Knowledge-Based Systems 348 (2026) 116429

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Hyperspectral single-source domain generalization via structured data simulation and domain-disparity decorrelation

Haotian Hu [a,1], Yunpeng Zheng [a,1], Qian Liu [a,b,*], Shuai Yang [a,b], Biqi Wang [c], Xiaohui Yuan [d], Lichuan Gu [a,b]

[a] School of Information and Artificial Intelligence, Anhui Agricultural University, HeFei, 230036, China
[b] Anhui Provincial Engineering Research Center for Agricultural Information Perception and Intelligent Computing, HeFei, China
[c] Nanjing University of Posts and Telecommunications, Nanjing, 210003, China
[d] University of North Texas, Denton, TX, 76203, USA

A R T I C L E I N F O

Keywords:
Hyperspectral image classification
Single-source domain generalization
Structured data simulation
Domain-disparity decorrelation

A B S T R A C T

Cross-scene hyperspectral image (HSI) classification faces significant challenges due to domain shifts arising from structured spectral–spatial variations. While Single-Source Domain Generalization provides a viable pathway by learning from only one source domain, prevailing methods, particularly those built on generative-adversarial strategies, remain constrained by two issues. On the one hand, they predominantly depend on unstructured randomization, which ignores the inherent structured variations of cross-scene HSI. On the other hand, in striving for domain invariance, they tend to enforce strict alignment, which can suppress discriminative intra-domain features that are valuable for classification. To address these challenges, this paper proposes SD^{3}Net, a novel framework grounded in structured data simulation and domain-disparity decorrelation. Specifically, to ensure structural fidelity, we introduce the Structured Data Simulation Generator (SDSG), which shifts data generation method from stochastic perturbation to logic-driven emulation. It consists of a Semantic-Anchored Layout Simulator (SALS) to reorganize spatial topologies based on semantic logic, and a Multi-Scale Spectral Variation Simulator (MSVS) to emulate spectral distortions induced by sensor and atmospheric variations. Meanwhile, to resolve the dilemma between invariance and discriminability, we propose the Domain-Disparity Decorrelation Discriminator (D^3). Instead of enforcing strict alignment, this mechanism explicitly models feature residuals between source and simulated domains and maximizes their predictive entropy. This strategy effectively decouples domain-specific disparities from the feature representation while preserving essential intra-domain semantics. Extensive experiments on multiple cross-scene HSI datasets demonstrate that SD^{3}Net significantly outperforms state-of-the-art methods, validating the efficacy of the proposed simulation and decorrelation strategies.

1. Introduction

Hyperspectral images (HSIs) capture continuous and fine-grained spectral signatures across hundreds of bands, serving as a fundamental data source for precision earth observation tasks, such as environmental monitoring and agricultural analysis [1,2]. The core objective of HSI classification aims to assign accurate semantic labels to each pixel based on these rich material properties [3,4]. In this context, deep learning-based approaches have been extensively explored and have demonstrated state-of-the-art performance in the semantic segmentation of remote sensing imagery [5]. However, the real-world deployment of these models is severely hampered by domain shift. Unlike general optical imagery where variations are often stylistic, domain shifts in HSI are inherently physical and highly structured. These discrepancies usually manifest in two distinct forms. The first is spectral variation, characterized by systematic deviations in spectral response induced by sensor configurations, atmospheric attenuation, and fluctuating illumination; while the second is spatial structural disparity, where the topological layout and neighborhood context of the same land covers vary systematically across scenes. Consequently, it is necessary to construct robust hyperspectral classification models capable of mitigating the adverse performance impact caused by structured spectral–spatial discrepancies in cross-scene scenarios.

* Corresponding author at: School of Information and Artificial Intelligence, Anhui Agricultural University, HeFei, 230036, China.
E-mail addresses: Haosky@stu.ahau.edu.cn (H. Hu), zheng_yunp@stu.ahau.edu.cn (Y. Zheng), liuq@ahau.edu.cn (Q. Liu), yangs@mail.hfut.edu.cn (S. Yang), 20250121@njupt.edu.cn (B. Wang), xiaohui.yuan@unt.edu (X. Yuan), glc@ahau.edu.cn (L. Gu).
[1] The two authors contribute equally to this work.

https://doi.org/10.1016/j.knosys.2026.116429
Received 31 January 2026; Received in revised form 27 April 2026; Accepted 7 June 2026
Available online 12 June 2026

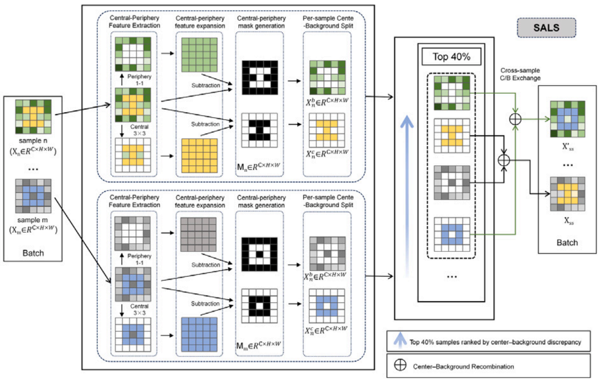


Fig. 2. Detailed structure of the SALS. It rearranges spatial layouts by swapping target and background components based on semantic masks.

2.1.1. Semantic-anchored layout simulator

Prevailing style-transfer methods, such as those using Adaptive Instance Normalization, primarily perform appearance-level perturbations by manipulating channel-wise statistics. While effective for altering low-level textures, these operations are insufficient to simulate the complex structural variations dictated by real-world geography. Therefore, our semantically guided SALS is designed to generate logically coherent spatial layouts by reorganizing image regions based on semantic cues, operating beyond mere style perturbation. As illustrated in Fig. 2, the SALS first decomposes an input patch into target and background components via semantic masking and then performs cross-sample region swapping to simulate realistic spatial rearrangements. Given a mini-batch of HSI patches $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, the SALS first identifies the semantic composition for each instance. For an arbitrary sample $\mathbf{X}_n$ in the batch, we compute the mean feature vectors of the central object region (μ_c^n), aggregated from a fixed neighborhood and the peripheral background (μ_e^n). These prototypes represent the target semantics and environmental context, respectively. A semantic binary mask$\mathbf{M}_n \in \{0,1\}^{C\times H\times W}$ is then generated by assigning each pixel spatial position to the nearest prototype in terms of distance as Eq. (1):

$$M_n(i,j) = \begin{cases} 1, & \text{if } \|\mathbf{x}_{i,j} - \mu_c^n\|_1 < \|\mathbf{x}_{i,j} - \mu_e^n\|_1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{x}_{i,j}$ represents the spectral feature vector at spatial coordinate (i,j) within the patch.

Using this mask $\mathbf{M}_n$, the patch $\mathbf{X}_n$ is decomposed into a target component $\mathbf{X}_n^c = \mathbf{X}_n \odot \mathbf{M}_n$ and a background component $\mathbf{X}_m^b = \mathbf{X}_n \odot (1 - \mathbf{M}_n)$, where $\odot$ denotes element-wise multiplication and 1 is an all-ones tensor.

To simulate plausible spatial rearrangements, we compute the center-to-background intensity disparity for all samples in a mini-batch. The top 40% of samples with the most pronounced disparities are selected for cross-sample swapping. The spatially augmented feature $\mathbf{X}_{ss}$ is obtained by combining the target component of one sample with the background of another:

$$\mathbf{X}_{ss} = \mathbf{X}_n^c + \mathbf{X}_m^b, \tag{2}$$

where $\mathbf{X}_{ss}$, $\mathbf{X}_n^c$, $\mathbf{X}_m^b \in \mathbb{R}^{C\times H\times W}$, $\mathbf{X}_m^b$ is the background component of another sample from this mini-batch that meets the requirements. This mechanism compels the model to learn features invariant to logical spatial context variations while preserving object identity.

2.1.2. Multi-scale spectral variation simulator

Spectral responses vary due to sensor characteristics, illumination, and atmospheric conditions. As illustrated in Fig. 3, the MSVS manipulates the spectral appearance by extracting and reassembling channel-wise statistics across samples. Since these statistics largely capture style-related information, their exchange allows us to synthesize novel, realistic spectral variations while preserving the underlying spatial content structure.

Multi-scale Channel Grouping Strategy. To comprehensively model both broad trends and fine details in spectral signatures, we process features at multiple granularities. This is achieved by randomly sampling channel subsets of varying sizes from the original 64-channel features to represent different spectral scales. This is implemented by grouping the spectral channels into sets of varying sizes, defined by a scale set $S = \{64, 32, 16, 8\}$. Here, each scale $s \in S$ specifies the number of adjacent channels treated as a cohesive group. A larger group size (e.g., 64) models coarser, interdependent spectral patterns, while a smaller group size (e.g., 8) captures finer-grained, localized spectral variations.

Statistics Computation and Exchange. For each channel group at scale s, we compute its mean μ and standard deviation σ across spatial dimensions, which serves as a concise descriptor of the group's spectral distribution:

$$\mu(\mathbf{F}_{s,c}) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{s,c,i,j} \tag{3}$$

$$\sigma(\mathbf{F}_{s,c}) = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(F_{s,c,i,j} - \mu(\mathbf{F}_{s,c})\right)^2 + \epsilon} \tag{4}$$

Here, $F_{s,c,i,j}$ denotes the feature value at spatial coordinates (i,j) within the cth channel at the sth scale, and ϵ is a small constant introduced to prevent division by zero and ensure numerical stability. The core of style simulation lies in exchanging these statistical descriptors between different samples within a training batch. For a sample, we replace its original statistics at each scale with those from a randomly selected sample, effectively transferring the multi-scale spectral style:

$$\mathbf{F}_{s,c}(n) = \sigma\big(\mathbf{F}_{s,c}(m)\big) \odot \left(\frac{\mathbf{F}_{s,c}(n) - \mu\big(\mathbf{F}_{s,c}(n)\big)}{\sigma\big(\mathbf{F}_{s,c}(n)\big)}\right) + \mu\big(\mathbf{F}_{s,c}(m)\big) \tag{5}$$

This operation yields a set of stylized feature maps $\mathbf{F}_s$ for each scale s.

Adaptive Multi-scale Fusion. The stylized features from different scales are integrated into a unified representation. We employ an attention mechanism to dynamically determine the importance of each scale at every spatial location. First, a lightweight convolutional block $g_{att}(\cdot)$ generates an unnormalized importance map for each $\mathbf{F}_s$. These maps are normalized across the scale dimension via a spatial softmax to produce attention weights $\boldsymbol{\alpha}_s$:

$$\boldsymbol{\alpha}_s = \mathrm{Softmax}\big(g_{att}(\mathbf{F}_s)\big), \quad \mathbf{F}_s \in \mathbb{R}^{C_s\times H\times W} \tag{6}$$

The final, synthesized spectral feature $\mathbf{X}_{ms}$ is computed as the attentively weighted sum across all scales, followed by a 1×1 convolution $g_{fus}(\cdot)$ to seamlessly blend the contributions:

$$\mathbf{X}_{ms} = g_{fus}\left(\sum_{s\in S}\boldsymbol{\alpha}_s\mathbf{F}_s\right), \quad \mathbf{X}_{ms} \in \mathbb{R}^{C\times H\times W} \tag{7}$$

This adaptive fusion ensures that the most salient spectral characteristics, whether coarse or fine, dominate the final output, leading to both realistic and diverse spectral simulations.

