$ by Eq. (8).
9: Evaluate the semantic-agreement, residual-separation, and code-balance constraints, and form $\mathcal{L}(\theta, \alpha)$ by Eq. (10).
10: Update θ by gradient descent and α by projected ascent.
11: **end for**
12: **return** p_0, encoders, codebooks, index alignment, and classifier.

At inference, for a target sample x_t, the source-selected transformation set is fixed before the target scene is observed and is applied independently to x_t. The preservation scores a_{tm} are recomputed by Eq. (2), and the consensus c_t follows Eq. (5). No target labels, dataset-level target statistics, parameter updates, or source–target alignment are used. Prediction then follows Eq. (7) with the identity-view residual $\bar{r}_t = \frac{1}{2}(r_t^{\lambda,0} + r_t^{s,0})$.

4 Experiments

4.1 Experimental setup

Dataset preparation. We evaluate TCDS on the Houston, Pavia, and HyRANK cross-scene benchmarks under the source-only domain generalization setting. Houston13→Houston18 uses prepared source and target scenes of 954×210 pixels each, with 48 spectral bands and seven shared categories retained for evaluation. Pavia University (PaviaU)→Pavia Center contains a 610×340 source scene and a 1096×715 target scene, both with 102 bands, from which seven shared categories are retained. The HyRANK task Dioni→Loukia contains 250×1376 and 249×945 pixels, respectively, with 176 bands and twelve shared categories. For all three tasks, each scene is scaled by its maximum

response and each pixel spectrum is further ℓ_2 normalized. At every labeled location, the central spectrum is combined with neighborhood summaries from 3×3, 7×7, and 13×13 windows. Source labels are split into 80% training and 20% validation samples by stratified sampling, whereas target labels are used only for final evaluation and never for training, early stopping, preservation scoring, or hyperparameter selection.

Implementation details. All hyperparameters are selected using only the held-out source split and deterministic source-derived pseudo-scenes. The constrained-learning parameters are fixed across all tasks as $(\delta, \varepsilon_{\mathrm{inv}}, \varepsilon_{\mathrm{orth}}, \eta) = (0.10, 0.50, 0.07, 2.0)$, with ν determined by δ through the KL constraint in Eq. (11). For Houston13→Houston18, PaviaU→Pavia Center, and Dioni→Loukia, (K, d, T_q, T_a) is set to $(96, 128, 0.20, 0.25)$, $(96, 128, 0.20, 0.25)$, and $(48, 96, 0.45, 0.25)$, while (ρ, T_c, γ_r) is set to $(0.30, 0.25, 1.15)$, $(0.30, 0.25, 1.15)$, and $(0.40, 0.35, 1.20)$, respectively. PaviaU→Pavia Center and Dioni→Loukia retain these settings at inference, whereas Houston13→Houston18 uses source-only calibration to select $(0.12, 0.28, 1.80)$ without target observations. At inference, Houston13→Houston18 and PaviaU→Pavia Center use four fixed source-derived views: the identity, spectral smoothing, adjacent-band mixing, and a spatial-context transformation. Dioni→Loukia adds random spectral convolution as a fifth view. Band masking is used only within source-side training augmentations and is not counted as an inference view. Class weights follow $w_c \propto n_c^{-0.5}$. Each model is retrained from scratch in three independent runs, and results are reported as mean±standard deviation using overall accuracy (OA), weighted F1 score, and Cohen's κ.

4.2 Comparison with Existing Methods

Compared methods. We compare TCDS with eight representative domain-generalization methods, including SDEnet (Zhang et al., 2023), FDGNet (Qin et al., 2025), S^{2}AMSNet (Chen et al., 2024), S^{2}ECNet (Dong et al., 2024), IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), min–max stylization and destylization (StyDeSty) (Liu et al., 2024), and spectral property-driven data augmentation (SPDDA) (Chen et al., 2026a). These methods cover spectral–spatial randomization, frequency disentanglement, synthetic-domain generation, representation enhancement, invariant-risk learning, and style-based source diversification. For a fair comparison, all methods use the same source split, spectral–spatial input, and two-layer multilayer perceptron backbone with layer normalization (LayerNorm) and gaussian error linear unit (GELU) activations. Their hyperparameters follow the recommended settings in the original papers or public implementations whenever available. Since IRM and VREx require multiple training environments, four source-derived environments are constructed using the identity view, spectral smoothing, adjacent-band mixing, and spatial mixing. Target labels are reserved solely for final evaluation and not for model selection or parameter tuning.

Table 1: Classification performance comparison on three cross-scene HSI benchmarks. All hyperparameters are selected using source data only. Values are mean±standard deviation.

