Skipping Domain Shifts: Domain Memory Retention Enhanced Hyperspectral Single-Source Domain Generalization

Taiqin Chen[1,2], **Hao Sha**[1], **Xiaochen Feng**[1], **Chunyu Chen**[3], **Xiangyu Chen**[1], **Ke Chen**[2], **Yongbing Zhang**[1*]

[1]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
[2]Pengcheng Laboratory
[3]School of Computer Science and Technology, Harbin Institute of Technology
chentq@stu.hit.edu.cn, ybzhang08@hit.edu.cn

Abstract

Hyperspectral single-source domain generalization aims to train a robust model capable of overcoming domain shifts using only one available source domain for cross-scene hyperspectral image (HSI) analysis. Existing methods typically adopt data augmentation to broaden decision boundaries or utilize style transfer techniques for test-time alignment. However, these studies either fail to completely bridge the domain gap or inadvertently disrupt the discriminative spectral information, leading to suboptimal generalization and degraded classification performance on the target domain. To overcome these limitations, we propose a novel method, termed the domain memory retention framework, which optimizes domain memory during training to explicitly encode the semantic structure of the source domain. During inference, the target features are projected to the source distribution by leveraging this memory, thereby skipping domain shifts and improving generalization performance across unseen scenes. Specifically, the memory is formulated as a codebook, consisting of multiple semantic codewords. We further develop a semantic richness constraint and a perturbation robustness constraint to enhance representation diversity while empowering domain invariance for the memory. Extensive experiments conducted on three remote sensing datasets demonstrate that the proposed method outperforms state-of-the-art methods.

1 Introduction

Hyperspectral image (HSI) provides a fine-grained material characterization by jointly encoding rich spectral signatures and spatial structures [1], enabling diverse applications in fields such as remote sensing and medical diagnosis [2]. With the advances in deep learning [3, 4], HSI analysis has obtained remarkable achievement [5]. However, existing methods commonly rest on an implicit assumption that the training and test data are independently and identically distributed [6]. In the practical cross-scene settings, this assumption is frequently violated by factors such as atmospheric conditions and sensor disparities [7]. These variations inevitably induce severe domain shifts, consequently leading to substantial performance degradation [8, 9]. For clarity, the training data is denoted as the source domain (SD), and the test data from other scenes is referred to as the target domain (TD).

To tackle this challenge, domain generalization (DG) [10] has been proposed to learn transferable representations with access only to SD, enabling the model to make reliable predictions even on TD

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

with domain shifts. In classical DG settings [11–13], the model has access to multiple SD during training to learn domain-invariant discriminative features by exploiting distributional diversity across various domains. Unfortunately, satisfying this multi-domain requirement is often impractical in HSI analysis due to the substantial acquisition costs compared to RGB images [14]. Single-source domain generalization (SDG) [15] has thus attracted increasing attention, given that it relies on single SD.

Current SDG methods typically generate challenging samples via standard data augmentation techniques [16] or learnable generative models [17] during training to simulate potential domain shifts [18], as shown in Fig. 1(a). This paradigm broadens the network's decision boundaries and mitigates its over-reliance on the statistical properties and representation biases of a single domain. However, these gains are inherently limited by the inaccessibility of TD. Performance may still degrade markedly when encountering unseen domains that differ substantially from SD. To mitigate this issue, recent studies [19] propose to construct a style space for SD, as illustrated in Fig. 1(b). Specifically, the domain style space is formulated as a set of mean-variance pairs derived from shallow features, which are obtained either via statistical estimation [20] or learnable parameterization [21] during training. In inference, AdaIN [22] is utilized to project unseen samples from TD into this modeled space, enabling predictions under a known distribution. However, this strategy is fundamentally ill-suited for HSI analysis. The statistical characteristic along the spectral dimension intrinsically encodes discriminative information, serving as the key feature for distinguishing different categories. Directly applying these methods cannot effectively achieve cross-domain style alignment, while inevitably damaging the discriminative spectral features and causing severe semantic bias.

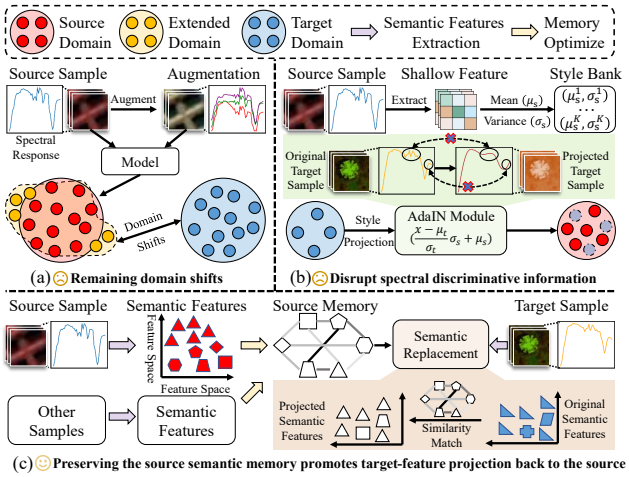


Figure 1: Comparison of existing studies and the proposed method. (a) Data augmentation expands decision boundaries but fails to fully resolve domain shifts. (b) Style transfer relies on shallow statistics (μ, σ) for alignment, risking the corruption of critical spectral discriminative features. (c) The proposed method leverages a domain memory to preserve source semantic structures, enabling robust target feature projection and enhanced generalization.

To address the above issue, we propose a domain memory retention framework (DMRF) for hyperspectral SDG, as illustrated in Fig. 1(c). Our key insight is to construct a domain memory to explicitly characterize the semantic structure of SD during training. Furthermore, this memory is employed to guide the projection of target-domain features back to source distribution during inference, thereby skipping the domain shifts. To establish a stable semantic representation, the proposed method formulates the domain memory as a codebook [23] comprising multiple discrete feature codewords. Briefly, semantic features is first extracted from the input HSI via an encoder. They are then quantized and projected into the semantic space represented by the codebook through similarity-based matching. A decoder generates final predictions based on the projected features. In addition, a semantic richness constraint is introduced to prevent the homogenization of discrete representations and avoid the under-utilization of the discrete codebook. We further introduce a perturbation robustness constraint to penalize codewords with high vulnerability to domain shifts, thereby yielding a domain-invariant codebook. The main contributions are as follow:

- We propose a novel framework for hyperspectral SDG, dubbed domain memory retention framework (DMRF). By constructing a domain memory to preserve source-domain semantic structure during training, DMRF effectively projects target-domain features back to source. This ensures that predictions are consistently grounded in a known distribution, substantially enhancing robustness against domain shifts.