2.1.3. Joint reconstruction and generation module

The final simulated sample $\mathbf{X}_G$ is synthesized by combining the spatially reorganized feature $\mathbf{X}_{ss}$ and the spectrally stylized feature $\mathbf{X}_{ms}$. The spectral features are first projected back to the spatial domain via transposed convolutions, then concatenated with the spatial features. A series of cascaded 3×3 convolutions followed by a sigmoid activation refine the fused representation to produce the output:

$$\mathbf{X}_G = \mathrm{Sigmoid}\Big(\mathrm{Conv}_{3\times 3}\big([\mathrm{Conv}_{ss}(\mathbf{X}_{ss}); \mathrm{Deconv}(\mathbf{X}_{ms})]\big)\Big) \tag{8}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. Through this joint spatial–spectral simulation, the SDSG generates diverse and realistic samples that effectively span the potential distribution of unseen target domains.

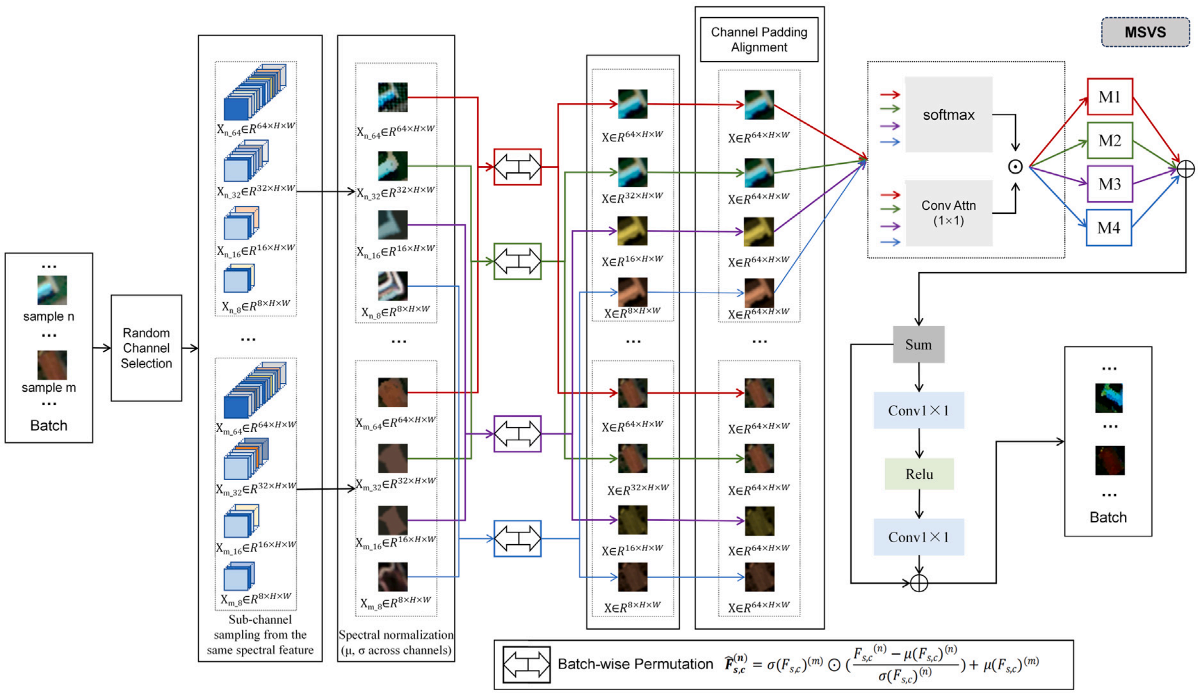


Fig. 3. Detailed structure of the MSVS. It emulates diverse spectral shifts by exchanging channel-wise mean and standard deviation across different scales.

2.2. Domain-disparity decorrelation discriminator

In cross-scene classification for HSI, the discriminator faces a dual challenge: it must maintain strong discriminative power for accurate classification while remaining robust to significant distribution shifts across unseen domains. To address this, we propose the D^3, which integrates MRCE with an explicit domain-disparity decorrelation constraint. As illustrated in Fig. 4, the D^3 is designed to learn representations that are sensitive to class-discriminative information yet invariant to domain-specific variations, thereby enhancing generalization without sacrificing discriminability. By combining multi-scale residual learning with a novel decorrelation objective, the D^3 effectively decouples domain-related discrepancies from the feature representation, enabling the model to preserve intra-domain discriminative attributes that are beneficial for classification.

2.2.1. Multiscale residual convolutional encoder

To effectively process both source and simulated samples, the MRCE employs a unified spectral–spatial encoder based on a three-stage multi-scale residual architecture, as shown in Table 1 This design captures the complex multi-scale characteristics of hyperspectral imagery, from fine-grained local details to high-level semantic contexts, through progressive feature extraction and cross-stage fusion.

MRCE proceeds in three sequential stages, each refining the representation at a different conceptual level. The first stage focuses on local spectral–spatial modeling. It uses 1×1 convolutions to compress spectral channels, mapping the input channels to a 128-dimensional space and reduce redundancy, followed by 3×3 convolutions to capture local spatial structures, further reducing the feature dimensionality to 64 while preserving the spatial resolution. On this basis, lightweight residual blocks are introduced to further refine the extracted features. Specifically, each residual block employs a bottleneck design comprising two 1×1 convolutional layers for channel compression and restoration. The first layer reduces the channel dimension from 64 to 16, and the second layer restores it back to 64, thereby reducing computational cost without compromising feature discriminability. This residual structure also facilitates gradient flow and enhances representational capacity.

The second stage enriches this representation by fusing the stage-1 features with the original input along the channel dimension, resulting in a concatenated channel size of $(inchannel{+}64)$. The fused features are then projected to 128 dimensions via a 1×1 convolution, followed by a 3×3 convolution and a residual block with the same structure as in the first stage. This skip connection ensures that discriminative low-level textures are preserved while higher-level semantics are developed.

The third stage, the output of the second stage is concatenated with the original input along the channel dimension. The resulting features are then processed by a 1×1 convolution followed by a 3×3 convolution. Unlike the previous stages, the third stage employs stride-based downsampling and padding-free convolutions to sharpen the spatial focus on central, semantically critical regions, thereby strengthening the model's attention to the most informative parts of the scene.

The transformation at each stage $k \in \{1, 2, 3\}$ can be formulated as:

$$\mathbf{f}_k = \mathcal{R}_k\big(\sigma(\mathbf{W}_k * \mathbf{Z}_k + \mathbf{b}_k)\big) \tag{9}$$

where $*$ denotes the 2D convolution operator, $\sigma(\cdot)$ is the non-linear activation function, $\mathcal{R}_k(\cdot)$ represents the residual unit, and $\mathbf{Z}_k$ is the input feature map to the kth stage.

The outputs from all stages are aggregated into a unified multi-scale representation through a two-step process: each stage's feature map f_k is first condensed via Global Average Pooling (GAP), and the resulting descriptors are then concatenated along the channel dimension. This is compactly expressed as:

$$\mathbf{f} = [\mathrm{GAP}(\mathbf{f}_1);\ \mathrm{GAP}(\mathbf{f}_2);\ \mathrm{GAP}(\mathbf{f}_3)] \tag{10}$$

where denotes the channel-wise concatenation operator across stages. The resulting representation $\mathbf{f}$ integrates hierarchical spectral–spatial patterns. Each stage produces a 64-dimensional vector, which are concatenated to form a 192-dimensional feature representation, enabling the discriminator to construct a feature space that is both discriminative and robust to domain variations.

2.2.2. Domain-disparity decorrelation constraint

To overcome the trade-off between strict domain invariance and intra-domain discriminability, we introduce a domain-disparity decorrelation constraint. This is particularly relevant for HSIs, which exhibit

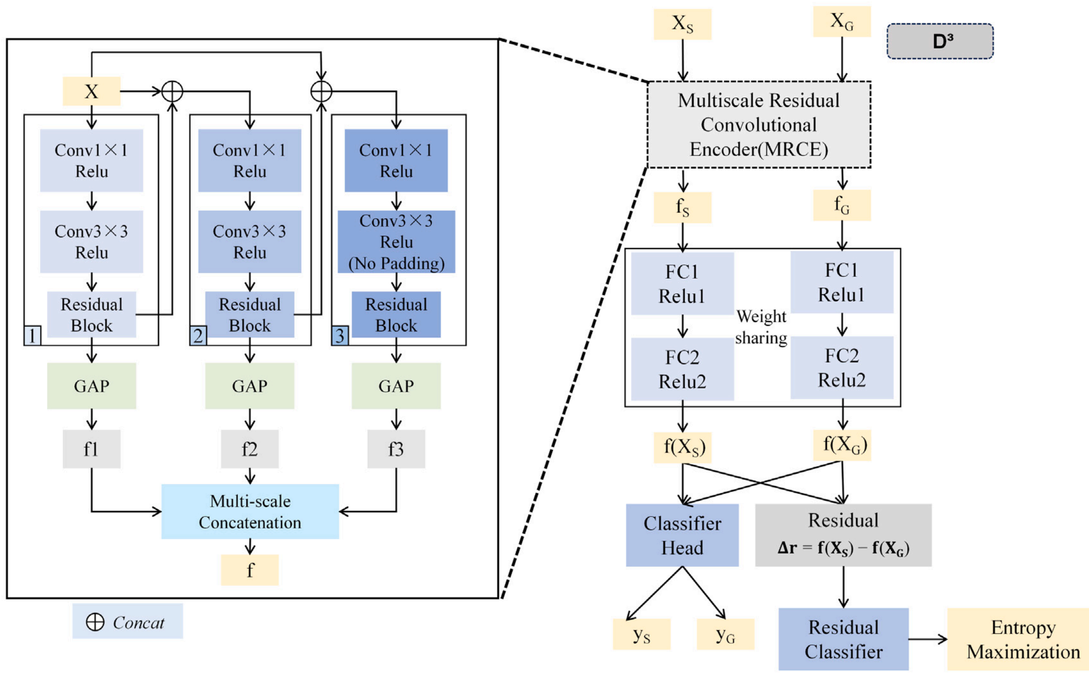


Fig. 4. Schematic diagram of the Domain-disparity Decorrelation Discriminator. The framework employs multi-scale residual blocks to enhance feature extraction capabilities and preserves intra-domain discriminative information beneficial for classification through decorrelation constraints.

Table 1
Structural configuration and parameter settings of each stage in MRCE.

Detailed Configuration of the MRCE Architecture							
Stage	Input channels	Operation	Kernel size	Output channels	Stride	Padding	Description
Stage 1	*inchannel*	Conv	1 × 1	128	1	0	Spectral compression
	128	Conv	3 × 3	64	1	1	Spatial modeling
	64	ResBlock	1 × 1 → 1 × 1	64	1	0	Bottleneck structure
Stage 2	*inchannel* + 64	Conv	1 × 1	128	1	0	Concatenated input
	128	Conv	3 × 3	64	1	1	Spatial modeling
	64	ResBlock	1 × 1 → 1 × 1	64	1	0	Bottleneck structure
Stage 3	*inchannel* + 64	Conv	1 × 1	128	1	0	Concatenated input
	128	Conv	3 × 3	64	1	0	spatial size reduction
	64	ResBlock	1 × 1 → 1 × 1	64	1	0	Bottleneck structure
Fusion	64 × 3	GAP	–	64	–	–	Stage-wise compression
	192	Concat	–	192	–	–	Multi-scale fusion

complex same object, different spectra and same spectra, different objects relationships. Conventional domain alignment methods typically pursue strict statistical consistency between source and target feature distributions. However, when applied to hyperspectral data, such rigid constraints often introduce a trade-off: in attempting to eliminate inter-domain discrepancies, the model may overlook subtle yet crucial discriminative features, leading to degraded intra-class discriminability. This module is designed to break this rigid alignment paradigm through a feature decoupling mechanism.