	Houston13→Houston18			PaviaU→Pavia Center			Dioni→Loukia		
Method	OA	F1	κ	OA	F1	κ	OA	F1	κ
SDEnet	73.10±2.51	0.7105±0.0319	0.5264±0.0465	61.51±1.03	0.5777±0.0033	0.5332±0.0128	43.62±0.94	0.4255±0.0111	0.3469±0.0044
FDGNet	72.15±1.44	0.6914±0.0315	0.5031±0.0439	61.60±0.89	0.5802±0.0072	0.5313±0.0109	48.32±0.39	0.4401±0.0028	0.3875±0.0080
S^{2}AMSNet	72.45±1.94	0.7094±0.0235	0.4999±0.0450	61.39±0.11	0.5753±0.0041	0.5308±0.0030	49.10±2.44	0.4735±0.0228	0.3937±0.0198
S^{2}ECNet	71.92±2.68	0.6969±0.0329	0.5056±0.0379	61.34±1.00	0.5753±0.0044	0.5303±0.0108	45.91±1.96	0.4392±0.0240	0.3660±0.0180
IRM	64.58±0.92	0.5493±0.0154	0.2239±0.0477	61.26±0.15	0.5871±0.0011	0.5265±0.0010	44.43±1.29	0.4237±0.0114	0.3543±0.0089
VREx	63.47±1.04	0.5300±0.0129	0.1756±0.0512	60.11±0.77	0.5820±0.0034	0.5145±0.0100	45.09±1.30	0.4275±0.0101	0.3587±0.0118
StyDeSty	67.78±3.95	0.6115±0.0360	0.3916±0.0642	61.41±0.20	0.5683±0.0025	0.5269±0.0037	50.88±1.27	0.4891±0.0174	0.4145±0.0156
SPDDA	68.71±0.50	0.6347±0.0137	0.3772±0.0258	62.67±1.31	0.5886±0.0053	0.5440±0.0152	43.81±1.87	0.4160±0.0142	0.3466±0.0167
TCDS (Ours)	**74.40±1.78**	**0.7116±0.0312**	**0.5304±0.0568**	**66.32±1.67**	**0.6263±0.0162**	**0.5955±0.0223**	**56.09±1.34**	**0.5077±0.0040**	**0.4519±0.0129**

Comparison results. Table 1 reports the OA, weighted F1, and κ results on the three cross-scene tasks. TCDS achieves the highest mean performance on Houston13→Houston18, PaviaU→Pavia Center, and Dioni→Loukia, with OA values of 74.40 ± 1.78, 66.32 ± 1.67, and 56.09 ± 1.34, respectively. These results exceed the strongest competing method by 1.30, 3.65, and 5.21 percentage points, while the weighted F1 and κ scores show the same overall advantage. The improvement is attributed to the joint effect of preservation-aware transformation weighting, semantic index alignment, and the consensus–residual representation. Transformations that weaken class evidence are assigned smaller contributions, while the aligned discrete assignments establish a common semantic space across views. The residual branch further retains complementary information that is not captured by the shared semantic codes. Fig. 3 provides representative classification maps on PaviaU→Pavia Center. Compared with the competing methods, TCDS produces more continuous regions with fewer

scattered errors near class boundaries and small man-made structures, yielding a spatial distribution closer to the ground truth.

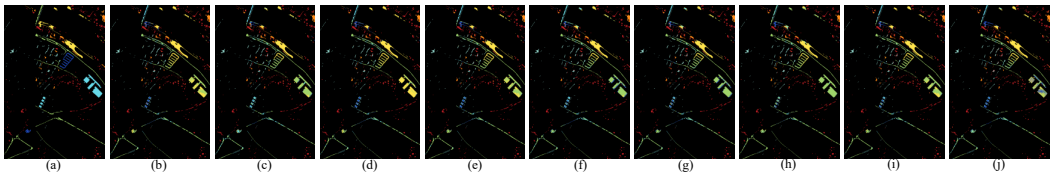


Figure 3: Representative classification-map comparison on PaviaU→Pavia Center. Panels (a)–(j) show ground truth, SDEnet, FDGNet, S^{2}AMSNet, S^{2}ECNet, IRM, VREx, StyDeSty, SPDDA, and TCDS, respectively. Only pixels belonging to the shared evaluation classes are displayed.