- We develop a robust memory optimization strategy by coupling a semantic richness constraint with a perturbation robustness constraint. This collaborative mechanism effectively promotes representational diversity while ensuring domain invariance against perturbations.
- We conduct extensive experiments to demonstrate that the proposed method achieves state-of-the-art (SOTA) performance compared to existing SDG methods.

2 Related Work

2.1 Hyperspectral Single-Source Domain Generalization

Existing single-source domain generalization studies [24] attempt to simulate potential domain shifts during training via data augmentation. SDENet [14] generates diverse augmented samples by integrating spatial-spectral randomization with morphological features, and introduces contrastive learning [25] to extract domain-invariant representations. FSSG [26] further leverages continuous interpolation in the frequency space and pixel-wise semantic consistency constraints to obtain extended samples, which mitigates domain shifts caused by the deviation of content semantics. These methods overemphasize the diversity of generated samples, while neglecting the damage of unrealistic samples on performance. To tackle this issue, FDGNet [27] integrates data-geometric constraints with a class-wise manifold mechanism to ensure authenticity during augmentation. SPDDA [28] further designs a spectral property-driven data augmentation strategy to address the trade-off between diversity and realism in generated samples. Overall, while the above studies improve generalization by broadening decision boundaries via data augmentation, they inherently struggle to cover arbitrary domain shifts due to the inaccessible target distribution.

2.2 Style Transfer-based Domain Generalization

Different from augmentation-based methods, some pioneering studies construct a style space during training. By aligning both source domain (SD) and target domain (TD) within this space, they enable predictions under a known distribution. Briefly, they employ the mean and variance of shallow features to characterize sample style, termed style statistics. For example, TSB [20] constructs a style center for each SD by averaging the style statistics over training samples. During inference, adaptive instance normalization (AdaIN) [22] is employed to project target-domain features into the closest SD. SPCNet [19] employs a momentum update mechanism to aggregate training samples for obtaining style bases of different SD. Then, a style transfer at test time is completed by adaptively combining these bases according to distributional similarity. TTDG [21] further employs a learnable way to construct style bases of each SD. ConstStyle [29] builds a unified style domain via clustering the style statistics of training samples, and further provides theoretical justification for this design. While these studies have achieved notable success on RGB images, they typically struggle to generalize to HSI. Specifically, as continuous spectral signatures in HSI inherently encode discriminative information, coarse-grained replacement or alignment may disrupt intrinsic spectral correlations and induce severe semantic distortion. Although DDG [30] introduces discrete codebook learning to mitigate such distortion, its frequency-driven update mechanism may bias the learned codewords toward dominant source patterns, failing to preserve category-specific spectral semantics.

3 Method

3.1 Overview of domain memory retention framework

While generating challenging samples helps bridge the domain gap, the unavailability of the target domain (TD) prevents completely eliminating domain shifts. To tackle this issue, style transfer-based methods align shallow feature styles between the source domain (SD) and TD. However, directly applying this paradigm to hyperspectral images (HSI) corrupts essential discriminative information along the spectral dimension, severely impairing analytical performance. In contrast, the proposed domain memory retention framework (DMRF) constructs a domain memory to preserve the semantic structure of SD, and further projects target features back into the source distribution, achieving skipping domain shifts. The proposed DMRF is illustrated in Fig. 2.

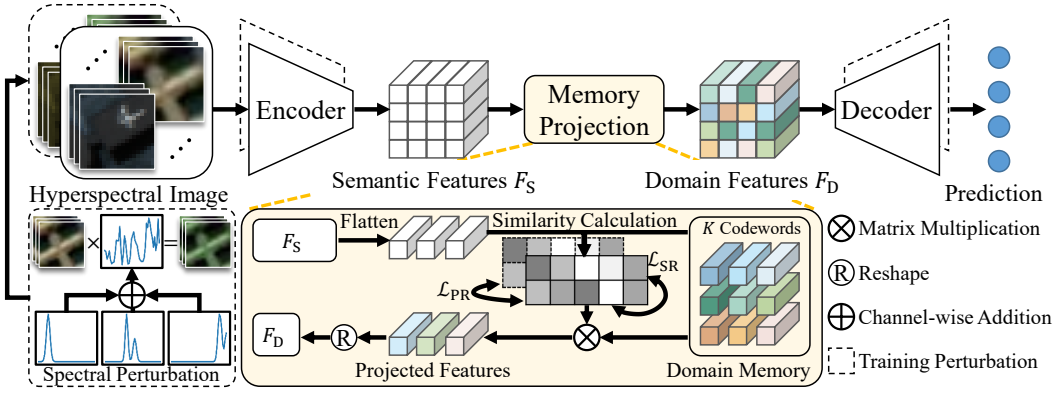


Figure 2: Overview of the proposed domain memory retention framework (DMRF). In DMRF, the model decouples into an encoder and a decoder, bridged by a domain memory. Semantic features F_{S} are projected into this memory space to yield domain features F_{D}. To regularize the memory, a semantic richness constraint $\mathcal{L}_{\mathrm{SR}}$ is proposed to maintain codeword diversity. Furthermore, a spectral perturbation is injected into training samples, while collaborating with a perturbation robustness constraint $\mathcal{L}_{\mathrm{PR}}$ to drive the encoding of domain-invariant representations.

In DMRF, the analytical model is decoupled into an encoder and a decoder, with a memory projection embedded between them. Specifically, given a HSI, DMRF first utilizes the encoder to extract semantic features $F_{\mathrm{S}} \in \mathbb{R}^{h\times w\times D}$, where D denotes the number of channels, h and w represent the spatial size. Memory projection is then performed by matching the features with the codebook based on their similarity, and obtaining domain features $F_{\mathrm{D}} \in \mathbb{R}^{h\times w\times D}$. Finally, these features are fed into the decoder to generate the prediction.