Let f_S and f_G denote the multi-scale feature representations of the source sample $\mathbf{X}_S$ and the generated sample $\mathbf{X}_G$ extracted by the encoder, respectively. Both features are further fed into a shared two-layer fully connected network with ReLU activations to obtain refined representations $f(\mathbf{X}_S)$ and $f(\mathbf{X}_G)$. The domain disparity is then explicitly modeled as the residual $\mathbf{r} = f(\mathbf{X}_S) - f(\mathbf{X}_G)$. Unlike alignment strategies that directly minimize numerical distances between features, we allow for the presence of reasonable discrepancies and model them as residual components that encode domain-specific information. To prevent this residual from carrying class-relevant information, we pass it through a lightweight auxiliary classifier to obtain a class-probability distribution $\mathbf{P}_r = D_c(\mathbf{r})$. The objective is to maximize the entropy of this distribution to further relax the consistency constraint:

$$H(\mathbf{r}) = -\sum_{c} \mathbf{P}_r \log\left(\mathbf{P}_r + \epsilon\right) \tag{11}$$

where ϵ is a small constant for numerical stability. During iterative training, optimize the decorrelation loss. From an information-theoretic perspective, maximizing the residual prediction entropy $H(r)$ enforces the removal of class-discriminative information from r, driving it to become a purely class-agnostic component. In the representation space, this mechanism can be interpreted as an orthogonalization constraint, where domain-specific variations are pushed into a complementary subspace that does not interfere with the discriminative feature space. Through this non-aggressive alignment strategy, the core semantic space is protected from the adverse effects of over-alignment, enabling effective reduction of inter-domain discrepancies while preserving well-defined class boundaries. This ultimately achieves a favorable balance between domain invariance and discriminability. The decorrelation loss is defined as the negative entropy averaged over a batch of N sample pairs:

$$\mathcal{L}_{\text{decorr}} = -\frac{1}{N}\sum_{i=1}^{N} H(\mathbf{r}_i) \tag{12}$$

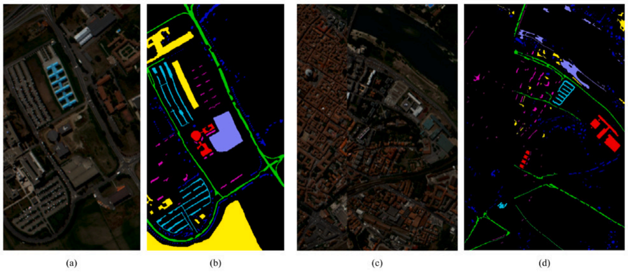


Fig. 5. Pseudo-color image and ground truth map of Pavia dataset. (a) Pseudo-color image of Pavia U. (b) Ground truth map of Pavia U. (c) Pseudo-color image of Pavia C. (d) Ground truth map of Pavia C.

Minimizing $\mathcal{L}_{\text{decorr}}$ encourages the residual r to be uninformative with respect to class labels, effectively pushing domain-specific variations into a subspace orthogonal to the discriminative feature directions. This allows the model to retain valuable intra-domain distinctions while still learning representations that generalize across domains.

2.3. Loss function

The overall objective of SD^{3}Net is to learn a discriminative yet domain-robust feature representation using only single-source supervision. The total loss function integrates a classification objective with the proposed domain-disparity decorrelation constraint, enabling the model to preserve intra-class discriminability while suppressing domain-sensitive variations.

To handle the class imbalance commonly present in hyperspectral datasets, we adopt a weighted cross-entropy loss. For a sample with ground truth, one-hot label y_i and predicted probability vector p_i from the discriminator's classification head $D_c(\cdot)$, the per-sample loss is:

$$\mathcal{L}_{\text{ce}}(\mathbf{p}_i, \mathbf{y}_i) = -\sum_c y_i^c \log p_i^c \tag{13}$$

Both source samples $\mathbf{X}_S$ and simulated samples $\mathbf{X}_G$ are used for classification training, ensuring that the model learns consistent categorical semantics across simulated domain shifts. The source-domain classification loss and the simulated-domain classification loss are respectively defined as:

$$\mathcal{L}_{\text{sd}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{ce}}(D_c(\mathbf{X}_S^i), \mathbf{y}_i) \tag{14}$$

$$\mathcal{L}_{\text{sim}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{ce}}(D_c(\mathbf{X}_G^i), \mathbf{y}_i) \tag{15}$$

The overall classification loss is a weighted combination of the two:

$$\mathcal{L}_{\text{cls}} = \lambda_1 \mathcal{L}_{\text{sd}} + \lambda_2 \mathcal{L}_{\text{sim}} \tag{16}$$

where λ_1, λ_2 balance the contribution of real and simulated samples.

The final loss for the discriminator is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_3 \mathcal{L}_{\text{decorr}} \tag{17}$$

where λ_3 controls the strength of the decorrelation regularization. During training, $\mathcal{L}_{\text{cls}}$ ensures basic discriminative performance on both source and simulated data, while $\mathcal{L}_{\text{decorr}}$ actively isolates domain-related variations from the classification-relevant feature subspace. This combined objective allows the model to maintain high discriminability within domains while achieving robust generalization across unseen target scenarios.

Table 2
Labeled sample statistics for Pavia data sets.

Color	Class Name	Source Domain (Pavia U)	Target Domain (Pavia C)
		Number of sample	
Blue	Trees	3064	7598
Green	Asphalt	6631	9248
Cyan	Bricks	3682	2685
Red	Bitumen	1330	7287
Magenta	Shadows	947	2863
Yellow	Meadow	18649	3090
Light purple	Bare soil	5029	6584
	Total	39332	39355

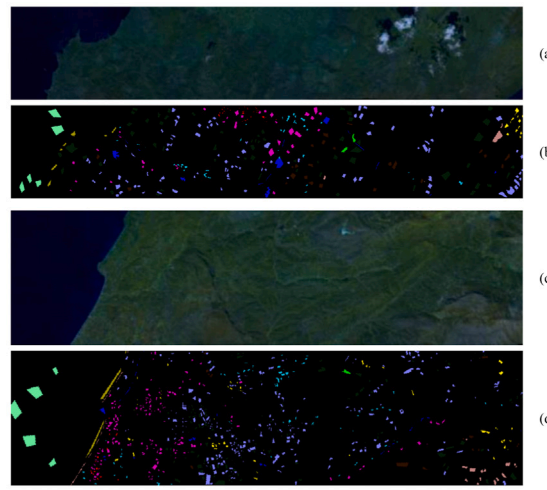


Fig. 6. Pseudo-color image and ground truth map of HyRANK dataset. (a) Pseudo-color image of Dioni. (b) Ground truth map of Dioni. (c) Pseudo-color image of Loukia. (d) Ground truth map of Loukia.

3. Experimental results and discussion

3.1. Datasets and evaluation metrics

3.1.1. Dataset description

To evaluate the proposed method under SSDG settings, we conduct systematic cross-scene HSI classification experiments on three public benchmarks: the Pavia, HyRANK, and WHU-Hi datasets. These datasets exhibit pronounced differences in sensors, spatial resolutions, spectral ranges, and land-cover categories, offering a rigorous testbed for assessing generalization capability across substantial domain shifts.

The Pavia dataset consists of two urban scenes acquired by the ROSIS sensor: Pavia University (PU) and Pavia Centre (PC). The spectral range is 430–860 nm. After removing noisy and water-absorption bands, 103 bands are retained for PU and 102 bands for PC. Both images have a spatial resolution of 1.3 m, with PU measuring 610 × 340 pixels and PC 1096 × 715 pixels. The detailed distribution of labeled samples for the source and target domains is provided in Table 2. Fig. 5 shows the corresponding false-color composites and ground-truth maps.

The HyRANK dataset contains two suburban/agricultural scenes from the Hyperion sensor: Dioni (250 × 1376 pixels) and Loukia (249 × 945 pixels). It comprises 176 spectral bands covering visible to short-wave infrared ranges, with a spatial resolution of 30 m. We adopt 12 common land-cover categories for evaluation, ensuring semantic consistency across domains. The labeled sample statistics are listed in Table 3, and the false-color images, along with the ground truth, are displayed in Fig. 6.

The WHU-Hi dataset includes two UAV-based scenes: HongHu (HH) and HanChuan (HC). HH has 940 × 475 pixels at 0.043 m resolution,

Table 3
Labeled sample statistics for HyRANK data sets.

Color	Class Name	Number of sample Source Domain (Dioni)	Target Domain (Loukia)
	Dense urban fabric	1262	206
	Mineral extraction sites	204	54
	Non-irrigated arable land	614	426
	Fruit trees	150	79
	Olive groves	1768	1107
	Coniferous forest	361	422
	Dense sclerophyllous vegetation	5035	2996
	Sparse sclerophyllous vegetation	6374	2361
	Sparsely vegetated areas	1754	399
	Rocks and bare soil	492	453
	Water bodies	1612	1393
	Coastal water	398	421
Total		20024	10317

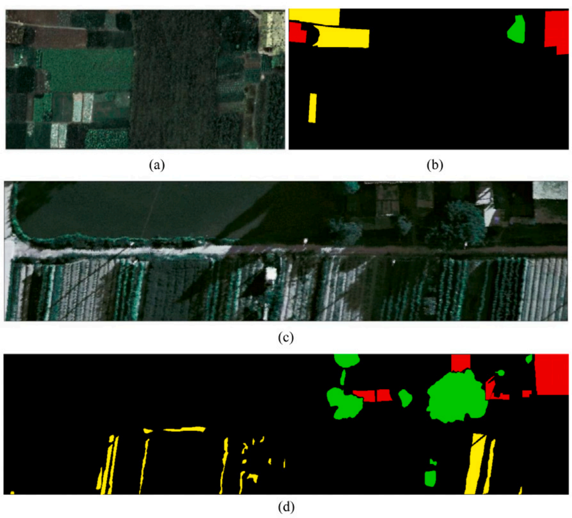


Fig. 7. Pseudo-color image and ground truth map of WHU dataset. (a) Pseudo-color image of HongHu. (b) Ground truth map of HongHu. (c) Pseudo-color image of HanChuan. (d) Ground truth map of HanChuan.

Table 4
Labeled sample statistics for WHU data sets.

Color	Class Name	Number of sample Source Domain (Honghu)	Target Domain (Hanchuan)
	Trees	4040	15558
	Bare soil	21821	9116
	Red roof	14041	10516
Total		39902	35190

while HC has 1217 × 303 pixels at 0.109 m resolution. Spectral coverage is 400–1000 nm. To align spectral dimensions, we use the first 270 bands from both scenes. Three dominant classes — tree cover, red roofs, and bare soil — are selected for cross-domain evaluation. The sample statistics are summarized in Table 4, and the corresponding false-color and ground-truth maps are presented in Fig. 7.

3.1.2. Evaluation metrics

To evaluate classification performance, we adopt Overall Accuracy (OA) and the Kappa coefficient as the primary quantitative metrics. OA measures the percentage of correctly classified pixels, while Kappa accounts for class imbalance and chance agreement, providing a more robust assessment of model consistency. For qualitative analysis, we visually compare the generated classification maps to assess spatial coherence, boundary sharpness, and structural preservation across different land-cover types under domain shift.

Table 5
Hyperparameter settings for different datasets.

Dataset	Patch size	λ_1	λ_2	λ_3
Pavia (Pavia U & Pavia C)	13× 13	0.1	1	1
HyRANK (Dioni & Loukia)	7× 7	0.1	0.1	1
WHU (Honghu & Hanchuan)	7× 7	0.1	0.3	1

3.2. Comparison methods and experimental setup

3.2.1. Comparison methods

To rigorously evaluate the proposed SD^{3}Net, we compare it with six recent state-of-the-art SSDG methods that represent the mainstream technical directions in HSI classification. These methods cover key paradigms, including generative augmentation, spectral–spatial decoupling, causal inference, style randomization, graph alignment, and feature decomposition, thereby forming a comprehensive and competitive benchmark for SSDG.