4.3 Ablation Study

In this subsection, we conduct ablation experiments to investigate how each component of TCDS affects its cross-scene classification performance. Fig. 4 reports the ablation results on PaviaU→Pavia Center, where the complete model obtains an OA of $66.32 \pm 1.67\%$. We first replace the dual codebooks with a shared codebook, which reduces the mean OA to 63.52%. We then remove semantic index alignment, preservation weighting, code balance, and the transformation bank separately, obtaining mean OA values of 63.45%, 63.97%, 64.12%, and 63.16%, respectively. As can be observed from Fig. 4, all ablated variants perform worse than the complete model, indicating that these components provide complementary contributions to the learned discrete semantics. The most pronounced degradation occurs when the residual branch is removed, for which the mean OA drops to 59.37%. This result indicates that the shared semantic codes alone do not retain all class-discriminative information and that the residual representation is important for preserving complementary information required for cross-scene classification.

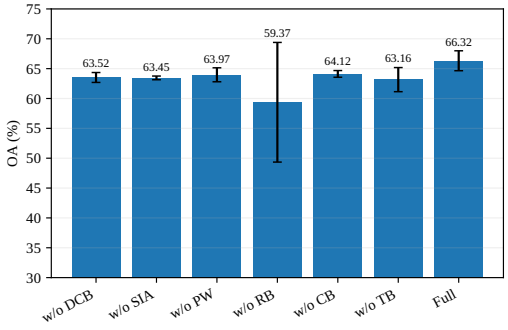


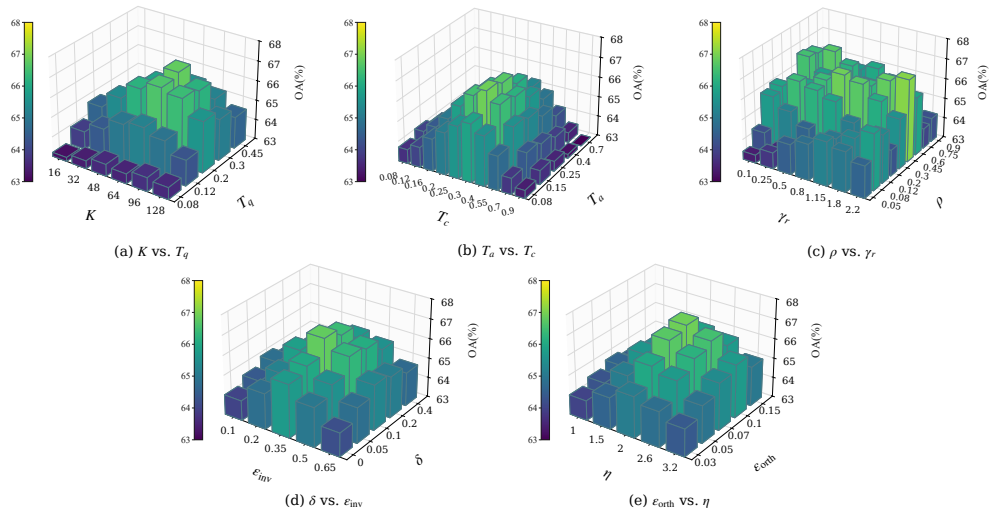
Figure 4: Ablation results on PaviaU→Pavia Center. Bars denote mean OA and error bars denote standard deviation. Dual codebooks (DCB), semantic index alignment (SIA), preservation weighting (PW), residual branch (RB), code balance (CB), and transformation bank (TB) denote the ablated components.

4.4 Pairwise Parameter Sensitivity Analysis

We further organize the ten analyzed parameters into five related pairs to provide a compact view of their coupled sensitivity patterns. Fig. 5 considers (K, T_q), (T_a, T_c), (ρ, γ_r), $(\delta, \varepsilon_{\text{inv}})$, and $(\varepsilon_{\text{orth}}, \eta)$ on PaviaU→Pavia Center. The first pair exhibits a broad high-performing region once the codebook has sufficient capacity, while the assignment temperature remains most favorable in the moderate range. The two temperature parameters form a tilted response region rather than a symmetric optimum, reflecting their different roles in assignment and consensus formation. The preservation scale and residual contribution show a broader asymmetric ridge, whereas the two constrained-learning pairs present comparatively flatter plateaus with moderate local variation. These paired views summarize the interaction patterns of the analyzed parameters, while target labels are used only for reporting and are not involved in source-only parameter selection.