In the proposed method, the domain memory is formed as a codebook composed of multiple feature codewords, denoted by $\mathcal{C} \in \mathbb{R}^{K\times D}$, K denotes the number of codewords. To project F_{S} into the semantic space represented by $\mathcal{C}$, we first flatten the spatial dimensions of F_{S} to obtain $F_{\mathrm{SF}} \in \mathbb{R}^{N\times D}$, where $N = h \times w$. Then the cosine similarity S between each spatial feature in F_{SF} and each codeword in $\mathcal{C}$ is utilized to measure semantic distance. We further employ a softmax function $\mathrm{Softmax}(\cdot)$ in S to obtain the normalized projection weights $W \in \mathbb{R}^{N\times K}$, which are formed as:

$$W_{i,j} = \frac{\exp(S_{i,j})}{\sum_{k=1}^{K}\exp(S_{i,k})}, \tag{1}$$

where $S_{i,j}$ denotes the cosine similarity between the i-th feature in F_{SF} and the j-th codeword in $\mathcal{C}$. Subsequently, the projected features $\hat{F}_{\mathrm{S}}$ are derived via a weighted sum of the codewords, given by $\hat{F}_{\mathrm{S}} = W\mathcal{C}$. Reshaping $\hat{F}_{\mathrm{S}}$ back to the original dimensions yields the final domain features F_{D}.

While the aforementioned procedure enables effective feature reconstruction, directly updating $\mathcal{C}$ via the vanilla optimization objective may lead to representation degradation due to a lack of semantic diversity with the learned memory space. Specifically, the model's bias towards dominant patterns induces codeword redundancy, where multiple codewords encode overlapping semantics. Meanwhile, codewords handling less frequent features are rarely activated and gradually die out, triggering codebook collapse. Compounding these issues, the optimized representations easily overfit the specific distribution of the source domain, causing semantic concepts to be entangled with domain-specific bias. To mitigate these challenges, we develop a semantic richness constraint $\mathcal{L}_{\mathrm{SR}}$ and a perturbation robustness constraint $\mathcal{L}_{\mathrm{PR}}$.

3.2 Semantic Richness Constraint

Formally, $\mathcal{L}_{\mathrm{SR}}$ consists of a semantic orthogonal loss $\mathcal{L}_{\mathrm{ort}}$ and an entropy-based activating objective $\mathcal{L}_{\mathrm{act}}$, given by $\mathcal{L}_{\mathrm{SR}} = \mathcal{L}_{\mathrm{ort}} + \mathcal{L}_{\mathrm{act}}$. The former is designed to mitigate codeword redundancy by enforcing orthogonality within the memory space, which is formulated as:

$$\mathcal{L}_{\mathrm{ort}} = \frac{2}{K(K-1)}\sum_{1\le i<j\le K}\max\left(|R_{i,j}| - \epsilon,\, 0\right), \tag{2}$$

where $R_{i,j}$ denotes the cosine similarity between the i-th codeword c_i and the j-th codeword c_j, ϵ denotes a predefined margin. By penalizing pairwise similarities that exceed ϵ, $\mathcal{L}_{\text{ort}}$ explicitly forces distinct codewords to capture diverse and non-overlapping semantic patterns.

Furthermore, we introduce $\mathcal{L}_{\text{act}}$ to mitigate the codebook collapse issue, which is formed as:

$$\mathcal{L}_{\text{act}} = \mathcal{H}_{\text{ins}}(W) - \mathcal{H}_{\text{glo}}(\bar{W}) = (-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} W_{i,j}\log W_{i,j}) - (-\sum_{j=1}^{K}\bar{W}_j \log \bar{W}_j), \tag{3}$$

where $\mathcal{H}_{\text{ins}}$ represents instance entropy, $\mathcal{H}_{\text{glo}}$ represents global entropy, $\bar{W}$ denotes the average of W along the spatial dimensions, given by $\bar{W} = \frac{1}{N}\sum_{i=1}^{N} W_i$.

We maximize the global entropy $\mathcal{H}_{\text{glo}}$ to explicitly force the model to maintain a uniform distribution of codeword assignments, thereby activating potentially dead codewords. However, blindly pursuing this global uniformity risks over-smoothing the class-discriminative information. To counteract this side effect, we minimize the instance entropy $\mathcal{H}_{\text{ins}}$, which acts as a sharpness regularization for individual features. This compels each feature to produce a deterministic assignment, effectively preserving the intrinsic discriminative information during projection. Further theoretical analysis is presented in Appendix A.2.

3.3 Perturbation Robustness Constraint

Beyond the issue of semantic diversity, the learned codeword representation may be entangled with the source-domain distribution, causing a significant semantic shift during the projection of target-domain features. To mitigate this challenge, we introduce a perturbation robustness constraint $\mathcal{L}_{\text{PR}}$. The core intuition is that a robust codebook should assign similar projection weights to samples from the same semantic category. To achieve this, we inject perturbations into the training samples, whose details are presented in the Appendix A.1. Then domain-sensitive codewords are identified by comparing the projections of clean and perturbed features. We further explicitly optimize them accordingly to learn a more robust codebook.

Specifically, we first define the sensitive and insensitive weights via a softmax operation, referred to as π^+ and π^-, respectively, which is formulated as:

$$\pi^s_{i,k} = \text{Softmax}(s\alpha\Delta W_{i,k}), \quad s \in \{+,-\}, \tag{4}$$

where $s \in \{+,-\}$ serves as a sign indicator, i and k index the spatial location and the codeword, respectively. α denotes a scaling hyperparameter, and $\Delta W = W^{\text{pert}} - W^{\text{clean}}$ denotes the discrepancy between the projection weights of perturbed and clean features.

We further measure the codeword-wise projection energy $E \in \mathbb{R}^{N\times K}$ to evaluate the impact of perturbations on semantic assignments, which is formulated as:

$$E_{i,k} = \frac{(\Delta f_i \cdot c_k)^2}{\|\Delta f_i\|\|c_k\|}, \tag{5}$$

where $\Delta f = F_{\text{S}}^{\text{pert}} - F_{\text{S}}^{\text{clean}}$ denotes the variation in semantic features extracted from the perturbed training sample. Consequently, $\mathcal{L}_{\text{PR}}$ is further formulated as:

$$\mathcal{L}_{\text{PR}} = \sum_{i=1}^{N}\left[\frac{\exp(w_i^{\text{c}})}{\sum_{j=1}^{N}\exp(w_j^{\text{c}})}\right]\log\left(1 + \frac{\sum_k \pi^+_{i,k}E_{i,k}}{\sum_k \pi^-_{i,k}E_{i,k}}\right), \tag{6}$$

where $w_i^{\text{c}} = \max(\frac{1}{2}(W_i^{\text{clean}} + W_i^{\text{pert}}))$ denotes the semantic confidence of the i-th pixel, which is employed to prevent the supervised information from being dominated by irrelevant pixels.