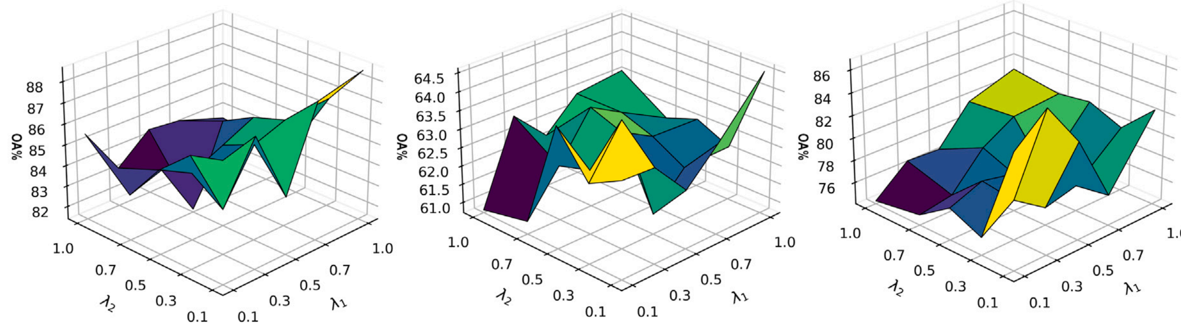
3.2.2. Experimental settings

We align the spectral dimensions between the source and target domains to ensure a fair cross-scene evaluation. For Pavia, Pavia University (PU) is trimmed to 102 bands to match Pavia Centre (PC). WHU-Hi datasets are aligned to 270 common bands. The HyRANK datasets retain their native 176 bands. During training, random flipping and radiometric noise are applied for augmentation, and spatial padding is used to maintain patch-size consistency.

All datasets follow an 80%–20% training–validation split. The SDSG and D^3 are optimized alternately in an adversarial manner: the D^3 is updated with the combined classification and decorrelation loss, followed by the SDSG using the corresponding generative objective. Training continues until validation accuracy converges, and the best model is selected for testing.

We use a learning rate of 1×10^{-3}, batch size 256, and train for 400 epochs across all experiments. Patch sizes and loss weights are tuned per dataset as shown in Table 5. All baselines are run with their officially reported settings. Results are averaged over 10 independent runs and reported as mean deviation. Experiments are performed on Ubuntu 22.04.5 LTS with PyTorch 2.4.1 and Python 3.8.20, using one NVIDIA RTX 3090 Ti GPU for acceleration.

We conduct a parameter sensitivity analysis on the essential hyperparameters and in the objective function, with quantitative results intuitively illustrated in the 3D surface plot (Fig. 8). Since the optimization depends primarily on the relative ratios among the loss components, we adopt the standard practice of fixing as an anchor to shrink the combinatorial search space. Subsequently, and are precisely adjusted within the range of [0.1,1] to evaluate their joint impact on the overall classification accuracy. The results show that the model maintains stable performance across a wide range of weight variations, and achieves the best performance under the parameter settings reported in Table 5.

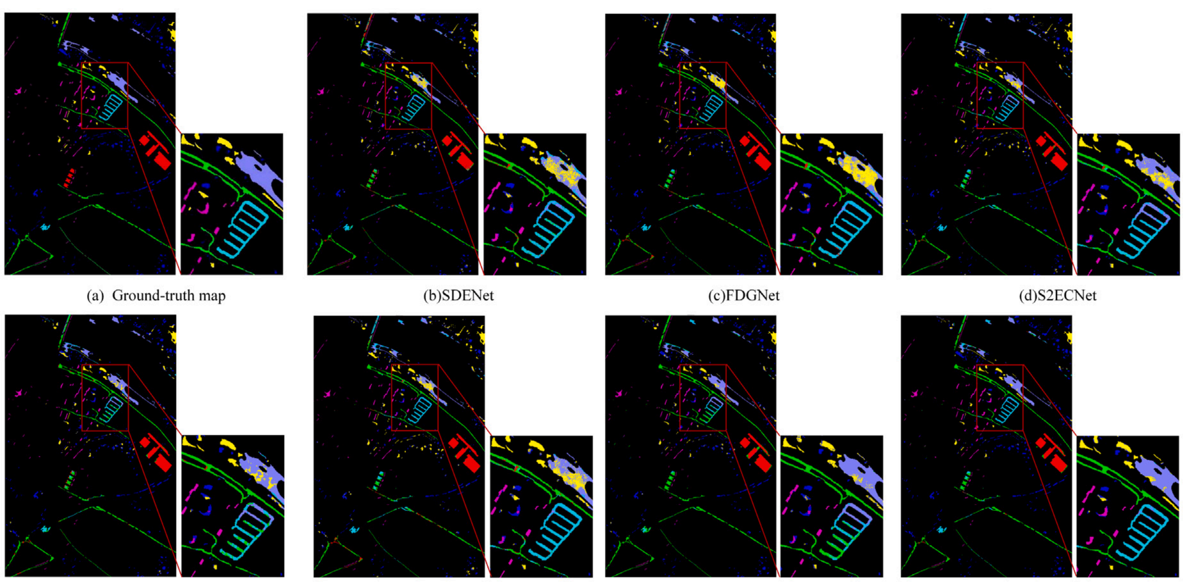


(a) Pavia (b) HyRANK (c) WHU

Fig. 8. Joint Sensitivity Analysis of λ_1 and λ_2 Across Multiple Datasets.

Table 6
Classification results on Pavia dataset (PU→PC).

Categories	SDENet	FDGNet	S2ECNet	LLURNet	TSTNet	NSGNet	SD^{3}Net
Trees	79.59	86.88	89.27	91.33	72.86	86.68	**97.28**
Asphalt	81.63	86.16	81.25	87.48	90.17	86.81	**91.44**
Bricks	65.14	86.44	46.15	81.08	56.05	70.99	**94.08**
Bitumen	84	84.60	82.01	**86.88**	67.61	79.36	82.53
Shadows	81.07	78.80	**98.29**	83.93	91.13	86.59	93.61
Meadow	79.09	77.18	79.09	70.61	**95.31**	78.06	77.02
Bare soil	80.92	65.16	81.79	86.32	69.84	74.88	**81.88**
OA	80.19	81.35	81.70	85.90	76.46	81.63	**88.52**
Kappa	76.25	77.69	77.99	83.01	71.17	77.93	**86.18**



(a) Ground-truth map (b)SDENet (c)FDGNet (d)S2ECNet

(e)LLURNet (f)TSTNet (g)NSGNet (h)SD³Net

Fig. 9. Ground truth and classification results of all methods on PaviaC.

3.3. Experimental results and analysis

Quantitative results are summarized in Tables 6–8, and the corresponding thematic maps are illustrated in Figs. 9–11.

In the cross-scene experiment on the Pavia dataset, SD^{3}Net achieves an OA of 88.52% and a Kappa coefficient of 86.18%, significantly outperforming all comparative methods in overall performance and exhibiting more stable classification results across multiple land-cover categories. In the Pavia dataset, images from urban areas and university scenes exhibit not only pronounced discrepancies in spatial structures but also notable variations in spectral responses induced by differences in atmospheric and illumination conditions. These factors make it challenging for methods that rely solely on feature alignment or simple data augmentation strategies to adequately capture the true target-domain distribution. By jointly simulating spatial structural variations and spectral distribution shifts through SALS and MSVS in the generation stage. The model obtains a richer set of potential target-domain samples. The introduced Domain-Disparity Decorrelation Constraint mechanism prevents excessive Suppression of domain disparities during feature learning, ensuring that classification-beneficial discriminative cues are not discarded due to an overemphasis on strict domain invariance. Consequently, the proposed method is able to simultaneously maintain sufficient distribution coverage and discriminative feature

Table 7
Classification results on HyRANK dataset (Dioni→Loukia).

Categories	SDENet	FDGNet	S2ECNet	LLURNet	TSTNet	NSGNet	SD^{3}Net
Dense urban fabric	28.16	14.08	15.05	**38.35**	24.76	11.65	29.61
Mineral extraction sites	**100**	79.63	98.15	1.85	11.11	31.48	18.52
Non-irrigated arable land	21.60	0	4.46	46.01	27.23	45.31	**48.59**
Fruit trees	1.27	0	0	15.19	**24.05**	7.59	1.27
Olive groves	**29.45**	3.52	22.22	0.27	20.23	8.13	0.00
Coniferous forest	**40.28**	36.73	21.80	13.74	2.37	26.78	34.6
Dense sclerophyllous vegetation	69.63	72.13	**87.52**	79.77	77.77	73.36	78.07
Sparse sclerophyllous vegetation	58.24	69.42	58.11	65.31	73.06	**78.53**	74.12
Sparsely vegetated areas	75.94	**82.43**	66.67	72.68	40.85	39.85	20.3
Rocks and bare soil	39.96	0	2.43	39.07	0.44	3.75	**55.41**
Water bodies	99.93	100	100	100	100	100	**100**
Coastal water	100	100	82.9	100	100	100	**100**
OA	62.61	60.18	62.56	63.60	62.62	62.86	**64.55**
Kappa	54.69	50.75	53.28	54.77	53.19	53.77	**55.56**

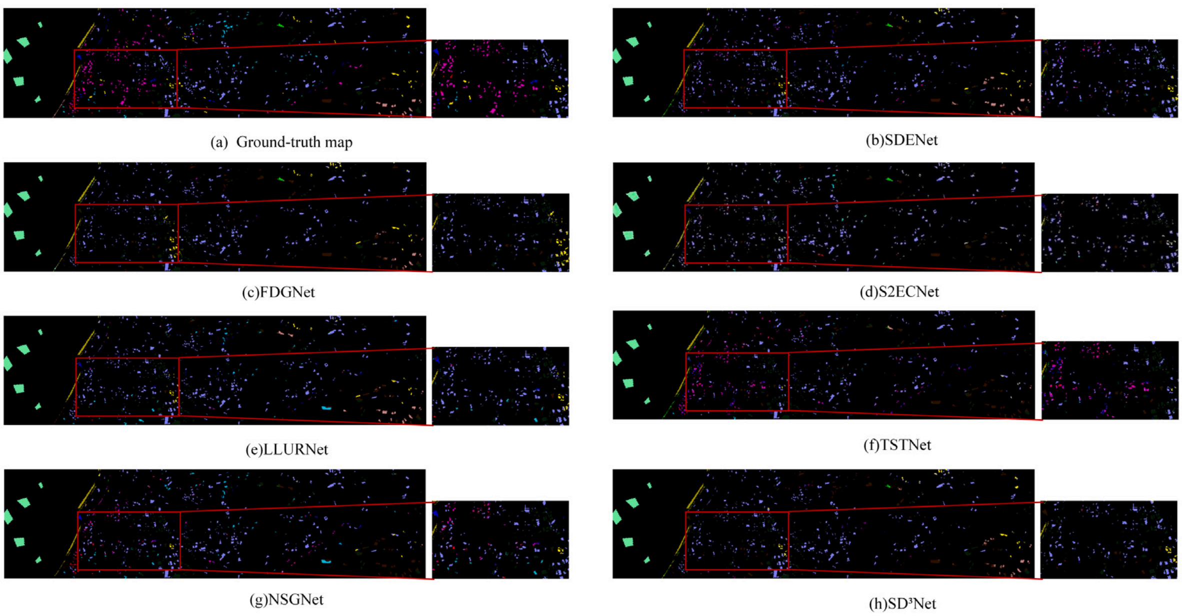


Fig. 10. Ground truth and classification results of all methods on the Loukia.

integrity under cross-urban scenarios, leading to more reliable domain generalization performance.