4.5 Visualization of Preservation-Aware Discrete Semantic Learning

To examine preservation-aware discrete semantic learning, we analyze held-out PaviaU→Pavia Center samples. Preservation scores are produced by the warm-start source classifier p_0, while Fig. 6(a) uses an independent source probe. Quartiles are formed within each class–transformation stratum after excluding the identity view, giving $7 \times 3 = 21$ strata, and the errors are macro-averaged across them. The balanced probe error drops from 14.73% in Q_1 to 3.85%, 1.53%, and 1.40% in Q_2–Q_4,

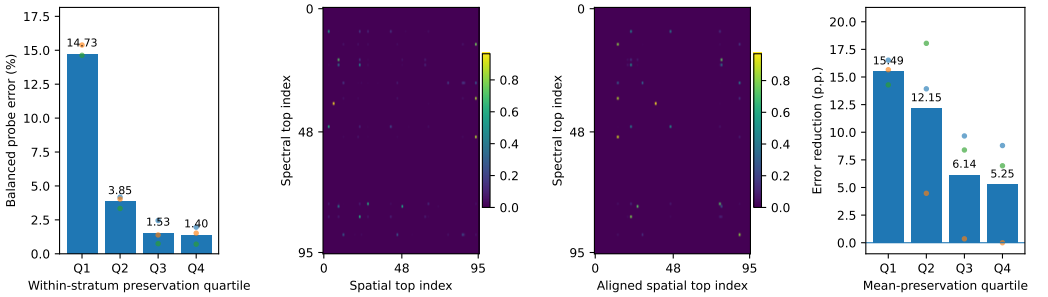


(a) K vs. T_q (b) T_a vs. T_c (c) ρ vs. γ_r

(d) δ vs. ε_{inv} (e) $\varepsilon_{\text{orth}}$ vs. η

Figure 5: Pairwise parameter-sensitivity visualization on PaviaU→Pavia Center for (a) K and T_q, (b) T_a and T_c, (c) ρ and γ_r, (d) δ and ε_{inv}, and (e) $\varepsilon_{\text{orth}}$ and η.

showing fewer transformed-sample errors at higher preservation. Fig. 6(b)–(c) shows that semantic index alignment raises the diagonal mass from 2.50% to 44.48%, yielding clearer spectral–spatial correspondence. In Fig. 6(d), restoring the residual branch reduces mean classification error by 15.49, 12.15, 6.14, and 5.25 percentage points from Q_1 to Q_4. The larger average gain at low preservation indicates that the residual pathway contributes most when the shared discrete representation retains less class-relevant information.



(a) Balanced preservation diagnostic (b) Before alignment (c) After alignment (d) Residual-branch benefit

Figure 6: Visualization of preservation-aware discrete semantic learning on PaviaU→Pavia Center. (a) Independent-probe error across preservation quartiles. (b)–(c) Spectral–spatial top-index correspondence before and after semantic index alignment. (d) Classification-error reduction provided by the residual branch across preservation levels.

5 Conclusion

This paper presents TCDS for cross-scene HSI domain generalization, which learns discrete semantic representations that preserve class-relevant information under source-derived scene variations. More specifically, the semantic preservation of each transformation is estimated from spectral and predictive consistency, and separate spectral and spatial-context codebooks are aligned at the semantic-index level to establish a common semantic coordinate system. Based on the estimated preservation, the aligned assignments are integrated through a weighted consensus so that transformations retaining more class evidence contribute more strongly to the shared representation. A complementary residual branch is further introduced to retain discriminative information outside the shared semantic codes, while constrained learning regulates semantic agreement, residual separation, code usage, and virtual-scene weighting. Extensive experimental results on Houston13→Houston18, PaviaU→Pavia Center, and Dioni→Loukia demonstrate the effectiveness of TCDS for cross-scene HSI classification.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2022.

Taiqin Chen, Yifeng Wang, Xiaochen Feng, Zhilin Zhu, Hao Sha, Yingjian Li, and Yongbing Zhang. Spectral property-driven data augmentation for hyperspectral single-source domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3038–3046, 2026a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pp. 1597–1607, 2020.

Xi Chen, Lin Gao, Maojun Zhang, Chen Chen, and Shen Yan. Spectral-spatial adversarial multidomain synthesis network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Xi Chen, Maojun Zhang, Shen Yan, Yuxiang Liu, and Chen Chen. Decoupling and aligning modality-shared semantics for single-domain generalization in hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:25833–25844, 2025.

Xi Chen, Maojun Zhang, Yuxiang Liu, Chen Chen, and Shen Yan. Cross-specl: Cross-scene hyperspectral image classification via spectral stability exploiting and causal learning. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026b.

Lijia Dong, Jie Geng, and Wen Jiang. Spectral-spatial enhancement and causal constraint for hyperspectral image cross-scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

Zhe Gao, Bin Pan, and Zhenwei Shi. C^3DG: Conditional domain generalization for hyperspectral imagery classification with convergence and constrained-risk theories. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026.