3.4 Training Objective

In addition to $\mathcal{L}_{\text{SR}}$ and $\mathcal{L}_{\text{PR}}$, we also employ a semantic consistency constraint on the codebook, following conventional vector quantization methods [31]. However, unlike these methods that typically rely on large-scale pretraining to learn a highly expressive codebook, the proposed method is optimized end-to-end using source-domain data. Consequently, adopting vanilla constraint that

directly enforces semantic consistency between F_{S} and F_{D} would drive the codebook to overfit the source-domain distribution.

To alleviate this issue, we introduce a teacher-guided semantic consistency loss $\mathcal{L}_{\mathrm{con}}$. Briefly, we first construct an intermediate sample $X^{\mathrm{inter}} = \beta X^{\mathrm{clean}} + (1-\beta)X^{\mathrm{pert}}$ using a random scalar $\beta \in (0,1)$, where X^{clean} and X^{pert} denote the original and the perturbed samples, respectively. The corresponding domain features extracted from X^{pert} and X^{inter} are denoted as $F_{\mathrm{D}}^{\mathrm{pert}}$ and $F_{\mathrm{D}}^{\mathrm{inter}}$, respectively. Furthermore, $\mathcal{L}_{\mathrm{con}}$ is formulated as:

$$\mathcal{L}_{\mathrm{con}} = \sum_{F_{\mathrm{D}}^{(v)} \in \mathcal{F}} (1 - \langle F_{\mathrm{D}}^{(v)}, \bar{F} \rangle), \tag{7}$$

where $\mathcal{F} = \{\hat{F}_{\mathrm{D}}^{\mathrm{clean}}, \hat{F}_{\mathrm{D}}^{\mathrm{inter}}, \hat{F}_{\mathrm{D}}^{\mathrm{pert}}\}$ denotes set consisting of the normalized domain features, $\langle \cdot, \cdot \rangle$ represents the inner product, $\bar{F} = \mathrm{sg}(\frac{\sum \mathcal{F}}{3})$ establishes a unified semantic anchor designed to provide a consistent teacher-guided supervisory information for samples subjected to varying degrees of perturbation, $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. In a word, total loss $\mathcal{L}$ is formed as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{SR}} + \lambda_2 \mathcal{L}_{\mathrm{PR}} + \lambda_3 \mathcal{L}_{\mathrm{con}} + \mathcal{L}_{\mathrm{task}}, \tag{8}$$

where λ_1, λ_2, and λ_3 are the scalar coefficients, which are both set as 1 by default. $\mathcal{L}_{\mathrm{task}}$ denotes the task-oriented constraint, such as the cross-entropy loss for classification.

Table 1: Quantitative Results on Cross-scene Hyperspectral Image Classification for Remote Sensing. The optimal results are bold, and the sub-optimal results are underlined.

Methods	Houston			Pavia			HyRank		
	OA ↑	F1 ↑	Kappa ↑	OA ↑	F1 ↑	Kappa ↑	OA ↑	F1 ↑	Kappa ↑
FDGNet	75.44	0.7329	0.5360	78.28	0.7736	0.7371	56.81	0.5035	0.4554
S2AMSNet	71.13	0.7291	0.5541	73.42	0.7303	0.6792	59.46	0.5561	0.4940
S2ECNet	72.84	0.7355	0.5641	78.05	0.7828	0.7358	60.21	0.5617	0.5047
SPDDA	<u>78.60</u>	<u>0.7826</u>	0.6252	<u>83.03</u>	<u>0.8348</u>	<u>0.7963</u>	60.70	0.5316	0.5024
ProRandConv	75.71	0.7635	0.5911	78.27	0.7763	0.7375	57.76	0.5295	0.4753
StyDeSty	70.99	0.7155	0.5260	70.33	0.7075	0.6432	55.25	0.4559	0.4255
TF-Cal	65.61	0.6831	0.4834	73.54	0.7403	0.6825	57.53	0.4927	0.4536
SPCNet	69.26	0.7147	0.5349	67.36	0.6703	0.6071	59.62	0.5162	0.4814
ConstStyle	58.61	0.6236	0.4158	72.64	0.7326	0.6709	60.56	<u>0.5832</u>	0.5138
DDG	77.77	0.7809	<u>0.6337</u>	71.99	0.7192	0.6626	<u>61.63</u>	0.5693	<u>0.5141</u>
Ours	**80.37**	**0.7927**	**0.6532**	**85.76**	**0.8591**	**0.8281**	**63.03**	**0.6134**	**0.5438**

4 Experiment

4.1 Experimental Settings

4.1.1 Dataset

To demonstrate the advantages of our method, we adopt cross-scene remote sensing image classification as validation tasks. The employed datasets consists of Houston, Pavia, and HyRank [32, 33], where each dataset comprises distinct source domain (SD) and target domain (TD). Both Houston and Pavia share 7 common land-cover categories across domains. Unlike the other datasets, HyRank contains 12 classes and consists of satellite-acquired imagery with a coarser spatial resolution, making its classification more challenging. Furthermore, substantial domain shifts exist between SD and TD within each dataset owing to multiple factors, such as acquisition times or sensor discrepancy. In addition, we follow prior works [34] by explicitly aligning the number of spectral channels between SD and TD to standardize the input dimensions.

4.1.2 Evaluation Metrics

We utilize several evaluation metrics to quantify results of each method, including the overall accuracy (OA), the weighted F1 score (F1), and the kappa coefficient (Kappa).

4.1.3 Implementation Details

In all experiments, we employ LeNet [35] as the backbone following the prior studies [28]. The final convolution layer is designated as the boundary to separate the encoder and the decoder. The margin ϵ defined in Equ. 2 is set to 0.05, and the scaling parameter defined in Equ. 4 is set to 40. During training, the chosen optimizer is Adam with the initial learning rate is 1×10^{-3}. Detailed training procedure is presented in Appendix A.3.