On the HyRANK dataset, SD^{3}Net attains an OA of 64.55% and a Kappa coefficient of 55.56%. This dataset comprises wide-band hyperspectral imagery covering the visible to short-wave infrared spectrum, where the uneven spectral discriminability among classes results in imbalanced inter-class separability. Most baseline methods suffer from severe performance degradation — even reaching zero accuracy in certain categories (e.g., Classes 3 and 10) — our network maintains competitive advantages of 48.59% and 55.41%, respectively. This resilience is credited to the MSVS. By exchanging multi-scale channel statistics, MSVS systematically simulates spectral shifts caused by sensors and atmospheric conditions. This "structured simulation" strategy covers potential target domain distributions more effectively than simple feature style transfers. This allows the model to retain the fine-grained spectral information needed to distinguish similar objects even under significant distribution shifts, achieving more reliable generalization in harsh conditions. As a result, SD^{3}Net significantly outperforms comparative methods on multiple challenging classes, demonstrating stronger generalization ability. However, compared with the other two datasets, the OA and Kappa on the HyRANK dataset are relatively lower. Through comparative analysis, we observe that HyRANK differs substantially from datasets such as Pavia or Houston. Specifically, it exhibits a pronounced long-tail distribution (i.e., severe class imbalance) in land-cover categories. Since single-source domain generalization methods rely heavily on prior knowledge extracted from a single source domain, extreme discrepancies in class composition and imaging conditions between the source and target domains impose inherent limitations on the coverage of simulated domains. Nevertheless, although the performance gain of SD^{3}Net on this dataset is smaller than that on the other datasets, its classification accuracy still significantly surpasses that of existing baseline models. This demonstrates that even in highly complex scenarios, the structured simulation strategy of SD^{3}Net can capture more robust features than generic augmentation methods.

In the WHU cross-scene experiment, SD^{3}Net exhibits the most pronounced overall advantage, achieving an OA of 86.67% and a Kappa coefficient of 79.46%, and maintaining a clear lead over all comparative methods. In the WHU dataset, classes such as Red roof and Bare soil exhibit severe spectral overlap, which, together with scene-induced spatial structural variations, makes cross-domain classification particularly challenging. By jointly generating samples through spatial structure simulation and spectral distribution simulation, SD^{3}Net effectively expands the coverage of coupled spatial–spectral variations in the potential target domain. Nevertheless, distribution simulation alone is insufficient to resolve the ambiguity among highly confused classes, as overly strong domain alignment tends to weaken the modeling of

Table 8
Classification results on WHU dataset (HH→HC).

Categories	SDENet	FDGNet	S2ECNet	LLURNet	TSTNet	NSGNet	SD^{3}Net
Trees	92.89	95.47	**95.48**	95.36	94.89	93.67	92.58
Bare soil	**94.10**	82.72	85.23	80.91	78.67	83.33	89.53
Red roof	35.71	34.55	35.80	33.85	47.66	30.26	**75.48**
OA	77.19	75.35	76.30	74.66	77.76	76.43	**86.67**
Kappa	64.71	61.30	62.84	60.20	64.95	62.96	**79.46**

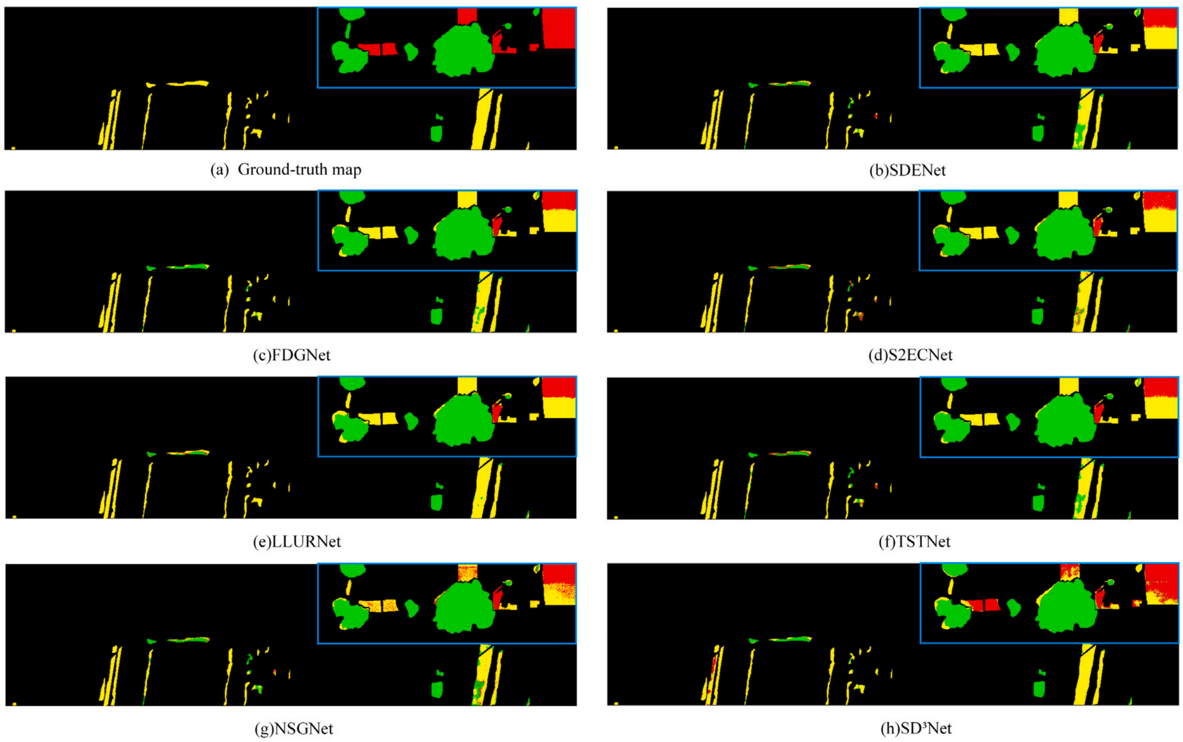


Fig. 11. Ground truth and classification results of all methods on the Hanchuan.

fine-grained class differences. The Domain-disparity Decorrelation Constraint mechanism explicitly mitigates the coupling between domain-related variations and class-discriminative factors, preventing features that are critical for distinguishing similar classes from being excessively suppressed. Consequently, the proposed method maintains sensitivity to key discriminative information under severe spectral confusion, leading to significant accuracy improvements on classes such as Red roof.

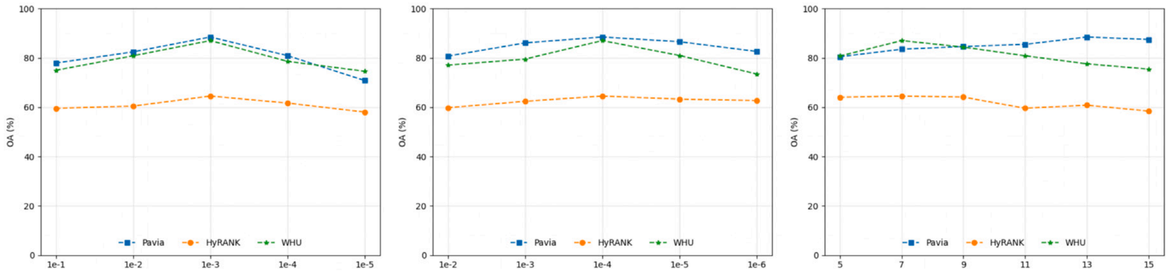
These results indicate that SD^{3}Net simultaneously achieves effective adaptation to cross-scene distribution shifts and stable preservation of discriminative capability under the single-source setting, thereby delivering consistent and reliable domain generalization performance across different datasets.

3.4. Parameter analysis

To assess the stability and robustness of SD^{3}Net under various configurations, a systematic sensitivity analysis is conducted regarding the base learning rate η, the regularization coefficient λ, and the spatial patch size. As illustrated in Fig. 12, the proposed network demonstrates consistent performance trends across the Pavia, HyRANK, and WHU datasets.

The base learning rate is pivotal for balancing the adversarial optimization between the SDSG and the discriminative classifier. Fig. 12(a) reveals that the classification accuracy across all target scenes reaches its peak at $\eta = 10^{-3}$. An excessively small learning rate leads to insufficient convergence, hindering the generator from adequately capturing the distribution expansion information introduced by the spatial and spectral simulation modules. Conversely, a large learning rate disrupts the dynamic equilibrium between the generator and the discriminator. In particular, excessive gradient updates render the categorical entropy constraint on residual components unstable, leading to a significant degradation in classification accuracy. For the regularization parameter λ, as shown in Fig. 12(b), the optimal value for all three datasets is 10^{-4}. When λ is below this threshold, the efficacy of the Domain-disparity decorrelation is diminished, causing the model to converge toward traditional domain generalization paradigms that rely on strict feature invariance. This tendency risks suppressing discriminative intra-domain attributes beneficial for classification. On the other hand, an excessively large λ imposes an over-stringent category-independence constraint on the residual components, which may interfere with the extraction of discriminative spatial–spectral features and subsequently weaken the overall classification capability.

The selection of spatial patch size directly influences the capacity of the SALS to model local contextual relationships and determines the efficiency of the discriminator in joint spatial–spectral feature extraction. Fig. 12(c) demonstrates that the optimal patch size varies according to the inherent spatial characteristics of the datasets. For datasets with coarser spatial resolution, such as HyRANK, a moderate patch size is sufficient to capture meaningful neighborhood context, since each pixel already covers a relatively large ground area, and excessively large patches tend to introduce redundant background information. In contrast, for high-resolution datasets, including Pavia and WHU, larger spatial patches are more beneficial as they encompass richer



(a) Base learning rate (b) Regularization parameter (c) Patch size

Fig. 12. Sensitivity analysis of SD^{3}Net with respect to (a) base learning rate, (b) regularization coefficient, and (c) spatial patch size on the Pavia, HyRANK, and WHU datasets.

Table 9
Ablation study of the SALS, MSVS, and domain-disparity decorrelation constraint modules for three cross-scene hyperspectral image classification tasks.

	Generator		Discriminator	PAVIA	HYRANK	WHU
Class	SALS	MSVS	Domain-disparity Decorrelation Constraint	OA	OA	OA
1	×	✓	✓	87.88±0.90	63.85±0.47	84.72±1.25
2	✓	×	✓	87.01±0.82	63.48±0.42	85.09±1.41
3	✓	✓	×	85.03±0.99	63.16±0.34	78.10±1.77
4	✓	✓	✓	**88.52±1.03**	**64.55±0.92**	**86.67±0.47**

fine-grained structural details and object-level spatial relationships. In particular, for the Pavia dataset with complex urban layouts, a patch size of 13 × 13 provides a favorable balance between contextual coverage and noise suppression, enabling spatial structure perturbations to more accurately reflect real-world land-cover distributions and yielding optimal accuracy. However, for the WHU dataset with extremely high spatial resolution, overly large patches may introduce excessive local variability and background clutter, which can hinder the extraction of discriminative representations. These results suggest that tailoring the patch size to the inherent spatial scale of disparate scenes is crucial for fully leveraging the distribution simulation mechanism. In summary, the sensitivity analysis presented in Fig. 12 empirically validates the stability and robustness of the proposed distribution simulation and Domain-disparity decorrelation strategies. The balanced performance across a wide range of settings underscores that the effectiveness of the proposed framework stems from its integrated design in semantic-anchored layout, multi-scale spectral variation, and domain-disparity decorrelation.