Pedram Ghamisi, Naoto Yokoya, Jun Li, Wenzhi Liao, Sicong Liu, Javier Plaza, Behnood Rasti, and Antonio Plaza. Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):37–78, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Haotian Hu, Yunpeng Zheng, Qian Liu, Shuai Yang, Biqi Wang, Xiaohui Yuan, and Lichuan Gu. Hyperspectral single-source domain generalization via structured data simulation and domain-disparity decorrelation. *Knowledge-Based Systems*, 348:116429, 2026.

Lingbo Huang, Xin He, and Yushi Chen. Multilevel alignment-based domain generalization network for cross-scene hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19:7742–7755, 2026.

Xizeng Huang, Yanni Dong, Yuxiang Zhang, and Bo Du. Single-source frequency transform for cross-scene classification of hyperspectral image. *IEEE Transactions on Image Processing*, 34: 3000–3012, 2025.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation. In *Proceedings of the International Conference on Machine Learning*, pp. 5815–5826, 2021.

Jian Li, Zhen Zhang, Rui Song, Yushi Li, and Qian Du. Scformer: Spectral coordinate transformer for cross-domain few-shot hyperspectral image classification. *IEEE Transactions on Image Processing*, 33:840–855, 2024.

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

Zhengyin Liang, Hui Yin, Min Liang, Qianqian Du, Ying Yang, and Hua Huang. UniDxMD: Towards unified representation for cross-modal unsupervised domain adaptation in 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20346–20356, 2025.

Songhua Liu, Xin Jin, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Stydesty: Min-max stylization and destylization for single domain generalization. In *Proceedings of the International Conference on Machine Learning*, pp. 31293–31311, 2024.

Danyang Peng, Jun Wu, Tingting Han, Yuanyuan Li, Yi Wen, Guangyu Yang, and Lei Qu. Disentanglement-inspired single-source domain-generalization network for cross-scene hyperspectral image classification. *Knowledge-Based Systems*, 303:112413, 2024.

Yunxiao Qi, Dongyang Liu, Junping Zhang, and Ye Zhang. A shift reduction domain generalization network for hyperspectral image cross-domain classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 2025.

Boao Qin, Shou Feng, Chunhui Zhao, Bobo Xi, Wei Li, and Ran Tao. Fdgnet: Frequency disentanglement and data geometry for domain generalization in cross-scene hyperspectral image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10297–10310, 2025.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Boshan Shi, Guo Cao, Youqiang Zhang, Yanbo Liu, and Kai Yang. Fourier-based spectral–spatial generator for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–17, 2025.

Xiaobo Tang, Chao Li, and Yuxin Peng. Unsupervised joint adversarial domain adaptation for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Biqi Wang, Yang Xu, Zebin Wu, Shangdong Zheng, Zhihui Wei, and Jocelyn Chanussot. Hyperspectral images single-source domain generalization based on nonlinear sample generation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024a.

Haoyu Wang, Tianyang Duan, and Xiaomin Liu. SIBDNet: Spectrum information bottleneck decoupling network for domain generalization in cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026.

Xiaozhen Wang, Jiahang Liu, Yue Ni, Weijian Chi, and Yangyu Fu. Two-stage domain alignment single-source domain generalization network for cross-scene hyperspectral images classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024b.

Mengying Xie, Jingwen Han, Yantao Li, and Shaojiang Deng. Cross-domain hyperspectral image generalization based on joint variable consistency and anchor structure invariance. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026.

Kai Yang, Boshan Shi, Guo Cao, and Youqiang Zhang. Fourier phase preservation and spectral random augmentation for cross-scene hyperspectral image classification with ensemble networks. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–17, 2025.

Yuxiang Zhang, Wei Li, Ran Tao, Jiayi Peng, Qian Du, and Zhihua Cai. Cross-scene hyperspectral image classification with discriminative cooperative alignment. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9646–9660, 2021.

Yuxiang Zhang, Wei Li, Weidong Sun, Ran Tao, and Qian Du. Single-source domain expansion network for cross-scene hyperspectral image classification. *IEEE Transactions on Image Processing*, 32:1498–1512, 2023.

Chunhui Zhao, Bowen Qin, Shou Feng, Wei Zhu, Lefei Zhang, and Jinchang Ren. An unsupervised domain adaptation method towards multi-level features and decision boundaries for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16, 2022.

Hanqing Zhao, Lianlei Lin, Junkai Wang, Sheng Gao, and Zongwei Zhang. Adversarial decoupling domain generalization network for cross-scene hyperspectral image classification. *Knowledge-Based Systems*, 318:113432, 2025.