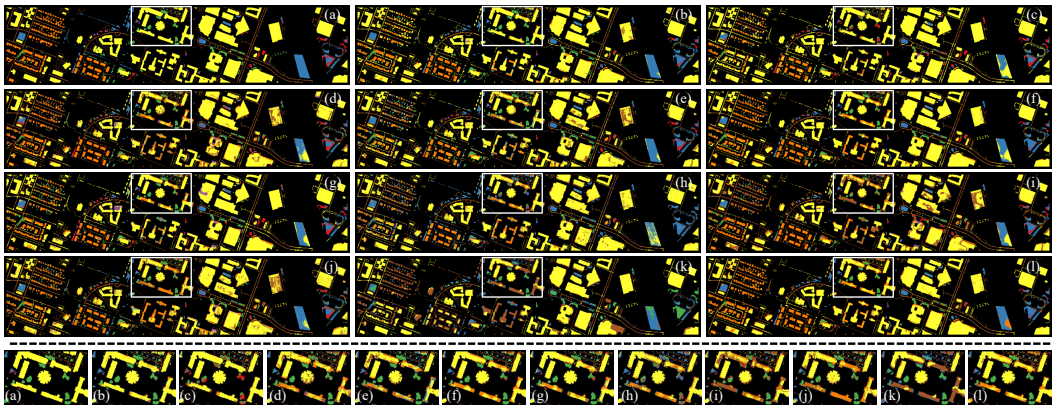


Figure 3: Visual comparison of classification results on Houston, which are spatially transposed for better display. The region indicated by the white rectangle is enlarged, with the corresponding zoomed-in view shown below the dashed line. (a) Ground Truth. (b) Ours. (c) FDGNet. (d) S2AMSNet. (e) S2ECNet. (f) SPDDA. (g) ProRandConv. (h) StyDeSty. (i) TF-Cal. (j) SPCNet. (k) ConstStyle. (l) DDG.

4.2 Comparison Study

To verify the effectiveness of the proposed method, we compare our DMRF against several state-of-the-art (SOTA) methods for hyperspectral single-source domain generalization, including FDGNet [27], S2AMSNet [36], S2ECNet [37], and SPDDA [28]. The SOTA methods designed for RGB images, namely ProRandConv [38] and StyDeSty [39], are also included as comparisons. Furthermore, to demonstrate the limitations of relying on style transformations, we also evaluate style transfer-based methods, such as TF-Cal [40], SPCNet [19], and ConstStyle [29]. The discrete codebook learning DDG [30] is also employed as comparison. For fairness, the style transfer-based methods and DDG are equipped with the identical data augmentation strategy used in the proposed method.

4.2.1 Quantitative Analysis

Comparative results are presented in Tab. 1. The proposed method outperforms comparative methods across all datasets, achieving an average improvement of over 2.0% in OA compared to the sub-optimal method. SPDDA addresses the inherent trade-off between generation diversity and fidelity by introducing physical priors like spectral heterogeneity and spectral mixing, yielding the most competitive performance among all baselines. However, the generated samples inevitably exhibit a domain shift relative to TD since target-domain data is inaccessible during training. A similar issue also exists in other augmentation-based methods. Style transfer-based methods have been proposed to tackle this issue, which is tailored for RGB images. These methods leverage a statistics-based domain style bank to align TD samples with SD during inference, attempting to ensure predictions within a known distribution. Unfortunately, such statistics-based transfer operations may disrupt the spectral discriminative information, thereby degrading the analytical performance of the model. Consequently, RGB-oriented style transfer-based methods struggle to outperform augmentation-based methods, as evidenced by the results of TF-Cal, SPCNet, and ConstStyle. Although DDG employs discrete codebook learning to mitigate this issue, its frequency-driven update mechanism causes codewords to entangle with SD distribution, as evidenced by the results on Pavia. In contrast to existing methods, our method constructs a domain memory to preserve the semantic structure of SD. A robust optimization strategy for the domain memory is further developed to mitigate semantic

Table 2: Impact of the number of codewords K on Houston.

K	OA ↑	F1 ↑	Kappa ↑
8	77.46	0.7790	0.6320
16	<u>79.29</u>	0.7806	0.6270
24	79.10	<u>0.7888</u>	<u>0.6455</u>
32	78.97	0.7814	0.6309
64	**80.37**	**0.7927**	**0.6532**
128	76.07	0.7662	0.5944
256	74.64	0.7575	0.5846
512	67.41	0.6993	0.5070

Table 3: Ablation study of the proposed constraints on Houston.

$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{SR}}$	$\mathcal{L}_{\text{PR}}$	OA ↑	F1 ↑	Kappa ↑
×	×	×	77.34	0.7792	0.6174
✓	×	×	78.05	0.7856	0.6362
×	✓	×	77.40	0.7761	0.6237
×	×	✓	78.94	0.7829	0.6314
✓	✓	×	77.92	0.7777	0.6240
✓	×	✓	78.93	**0.7931**	<u>0.6447</u>
×	✓	✓	<u>79.30</u>	0.7802	0.6319
✓	✓	✓	**80.37**	<u>0.7927</u>	**0.6532**

redundancy while ensuring domain invariance. During testing, target-domain features are projected back to the source distribution, achieving superior cross-scene performance.

4.2.2 Qualitative Analysis

We also visualize predictions generated by all methods on Houston, as shown in Fig. 3. More visual comparisons are presented in Appendix A.4.4. In addition, the Houston results were spatially transposed for better visual presentation. Compared with the Ground Truth (GT), the compared methods generally exhibit isolated misclassified pixels and spatial discontinuities. To facilitate a clearer comparison, we enlarge a complex region marked by a white rectangle, which predominantly consists of a single land-cover class but contains minor instances of other categories near its boundaries. Observing these enlarged regions, it is evident that the compared methods suffer from sever pixel-level misclassifications, leading to jagged boundaries that disrupt the spatial and structural continuity of the ground objects. In contrast, the classification map predicted by the proposed method is closer to GT, as demonstrated by the comparisons between Fig. 3(a) and Fig. 3(b). Our method effectively mitigates scattered misclassifications within the region, preserving the integrity and continuity of ground objects while producing smooth, precise boundaries.

4.3 Ablation Study

4.3.1 Hyperparameter Sensitivity Analysis

The number of codewords K in the codebook is a critical parameter for the proposed method. Thus, we conduct a sensitivity analysis by varying K across a range of values, whose results are presented in Tab. 2. When K is relatively small, the codebook lacks sufficient codewords to comprehensively characterize the semantic structures of the source domain, leading to a bottleneck in representation and suboptimal performance. In this regime, appropriately increasing the number of codewords can significantly enhance the model's discriminative ability. The model achieves optimal performance at $K = 64$. However, a pronounced performance degradation occurs when K becomes excessively large. This can be attributed to the codebook capacity exceeding the intrinsic complexity of the training datasets and the task, thereby inducing structural redundancy in the latent space. Such redundancy impairs the codebook's ability to distill representative prototypes, resulting in feature confusion and diminished discriminative power across classes. Consequently, we set K to 64 for Houston and Pavia, while increasing it to 128 for the more challenging HyRank.