3.5. Ablation experiment

This section verifies the necessity of the SALS, MSVS, and Domain Difference Decoupling (DDD) constraint through an ablation study. We tested the classification results under the following conditions: (1) without SALS, (2) without MSVS, (3) without DDD, and (4) the complete framework (SD^{3}Net). The experimental results are summarized in Table 9, showing the OA performance of each module across three cross-scene tasks.

The ablation results show that the complete framework achieves the best performance across all scenarios, validating the synergistic effects of the components in handling the coupled relationship between ground object structure and spectral response.

When the SALS is removed from the generator, the OA decreases by 0.64%, 0.70%, and 1.95% on the Pavia, HyRANK, and WHU datasets, respectively. In the absence of SALS, the model relies primarily on spectral-level transformations while failing to account for variations in spatial contextual layouts. By contrast, the introduction of semantically guided spatial reorganization transforms indiscriminate random perturbations into physically meaningful, structured simulations, ensuring topological consistency in the spatial configurations of generated samples. This mechanism facilitates the learning of more robust spatio-spectral representations and significantly enhances spatial generalization capability in unseen target domains.

Removing the MSVS leads to a pronounced degradation in classification performance, with OA values dropping to 87.01%, 63.48%, and 85.09% on the Pavia, HyRANK, and WHU datasets, respectively. Given the sensitivity of hyperspectral imagery to spectral distortions induced by sensor discrepancies and atmospheric conditions, the absence of the multi-scale statistical stylization mechanism substantially limits the model's ability to simulate dynamic spectral response variations. Consequently, the extracted features exhibit reduced robustness across spectral dimensions, resulting in inferior cross-domain generalization performance.

The most severe performance decline is observed when the Domain-disparity Decorrelation Constraint is removed from the discriminator. On the WHU dataset, the OA decreases from 86.67% to 78.10%, corresponding to a nearly 9% performance drop, with similarly significant degradation observed on the Pavia and HyRANK datasets. This result highlights the critical role of the Domain-disparity Decorrelation Constraint in decoupling domain-specific variations, allowing the model to preserve fine-grained discriminative information while enforcing cross-domain feature consistency.

Overall, the results across the three cross-scene experiments indicate that the performance improvements of SD^{3}Net do not stem from the isolated contribution of any single module, but rather from the synergy between the distribution simulation mechanism and the domain disparity decoupling strategy at the level of feature learning objectives. SALS and MSVS, along with their core sub-mechanisms, effectively expand the source-domain distribution, enabling the model to cover complex and diverse target-domain variations, while Domain-disparity Decorrelation Constraint prevents an excessive pursuit of strict domain invariance and ensures the preservation of domain-related differences that are beneficial for classification. This tripartite collaboration enables the model to more effectively approximate the underlying target-domain distribution, thereby substantially improving cross-scene classification reliability in unseen environments.

Table 10
Results of time complexity and parameter analysis.

Time Complexity and Parameter Analysis							
Categories	SDENet	FDGNet	S2ECNet	LLURNet	TSTNet	NSGNet	SD^{3}Net
Training Time (s)	9836.23	10 116.57	138 924.42	9260.39	22 500.85	840.78	806.67
Inference Time (s)	14.73	14.38	166.59	15.05	21.77	14.33	14.28
Parameters (M)	2.53	4.23	2.27	0.60	7.80	0.67	2.61

3.6. Time complexity and computational cost analysis

To comprehensively evaluate the deployment potential of SD^{3}Net in practical remote sensing scenarios, we further conduct a computational complexity analysis on the Pavia University dataset. The evaluation metrics include Training Time, Inference Time, and the number of model parameters. SD^{3}Net is compared with several mainstream domain generalization methods, including SDENet, FDGNet, S2ECNet, LLURNet, TSTNet, and NSGNet, and the results are reported in Table 10.

From the perspectives of Training Time and Inference Time, SD^{3}Net achieves 806.67 s and 14.28 s, respectively, demonstrating the best computational efficiency among all compared methods. This advantage mainly stems from the efficient structured image generation mechanism adopted in SALS and MSVS, which enhances the diversity of feature perturbations while promoting faster model convergence, thereby effectively reducing training cost. Meanwhile, these two modules are only active during training and are removed during inference, introducing no additional computational burden at test time. This further ensures low inference time, enabling the model to maintain strong generalization capability while achieving higher operational efficiency.

In terms of parameter scale, SD^{3}Net contains 2.61 M parameters, which is lower than FDGNet (4.23 M) and TSTNet (7.80 M), and only higher than S2ECNet, LLURNet, and NSGNet. Although SALS and MSVS are introduced, the additional parameter growth is effectively controlled, demonstrating that SD^{3}Net achieves superior performance while maintaining relatively low model complexity, thereby striking a favorable balance between accuracy and lightweight design.

3.7. Limitations analysis

Despite the excellent performance of SD^{3}Net in the task, inherent generalization limitations of single-source domain generalization still remain in practical deployment. The SALS and MSVS modules perform structured simulation of spatial and spectral distributions based on source-domain prior knowledge. However, under extreme scenarios — such as recognition tasks across long temporal gaps or highly heterogeneous conditions involving sensors with substantial spatial resolution discrepancies — the performance gains obtained by expanding source-domain priors may diminish. In addition, the performance observed on HyRANK indicates that, in more challenging scenarios, severe class imbalance, long-tailed distributions, and highly heterogeneous imaging conditions may affect the scalability of the system. These scenarios reflect a fundamental limitation of single-source generalization: when domain shifts exceed the scope of physical priors, its generalization capability may be constrained.

To further improve model performance under these scenarios, future research will explore integrating generalized prior knowledge from multimodal large models into the proposed framework [49]. Such models can provide richer and more transferable representations, thereby compensating for the lack of diversity in a single data source. Moreover, to address sensitivity issues related to specific datasets, we plan to adopt self-supervised pretraining strategies to further optimize feature extraction under complex conditions. By combining the structured simulation advantages of SD^{3}Net with broader global generalization prior knowledge, it is expected that the robustness and universality of hyperspectral image classification systems can be fundamentally enhanced to meet the demands of complex real-world application scenarios.

4. Conclusion

In this article,a novel SSDG framework, termed SD^{3}Net, is proposed to address the pronounced data distribution shifts in cross-scene HSI classification. By synergistically integrating the Structured Data Simulation Generator with the Domain-Disparity Decorrelation strategy, the proposed framework systematically characterizes structured cross-domain variations while relaxing the assumption of strict feature invariance. Experimental results across multiple challenging hyperspectral datasets demonstrate that our approach effectively approximates the underlying data distribution of unseen domains and preserves essential intra-domain discriminative attributes, achieving superior generalization performance without requiring prior information from the target domain. Future research will explore the integration of physical model-based priors with deep learning architectures, specifically focusing on dictionary unfolding mechanisms. By formulating hyperspectral SSDG as a sparse coding problem within a deep unfolding framework, we aim to achieve more interpretable and robust feature disentanglement, thereby further enhancing the reliability of hyperspectral interpretation in complex and unobserved environments.

CRediT authorship contribution statement

Haotian Hu: Writing – original draft, Software, Formal analysis, Conceptualization. **Yunpeng Zheng:** Writing – original draft, Software, Formal analysis, Conceptualization. **Qian Liu:** Writing – review & editing, Validation, Methodology, Funding acquisition. **Shuai Yang:** Writing – review & editing, Funding acquisition. **Biqi Wang:** Writing – review & editing. **Xiaohui Yuan:** Writing – review & editing. **Lichuan Gu:** Writing – review & editing, Funding acquisition.

Acknowledgments

This work is supported by the Natural Science Foundation of Anhui Province under Grants 2308085QF202 and 2308085MF217, the National Natural Science Foundation of China under Grants 62501015, 62306008 and 62301006, the Natural Science Research Project of Education Department of Anhui Province of China under Grant 202 4AH050449, the Key Project of Anhui Province's Science and Technology Innovation Tackle Plan under Grant 202423k09020040.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Module-wise ablation experiment analysis

To validate the independent contributions and necessity of each core sub-module, we further design three sets of ablation experiments: (1) a sample selection ratio ablation for SALS, (2) a multi-scale grouping ablation for the MSVS module, and (3) an entropy maximization constraint ablation for D^3. Such analyses are essential for disentangling and understanding the internal contribution mechanisms of SD^{3}Net.

It can be observed from Table A.11 that when the sample selection ratio in SALS varies, the overall impact on OA remains relatively stable. The current threshold of 40% is designed to balance sample

Table A.11
Comparative experiment with Percentile Threshold.

Comparative Experiment (Percentile Threshold) OA/Kappa					
Dataset	0.2	0.4	0.6	0.8	1.0
Pavia	88.31/85.95	88.52/86.18	86.97/84.33	85.92/83.07	87.75/84.85
HyRANK	61.39/51.60	64.55/55.56	62.48/53.39	63.50/54.33	62.47/53.49
WHU	77.20/63.89	86.67/79.46	76.37/62.17	76.54/63.41	84.26/74.92

Table A.12
Ablation experiment on spatial multiscale .

Ablation Experiment (Spatial Multiscale) OA/Kappa					
Dataset	Full Scale 64	64+32	64+32+16	64+32+16+8	64+32+16+8+4
Pavia	86.02/81.77	86.28/83.24	86.20/83.13	88.52/86.18	86.61/83.92
HyRANK	62.15/52.67	61.13/51.47	62.74/53.14	64.55/55.56	62.01/51.93
WHU	84.30/75.20	76.14/63.07	79.76/68.11	86.67/79.46	80.41/69.06

Table A.13
Ablation experiment on the discriminator.

Ablation Experiment (Discriminator) OA/Kappa		
Dataset	Original discriminator	Remove entropy loss
Pavia	88.52/86.18	84.03/80.91
HyRANK	64.55/55.56	61.37/51.54
WHU	86.67/79.46	75.92/62.51

diversity and physical plausibility. In hyperspectral imagery, the spatial correlation between central objects and their surrounding background constitutes an important structural prior. Blindly exchanging all samples would introduce a large number of noisy samples with disordered spatial logic and weak physical meaning. To address this, we compute and rank the feature differences between the center and background, selecting only the top 40% of samples with the most significant discrepancies for swapping. These experiments further demonstrate that this ratio most effectively simulates realistic object-boundary variations while avoiding training interference caused by low-quality recomposed images.

It can be observed from Table A.12 that when the multi-scale grouping strategy in MSVS is removed, the results show consistent performance degradation across all three cross-scene tasks. Specifically, the OA drops by approximately 2 percentage points on the Pavia and HyRANK datasets, and by around 5 percentage points on the WHU dataset. This indicates that, compared to conventional single-channel style transformations, the proposed multi-scale grouping strategy more comprehensively captures nonlinear distortions across different spectral bands, making it indispensable for modeling complex spectral domain shifts.

It can be observed from Table A.13 that when the entropy maximization constraint in the D^3 module is removed, the model performance significantly declines across all three cross-scene tasks. In particular, the OA decreases by 4.49 percentage points on the Pavia dataset, 3.18 percentage points on HyRANK, and 10.75 percentage points on WHU. These results verify that maximizing residual entropy in the discriminator plays a critical role in relaxing over-restrictive alignment constraints and preventing the collapse of discriminative semantics, thereby serving as the key driving force for effective feature decoupling.

Appendix B. Spectral visualization analysis

To intuitively verify whether the proposed Multi-Scale Spectral Variation Simulation (MSVS) module follows the physical principles of hyperspectral data while generating diversified samples, we conducted a visual comparison of spectral curves before and after transformation. Specifically, we extracted the original sample together with its transformed counterparts under four different scales ($s = 64, 32, 16, 8$), and mapped them back into the spectral domain for curve plotting. As shown in Fig. B.13, although MSVS introduces a channel grouping strategy, the spectral curves at different scales, after being restored to the original dimensionality (Fig. B.13, green, brown, and blue curves), maintain an overall shape that highly agrees with the original spectral envelope across full channels (red curve). No unreasonable band mutations or artifacts are observed at key absorption peaks or abrupt reflectance transition regions. This phenomenon confirms that MSVS does not arbitrarily perturb isolated bands, but instead preserves the intrinsic correlation of the spectral sequence through its multi-scale grouping strategy. Therefore, while effectively expanding the diversity of the data distribution, the proposed model strictly maintains the physical continuity of spectral characteristics.

Within the multi-scale fusion mechanism, the consistency of spectral curves across different scales further demonstrates that the proposed simulation strategy successfully avoids the physical mismatch problem often caused by conventional random perturbation methods, thereby laying a solid foundation for improving generalization ability in cross-scene tasks.

Appendix C. T-SNE visualization

To intuitively demonstrate how SD^3Net achieves cross-domain invariance while preserving discriminative semantics, we employ t-SNE to visualize the feature distribution of the Pavia dataset (as shown in Fig. C.14). In particular, in panels (a), (c), and (e), · denotes the source domain, while in panels (b), (d), and (f), · denotes the target domain and ×denotes the generated domain; different colors represent different classes.

As observed from Figs. C.14(a) and (b), the proposed structured variation generation strategy can effectively expand the source-domain data distribution, enabling generated-domain samples to cover the target-domain distribution range while preserving the semantic structure of the source domain. Meanwhile, while enlarging the distribution support range, the generated samples still maintain strong intra-class proximity and aggregation trends with the corresponding source-domain samples. This indicates that the generated samples produced by the proposed model not only exhibit higher diversity and can cover potential target-domain scenario variations, but also possess validity and semantic consistency.

By comparing Fig. C.14(a), (b) with Fig. C.14(c)- (f), it can be observed that class separability is significantly improved after being processed by the SD^3Net discriminator. Furthermore, by comparing Fig. C.14(c), (d) and Fig. C.14(e), (f), it can be found that, with the introduction of the D^3 module in Fig. C.14(c), (d), the source-domain, generated-domain, and target-domain features all exhibit higher intra-class compactness and clearer inter-class boundaries, and their clustering quality is evidently superior to that in Fig. C.14(e), (f) without the D^3 module. This demonstrates that the D^3 module can effectively

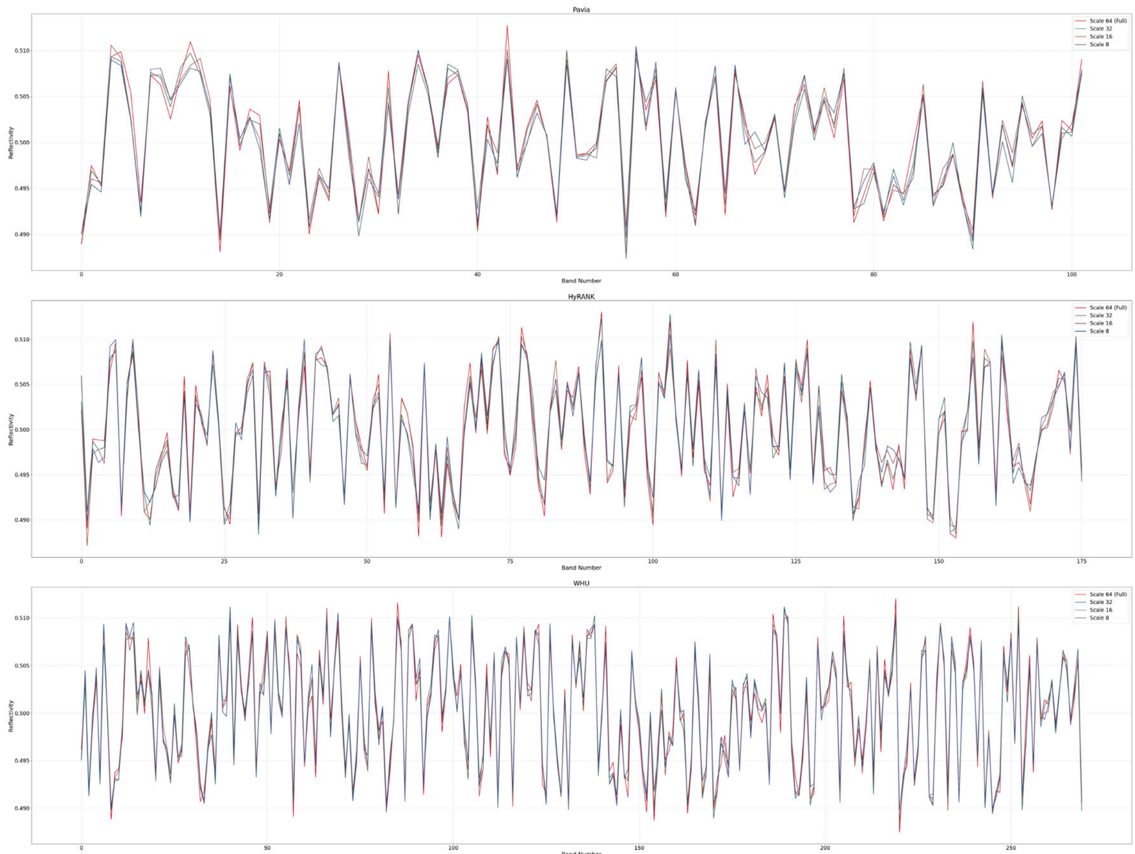


Fig. B.13. (a) illustrates the spectral curves across all channels for the Pavia dataset, (b) shows those for the HyRANK dataset, and (c) presents the corresponding curves for the WHU dataset. In each figure, the red curve represents the spectral profile of the original image, the green curve corresponds to the 32-channel representation, the brown curve to the 16-channel representation, and the blue curve to the 8-channel representation.

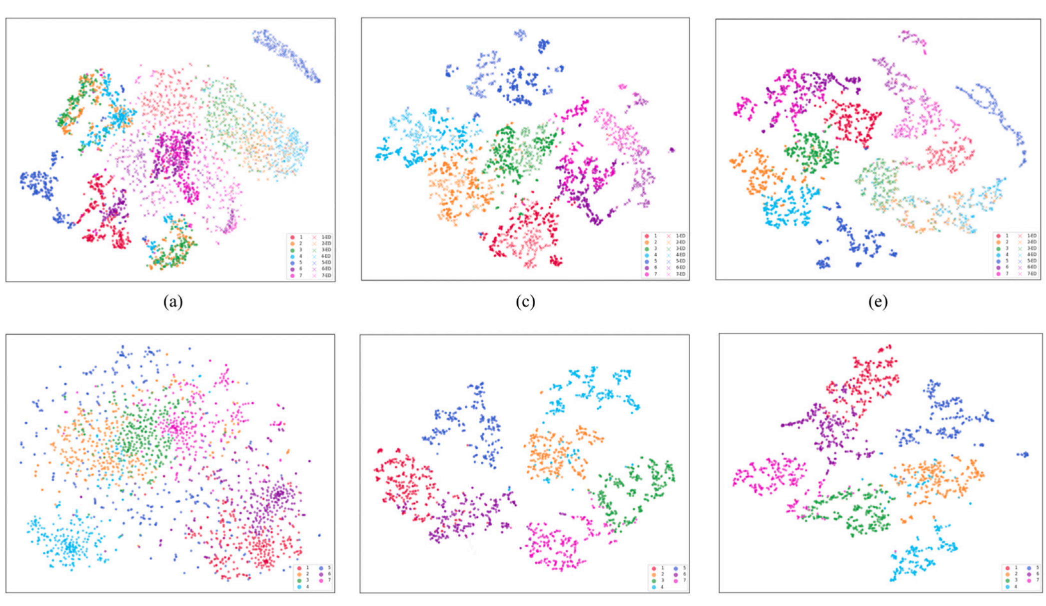


Fig. C.14. The above figure presents the t-SNE visualization results for the Pavia dataset, where (a) shows the original samples from SD and ED, and (b) shows the original samples from TD. (c) and (e) represent the SD features and ED features, while (d) and (f) represent the TD features. Specifically, (c) and (d) illustrate the t-SNE visualization results with the D^3 module applied, whereas (e) and (f) correspond to the results without the D^3 module.

suppress the influence of domain-specific interference on the discriminative semantic space, promote more reasonable feature alignment between the generated and real domains, enhance the discriminative capability of feature representations, and achieve more effective feature decoupling.

The results show that the structured generation mechanism proposed in SD^{3}Net can effectively expand the data distribution and cover potential target-domain variations while preserving intra-class semantic consistency; meanwhile, the D^3 module further improves the intra-class clustering quality and inter-domain separability of cross-domain features, jointly validating that the proposed method achieves domain invariance while preserving discriminative capability.

Appendix D. Theoretical derivation of D^3

Given the dataset $D_{\text{train}} = \{x \in X_{\text{train}}, y \in Y_{\text{train}}\}$, where X_{train} and Y_{train} denote the training image set and label set, respectively, we need to establish the mapping relationship between the data X_{train} and the ground-truth labels Y_{train}. Under the classical empirical risk minimization (ERM) framework, the training objective is to select a hypothesis $h : X \to Y$ from a predefined hypothesis space H, such that the empirical risk on D_{train} is minimized:

$$\inf_{h\in H} = E_{(x,y\sim D_{\text{train}})}[L(h(x), y)] \tag{D.1}$$

However, when the model is deployed on the test set D_{test}, due to possible domain shift between the training and test sets, i.e., $P(X_{\text{train}}) \neq P(X_{\text{test}})$, the model performance often degrades. To improve generalization ability, we need to train the model to be invariant to these domain-specific attributes.

For the source-domain feature defined as f_X, and the generated-domain feature as f_G, the residual component is defined as $r = f_X - f_G$ For the residual component r, we attempt to maximize the uncertainty of predictions made by r, such that it does not contain too much classification-related information. Since entropy can be regarded as a measure of predictive uncertainty, we maximize the conditional entropy $H(y|r)$ to amplify the predictive uncertainty of r. Therefore, our objective is:

$$L_{\text{rec}} = -H(y|r) = -H[\text{Softmax}(F(r))] \tag{D.2}$$

(1) Discriminability Preservation

In previous consistency regularization methods, the similarity between the class representation in classifier F and the source-domain feature is computed as follows:

$$\hat{y}_X^i = sim(p_i, f_X),\ \hat{y}_G^i = sim(p_i, f_G) \tag{D.3}$$

where $sim(\cdot)$ is a similarity function constructed by inner product, p_i is the feature of the ith class, and $\hat{y}_o^i$ and $\hat{y}_a^i$ denote the probabilities belonging to the ith class representation.

Then we can obtain:

$$\begin{aligned}\hat{y}_G^i &= sim(p_i, f_X - r)\\ &= sim(p_i, f_X) - sim(p_i, r)\\ &= \hat{y}_X^i - \hat{y}_r^i\end{aligned} \tag{D.4}$$

When the conditional entropy $H(y|r)$ is maximized, the residual component is assigned equal probabilities to every class, namely$\hat{y}_r^1 = \hat{y}_r^2 = \hat{y}_r^3 = \cdots = \hat{y}_r^C = K$Therefore, the relationship between $\hat{y}_G^i$ and $\hat{y}_X^i$ is:

$$\hat{y}_G^i = \hat{y}_X^i - K = f(\hat{y}_X^i, K) \tag{D.5}$$

Within the same iteration, K is a constant, and f is an order-preserving mapping function, which guarantees that if$\hat{y}_X^j > \hat{y}_X^k$, then$\hat{y}_G^j > \hat{y}_G^k$.