4.3.2 Constraints Ablation Analysis

We then perform ablation experiments to verify the effect of $\mathcal{L}_{\text{SR}}$, $\mathcal{L}_{\text{PR}}$, and $\mathcal{L}_{\text{con}}$. Regarding the ablation settings, the mean absolute error (MAE) between semantic features and domain features is employed as a alternative for $\mathcal{L}_{\text{con}}$. For $\mathcal{L}_{\text{SR}}$ and $\mathcal{L}_{\text{PR}}$, we perform a straightforward removal to evaluate their individual contributions to the proposed method. In addition, the perturbation injection is still utilized to generate extended data when $\mathcal{L}_{\text{PR}}$ is omitted to ensure fairness.

Results are presented in Tab. 3. Comparing with the vanilla version in terms of OA, the standalone integration of each proposed constraint consistently yields a performance gain, validating that each module contributes positively to the model's generalized ability. Notably, $\mathcal{L}_{\text{PR}}$ exhibits the most significant improvement, increasing the OA from 77.34% to 78.94%. This demonstrates that

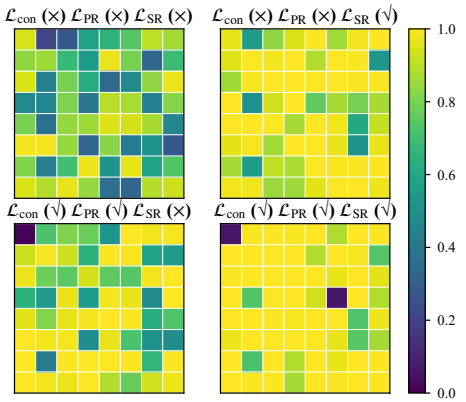


Figure 4: Visualization of the normalized activation frequency of 64 codewords, with yellow representing higher frequencies.

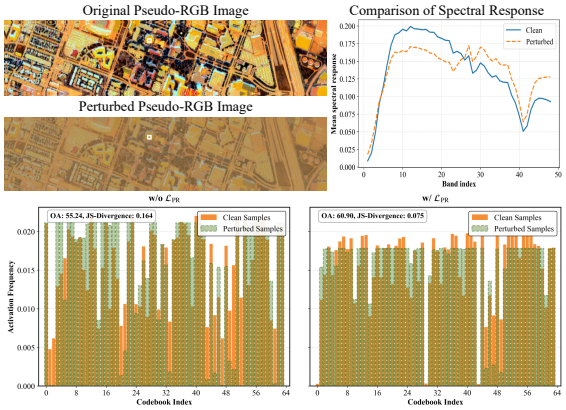


Figure 5: Visualization of the contribution of the proposed perturbation robustness constraint ($\mathcal{L}_{\mathrm{PR}}$) to domain-invariant feature learning.

encouraging the model to encode domain-invariant features within the codebook plays a pivotal role in enhancing generalization. Combining $\mathcal{L}_{\mathrm{PR}}$ with $\mathcal{L}_{\mathrm{SR}}$ further increases the OA to 79.30%. The boost can be attributed to the introduction of $\mathcal{L}_{\mathrm{SR}}$, which effectively optimizes the activation distribution of each codeword, thereby enriching the semantic diversity represented by the codebook.

To empirically verify this observation, we conduct a quantitative analysis of codeword activation frequency during inference. Based on the projection weights $W_{i,j}$, we employ the following criterion to determine whether the j-th codeword is activated when processing the i-th feature:

$$\mathbb{I}_{i,j} = \begin{cases} 1, & \text{if } W_{i,j} > \tau \cdot \bar{W}_i \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where $\bar{W}_i = \frac{1}{K}\sum_{k=1}^{K} W_{i,k}$ represents the mean projection weight of the i-th feature across the entire codebook, and τ is a proportional threshold, which is set to 0.8 in this experiment. Subsequently, we calculate the cumulative activation count for each codeword across all samples in the target domain, whose normalized statistical results are visualized in Fig. 4. The visualization intuitively demonstrates that $\mathcal{L}_{\mathrm{SR}}$ acts as a effective regularizer that discourages codeword under-utilization, yielding a more robust and expressive latent representation across the target dataset.

To validate the contribution of $\mathcal{L}_{\mathrm{PR}}$ to learning domain-invariant features, we also employ a quantitative analysis, which is shown in Fig. 5. During experiment, the contrast perturbations are artificially injected into the target-domain data. As illustrated in the top-left of Fig. 5, the original and perturbed pseudo-RGB images exhibit noticeable visual discrepancies. The perturbation induces a structural offset in the original spectral curve, which is observed by the top-right of Fig. 5 depicting the mean spectral response from the region highlighted by the white rectangle.

In cross-domain classification tasks, an ideal codeword should remain robust against such low-level spectral variations. Therefore, we investigate the codebook activation distributions before and after perturbation to evaluate whether the learned codewords effectively capture domain-invariant information. The Jensen-Shannon Divergence (JS-Divergence) [41] is introduced as a quantitative metric to assess the consistency between the original and perturbed distributions. To avoid interference from other factors, we compare the vanilla version with the version incorporating only $\mathcal{L}_{\mathrm{PR}}$. Results are illustrated in in the bottom of Fig. 5. Without the integration of $\mathcal{L}_{\mathrm{PR}}$, the activation distributions are significantly misaligned, yielding a high JS-Divergence of 0.164 and a degraded OA of 55.24%. This indicates that the vanilla codebook is highly vulnerable to domain variations. Conversely, with the incorporation of $\mathcal{L}_{\mathrm{PR}}$, the distributions of clean and perturbed samples become tightly aligned, dramatically reducing the JS-Divergence to 0.075 and boosting OA to 60.90%. The comparisons prove that $\mathcal{L}_{\mathrm{PR}}$ effectively forces the model to distill perturbation-invariant semantic structures within the codebook, thereby enhancing the robustness of cross-domain representations.