This mathematical verification fundamentally demonstrates the core discriminative semantics acquired from the source domain are perfectly preserved in the generated target domain without semantic collapse.

(2) Representation Consistency Regularization is a Special Case of Domain-Decorrelation Regularization

Most existing cross-domain generalization methods impose consistency regularization between f_X and f_G based on l_1 and l_2 losses, but we use $\hat{r} = f_X - f_G$to represent the unnormalized residual, and Δy^i denotes the prediction of r belonging to class i. Therefore, the objective of representation consistency regularization is to make $\hat{r}$ eventually approach a zero vector:

$$\hat{r} = 0 \Rightarrow sim(P_i, \hat{r}) = 0 \Rightarrow \Delta y^1 = \cdots = \Delta y^C = 0 \tag{D.6}$$

We believe such regularized representation is overly strict, while increasing the training difficulty of the model. Our goal is not to force $\hat{z}_n$ to approach a zero vector, but to ensure that $\hat{r}$ does not contain any task-related information, so we relax the constraint in the above equation:

$$\Delta y^1 = \cdots = \Delta y^C \tag{D.7}$$

It can be seen that Eq. (5) can also match the condition of Eq. (6), which makes representation consistency regularization a special case of this method.

(3) Making the Model Less Sensitive to Domain-Specific Attributes

The mutual information between the residual component and label attributes is:

$$\begin{aligned}I(R;Y) &= KL[p(r,y) \parallel p(r)p(y)]\\ &= \iint d_r d_y p(r,y)\log\frac{p(r,y)}{p(r)p(y)}\\ &= \iint d_r d_y p(r,y)\log\frac{p(y|r)p(r)}{p(r)p(y)}\\ &= \iint d_r d_y p(r,y)\log\frac{p(y|r)}{p(y)}\\ &= \iint d_r d_y p(r,y)\log\frac{1}{p(y)} - \iint d_r d_y p(r,y)\log\frac{1}{p(y|r)}\\ &= H(Y) - H(Y|R)\end{aligned} \tag{D.8}$$

where H denotes entropy, Y is the label set, R is the residual component set, $r \in R$, and $y \in Y$. Since $H(Y)$ is irrelevant to our optimization process, it can be ignored. Then we have:

$$\min_R I(R;Y) = \min_R -H(Y|R) = \max_R H(Y|R) \tag{D.9}$$

Therefore, by minimizing Eq. L_{rec}, we are only minimizing the mutual information between the residual component and label attributes. Since the generator imposes various task-irrelevant transformations to introduce domain-specific attributes into the original representation and correspondingly generate residual components, the residual component can be regarded as a proxy for domain-specific attributes. Minimizing the mutual information $I(R;Y)$ in the equation can decorrelate domain-specific attributes from label attributes.

Data availability

Data will be made available on request.

References

[1] Q. Liu, Z. Wu, Q. Du, Y. Xu, Z. Wei, Multiscale alternately updated clique network for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–15.

[2] Z. Shao, H. Fu, D. Li, O. Altan, T. Cheng, Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation, Remote Sens. Environ. 232 (2019) 111338.

[3] Z. Li, M. Liu, Y. Chen, Y. Xu, W. Li, Q. Du, Deep cross-domain few-shot learning for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–18, http://dx.doi.org/10.1109/TGRS.2021.3057066.

[4] C. Ding, Y. Chen, R. Li, D. Wen, X. Xie, L. Zhang, W. Wei, Y. Zhang, Integrating hybrid pyramid feature fusion and coordinate attention for effective small sample hyperspectral image classification, Remote. Sens. 14 (10) (2022) 2355.

[5] X. Yuan, J. Shi, L. Gu, A review of deep learning methods for semantic segmentation of remote sensing imagery, Expert Syst. Appl. 169 (2021) 114417.

[6] W.M. Kouw, M. Loog, A review of domain adaptation without target labels, IEEE Trans. Pattern Anal. Mach. Intell. 43 (3) (2019) 766–785.
[7] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, Z. Cai, Cross-scene hyperspectral image classification with discriminative cooperative alignment, IEEE Trans. Geosci. Remote Sens. 59 (11) (2021) 9646–9660.
[8] B. Deng, S. Jia, D. Shi, Deep metric learning-based feature embedding for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 58 (2) (2019) 1422–1435.
[9] Z. Feng, S. Tong, S. Yang, X. Zhang, L. Jiao, Pseudo-label-assisted subdomain adaptation for hyperspectral image classification, IEEE Trans. Circuits Syst. Video Technol. 34 (6) (2023) 4729–4744.
[10] Y. Zhang, W. Li, W. Jia, M. Zhang, R. Tao, S. Liang, Cross-domain hyperspectral image classification based on bi-directional domain adaptation, IEEE Trans. Circuits Syst. Video Technol. (2025).
[11] C. He, Y. Liu, Z. Jiang, Y. Li, C. Yan, B. Zheng, Patch-based cross-domain adaptation for few-shot remote sensing scene recognition, in: 2024 IEEE International Conference on Signal, Information and Data Processing, ICSIDP, IEEE, 2024, pp. 1–6.
[12] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, P.S. Yu, Generalizing to unseen domains: A survey on domain generalization, IEEE Trans. Knowl. Data Eng. 35 (8) (2022) 8052–8072.
[13] Y. Zhao, Z. Zhong, N. Zhao, N. Sebe, G.H. Lee, Style-hallucinated dual consistency learning for domain generalized semantic segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 535–552.
[14] X. He, B. Yang, F. Lyu, MMS: Morphology-mixup stylized data generation for single domain generalization in medical image segmentation, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 1981–1985.
[15] Y. Jia, J. Zhang, S. Shan, X. Chen, Single-side domain generalization for face anti-spoofing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8484–8493.
[16] M. Mancini, Z. Akata, E. Ricci, B. Caputo, Towards recognizing unseen categories in unseen domains, in: European Conference on Computer Vision, Springer, 2020, pp. 466–483.
[17] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1126–1135.
[18] J. Wang, Y. Zhai, Prototypical siamese networks for few-shot learning, in: 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication, ICEIEC, IEEE, 2020, pp. 178–181.
[19] M. Mancini, S.R. Bulo, B. Caputo, E. Ricci, Best sources forward: domain generalization through source-specific nets, in: 2018 25th IEEE International Conference on Image Processing, ICIP, IEEE, 2018, pp. 1353–1357.
[20] A. D'Innocente, B. Caputo, Domain generalization with domain-specific aggregation modules, in: German Conference on Pattern Recognition, Springer, 2018, pp. 187–198.
[21] G. Wu, S. Gong, Collaborative optimization and aggregation for decentralized domain generalization and adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6484–6493.
[22] F.M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, T. Tommasi, Domain generalization by solving jigsaw puzzles, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2229–2238.
[23] D. Kim, Y. Yoo, S. Park, J. Kim, J. Lee, Selfreg: Self-supervised contrastive regularization for domain generalization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9619–9628.
[24] S. Jeon, K. Hong, P. Lee, J. Lee, H. Byun, Feature stylization and domain-aware contrastive learning for domain generalization, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 22–31.
[25] G. Blanchard, A.A. Deshmukh, U. Dogan, G. Lee, C. Scott, Domain generalization by marginal transfer learning, J. Mach. Learn. Res. 22 (2) (2021) 1–55.
[26] T. Grubinger, A. Birlutiu, H. Schöner, T. Natschläger, T. Heskes, Domain generalization based on transfer component analysis, in: International Work-Conference on Artificial Neural Networks, Springer, 2015, pp. 325–334.
[27] C. Gan, T. Yang, B. Gong, Learning attributes equals multi-source domain generalization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 87–97.
[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[29] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (59) (2016) 1–35.
[30] R. Gong, W. Li, Y. Chen, L.V. Gool, Dlow: Domain flow for adaptation and generalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2477–2486.
[31] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, H. Qi, Topological structure and semantic information transfer network for cross-scene hyperspectral image classification, IEEE Trans. Neural Netw. Learn. Syst. 34 (6) (2021) 2817–2830.
[32] L. Dong, J. Geng, W. Jiang, Spectral–spatial enhancement and causal constraint for hyperspectral image cross-scene classification, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–13.
[33] Y. Zhang, W. Li, W. Sun, R. Tao, Q. Du, Single-source domain expansion network for cross-scene hyperspectral image classification, IEEE Trans. Image Process. 32 (2023) 1498–1512.
[34] Y. Zhang, S. Liang, W. Li, H. Ma, J. Xu, Y. Ma, J. Xie, W. Li, M. Zhang, R. Tao, et al., UniTS: Unified time series generative model for remote sensing, 2025, arXiv preprint arXiv:2512.04461.
[35] Y. Zhang, W. Li, M. Zhang, J. Han, R. Tao, S. Liang, Spectralx: Parameter-efficient domain generalization for spectral remote sensing foundation models, 2025, arXiv preprint arXiv:2508.01731.
[36] B. Qin, S. Feng, C. Zhao, B. Xi, W. Li, R. Tao, FDGNet: Frequency disentanglement and data geometry for domain generalization in cross-scene hyperspectral image classification, IEEE Trans. Neural Netw. Learn. Syst. 36 (6) (2024) 10297–10310.
[37] B. Wang, Y. Xu, Z. Wu, S. Zheng, Z. Wei, J. Chanussot, Hyperspectral images single-source domain generalization based on nonlinear sample generation, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–13.
[38] H. Zhao, J. Zhang, L. Lin, J. Wang, S. Gao, Z. Zhang, Locally linear unbiased randomization network for cross-scene hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–12.
[39] L. Xie, X. Yuan, A. Meyarian, Z. Qiao, Z. Wei, L. Gu, SPADe: Spatial plaid attention decoder for semantic segmentation of street views, IEEE Trans. Intell. Transp. Syst. (2025).
[40] Y. Deng, Z. Fang, S. Li, L. Chen, Beyond in-domain detection: SpikeScore for cross-domain hallucination detection, 2026, arXiv preprint arXiv:2601.19245.
[41] Y. Zhang, M. Zhang, W. Li, S. Wang, R. Tao, Language-aware domain generalization network for cross-scene hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–12.
[42] Q. Zhou, Q. Gu, J. Pang, X. Lu, L. Ma, Self-adversarial disentangling for specific domain adaptation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (7) (2023) 8954–8968.
[43] S. Sanyal, A.R. Asokan, S. Bhambri, A. Kulkarni, J.N. Kundu, R.V. Babu, Domain-specificity inducing transformers for source-free domain adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 18928–18937.
[44] D. Zhang, M. Chen, C. Li, L. Huang, J. Zhang, Aggregation of disentanglement: Reconsidering domain variations in domain generalization, 2023, arXiv preprint arXiv:2302.02350.
[45] Y. Chen, Y. Wang, Y. Pan, T. Yao, X. Tian, T. Mei, A style and semantic memory mechanism for domain generalization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9164–9173.
[46] Q. Liu, Z.B. Wu, Y. Xu, P. Zheng, S.D. Zheng, Z.H. Wei, Hyperspectral remote sensing image classification based on multidirectional adaptive aware network, Natl. Remote. Sens. Bull. 28 (1) (2024) 168–186, http://dx.doi.org/10.11834/jrs.20243292.
[47] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6023–6032.
[48] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.
[49] Y. Deng, X. Qiu, J. Chen, X. Tan, Reward guidance for reinforcement learning tasks based on large language models: The LMGT framework, Knowl.-Based Syst. 322 (2025) 113689.