5 Conclusion

In this paper, we propose the domain memory retention framework (DMRF) for hyperspectral single-source domain generalization. DMRF optimizes a memory to explicitly encode the semantic structure of the source domain, aligning target features into a consistent distribution. Furthermore, two novel constraints are developed to guarantee memory domain-invariance and semantic diversity. Extensive experiments demonstrate the effectiveness of the proposed method. Despite its robust performance, DMRF relies on the inherent assumption in domain generalization that the model can establish clear inter-class boundaries in the source domain. However, increased noise or categories weaken this separability, inevitably causing codeword redundancy and limiting generalization. Future work will explore advanced paradigms for domain shifts under complex distributions.

Acknowledgments

This research was partially supported by the National Key R&D Program of China (No.2023YFC3305600), the National Natural Science Foundation of China (62331011), the Shenzhen Science and Technology Project (ZDYJ20251211140359002), and the Major Key Project of Pengcheng Laboratory (PCL2025A02).

References

[1] Danfeng Hong, Chenyu Li, Naoto Yokoya, Bing Zhang, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Hyperspectral imaging. *Nature Reviews Methods Primers*, 6(1):19, 2026.

[2] Taiqin Chen, Hao Sha, Yifeng Wang, Yuan Jiang, Shuai Liu, Zikun Zhou, Ke Chen, and Yongbing Zhang. A channel-adaptive and plug-and-play framework for hyperspectral image analysis. *Information Fusion*, page 103770, 2025.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[4] Taiqin Chen, Zikun Zhou, Zheng Fang, Wenzhen Zou, Kangjun Liu, Ke Chen, Yongbing Zhang, and Yaowei Wang. Radiodun: A physics-inspired deep unfolding network for radio map estimation. *IEEE Transactions on Cognitive Communications and Networking*, 2026.

[5] 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.

[6] Zhilin Zhu, Xiaopeng Hong, Zhiheng Ma, Weijun Zhuang, Yaohui Ma, Yong Dai, and Yaowei Wang. Reshaping the online data buffering and organizing mechanism for continual test-time adaptation. In *European Conference on Computer Vision*, pages 415–433. Springer, 2024.

[7] Zhilin Zhu, Yabin Wang, Zhiheng Ma, Yaguang Song, Yaowei Wang, and Xiaopeng Hong. Dance across shifts: Forward-facilitation continual test-time adaptation through dynamic style bridging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 32322–32333, June 2026.

[8] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

[9] Zhilin Zhu, Zhiheng Ma, Yabin Wang, Yaguang Song, Yaowei Wang, and Xiaopeng Hong. Sample-aware knowledge association and enhancement for open-vocabulary continual learning. *International Journal of Computer Vision*, 134(7):332, 2026.

[10] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

[11] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[12] Sunny Gupta, Vinay Sutar, Varunav Singh, and Amit Sethi. Fedalign: Federated domain generalization with cross-client feature alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1801–1810, 2025.

[13] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

[14] 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.

[15] Qinwei Xu, Ruipeng Zhang, Yi-Yan Wu, Ya Zhang, Ning Liu, and Yanfeng Wang. Simde: A simple domain expansion approach for single-source domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4798–4808, 2023.

[16] Guangtao Zheng, Mengdi Huai, and Aidong Zhang. Advst: Revisiting data augmentations for single domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 21832–21840, 2024.

[17] Yuxiang Zhang, Mengmeng Zhang, Wei Li, Shuai Wang, and Ran Tao. Language-aware domain generalization network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

[18] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 834–843, 2021.

[19] Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Style projected clustering for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3061–3071, 2023.

[20] Jungwuk Park, Dong-Jun Han, Soyeong Kim, and Jaekyun Moon. Test-time style shifting: Handling arbitrary styles in domain generalization. In *International Conference on Machine Learning*, pages 27114–27131. PMLR, 2023.

[21] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–187, 2024.

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[23] Guotao Liang, Baoquan Zhang, Zhiyuan Wen, Junteng Zhao, Yunming Ye, Kola Ye, and Yao He. Towards improved text-aligned codebook learning: Multi-hierarchical codebook-text alignment with long text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4060–4069, 2025.

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[26] 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*, 2025.

[27] 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, 2024.

[28] 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*, volume 40, pages 3038–3046, 2026.

[29] Nam Duong Tran, Nam Nguyen Phuong, Hieu H Pham, Phi Le Nguyen, and My T Thai. Conststyle: Robust domain generalization with unified style transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3174–3183, 2025.

[30] Shaocong Long, Qianyu Zhou, Xikun Jiang, Chenhao Ying, Lizhuang Ma, and Yuan Luo. Domain generalization via discrete codebook learning. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[31] Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vq-gan to 100,000 with a utilization rate of 99%. *Advances in Neural Information Processing Systems*, 37: 12612–12635, 2024.

[32] Yuxiang Zhang, Wei Li, Mengmeng Zhang, Ying Qu, Ran Tao, and Hairong Qi. Topological structure and semantic information transfer network for cross-scene hyperspectral image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6):2817–2830, 2021.

[33] Yuxiang Zhang, Wei Li, Mengmeng Zhang, Shuai Wang, Ran Tao, and Qian Du. Graph information aggregation cross-domain few-shot learning for hyperspectral image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1912–1925, 2022.

[34] Puhong Duan, Shiyu Jin, Xiaotian Lu, Lianhui Liang, Xudong Kang, and Antonio Plaza. Domain-adaptive mamba for cross-scene hyperspectral image classification. *IEEE Transactions on Image Processing*, 2026.

[35] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[36] 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.

[37] 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.

[38] Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random convolutions for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10312–10322, 2023.

[39] Songhua Liu, Xin Jin, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Stydesty: Min-max stylization and destylization for single domain generalization. *arXiv preprint arXiv:2406.00275*, 2024.

[40] Xingchen Zhao, Chang Liu, Anthony Sicilia, Seong Jae Hwang, and Yun Fu. Test-time fourier style calibration for domain generalization. In *31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 1721–1727. International Joint Conferences on Artificial Intelligence, 2022.

[41] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 2002.

[42] Jurij Jemec, Franjo Pernuš, Boštjan Likar, and Miran Bürmen. Push-broom hyperspectral image calibration and enhancement by 2d deconvolution with a variant response function estimate. *Optics express*, 22(22):27655–27668, 2014.

[43] Hao Sha, Haoyang Li, Yongbing Zhang, and Shangguo Hou. Deep learning-enhanced single-molecule spectrum imaging. *Apl Photonics*, 8(9), 2023.

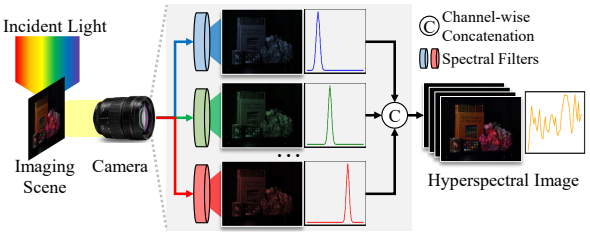


Figure 6: The imaging principle of push-broom hyperspectral cameras.

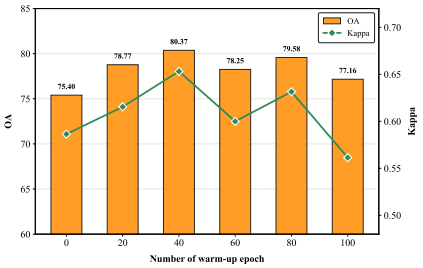


Figure 7: Impact of the warm-up epochs E_{w} on Houston.

A Technical appendices and supplementary material

A.1 Perturbation Injection

Inspired by the imaging mechanism of push-broom hyperspectral cameras [42], we introduce a spectral transmission simulation to inject perturbation into training samples. The detailed imaging procedure is illustrated in Fig. 6. Briefly, the camera uses a set of filters with different spectral responses to selectively transmit energy from incoming light and record corresponding images for each band. They are concatenated along the spectral dimension to form final HSI.

Based on the physical process, we serve the original hyperspectral images as the continuous energy distribution of incident light across the full spectrum. The spectral response of each channel is assumed to follow a Gaussian distribution [43]. In real-world scenarios, the spectral response of filters inevitably deviate from ideal transmission profiles due to multiple factors like manufacturing processes. For example, degraded out-of-band blocking capabilities can trigger energy leakage from adjacent bands, or even from additional secondary sidebands. To simulate the practical phenomenon, we formulate the filter response $g(\Delta b)$ as a mixture of M Gaussian distributions. Taking the c-th band as an example:

$$g(\Delta b) = \sum_{m=1}^{M} \rho_m \exp\left(-\frac{(\Delta b - \mu_m)^2}{2\sigma_m^2}\right), \tag{10}$$

where $\Delta b \in [-B, B]$ denotes the filtering range of the filter, $\mu_1 = 0$ is determined to as the primary peak, auxiliary interference peaks $\mu_m (m > 1)$ are randomly distributed within the range $[-B, B]$ to simulate random cross-band energy leakage. The amplitudes ρ_m decay exponentially to model the diminishing effect of severe leakage, given by a random decay factor $\gamma \in (0, 1)$ such that $\rho_m = \gamma^{m-1}$. The bandwidth of each peak is controlled by a randomly sampled standard deviation σ_m.

Subsequently, the synthetic filter is normalized to ensure energy conservation during the simulated imaging process. The c-th channel of perturbed image X_c^{pert} is formulated as:

$$X_c^{\mathrm{pert}} = \sum_{\Delta b=-B}^{B} \hat{g}(\Delta b) \cdot X_{c+\Delta b}, \tag{11}$$

where $\hat{g}(\cdot)$ denotes the normalized spectral response, X_c denotes the c-th channel of original images.

A.2 Theoretical Analysis of the Semantic Richness Constraint

The Semantic Richness Constraint is formulated based on the complementary characteristics of instance-level assignment uncertainty and global codeword utilization diversity. Although both terms are derived from the codeword assignment probability W, they characterize different statistical properties of the codeword assignments. To further understand the optimization behavior and theoretical properties of the proposed constraint, we analyze its information-theoretic interpretation and derive the corresponding objective bounds. Specifically, we establish an information-theoretic interpretation of the constraint and analyze its optimization characteristics through objective bounds and extremal cases.

Let U denote the feature instance index and Z denote the activated codeword index. Since W represents the normalized probability distribution over codewords, the conditional probability between a feature instance and its assigned codeword is defined as $p(Z = j|U = i) = W_{i,j}$, where each feature instance is sampled uniformly, given that $p(U = i) = \frac{1}{N}$.

Accordingly, the instance-level entropy term can be interpreted as the conditional entropy of codeword assignments:

$$\mathcal{H}_{\text{ins}}(W) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} W_{i,j}\log W_{i,j} = H(Z|U). \tag{12}$$

Meanwhile, the marginal probability of activating the j-th codeword is given by the global codeword utilization distribution, given that $p(Z = j) = \bar{W}_j$. Therefore, the global entropy term can be expressed as:

$$\mathcal{H}_{\text{glo}}(\bar{W}) = -\sum_{j=1}^{K} \bar{W}_j \log \bar{W}_j = H(Z). \tag{13}$$

The Semantic Richness Constraint is further reformulated as:

$$\mathcal{L}_{\text{act}} = \mathcal{H}_{\text{ins}}(W) - \mathcal{H}_{\text{glo}}(\bar{W}) = H(Z|U) - H(Z) = -I(U;Z), \tag{14}$$

where $I(U;Z)$ denotes the mutual information between feature instances and their activated codewords. Therefore, minimizing $\mathcal{L}_{\text{act}}$ is equivalent to maximizing the mutual information, which encourages codeword assignments to preserve feature-specific semantic information while promoting diverse codeword utilization.

The optimization range of the proposed objective can be further analyzed according to the properties of entropy. Since entropy is a concave function, Jensen's inequality [**?**] gives:

$$H\left(\frac{1}{N}\sum_{i=1}^{N} W_i\right) \geq \frac{1}{N}\sum_{i=1}^{N} H(W_i), \tag{15}$$

where W_i denotes the codeword assignment distribution of the i-th feature instance. According to the definitions of $\mathcal{H}_{\text{ins}}$ and $\mathcal{H}_{\text{glo}}$, we have $\mathcal{H}_{\text{glo}}(\bar{W}) \geq \mathcal{H}_{\text{ins}}(W)$. Consequently, the Semantic Richness Constraint satisfies $\mathcal{L}_{\text{act}} = \mathcal{H}_{\text{ins}}(W) - \mathcal{H}_{\text{glo}}(\bar{W}) \leq 0$. Furthermore, according to the entropy bounds of a discrete probability distribution, given that:

$$\mathcal{H}_{\text{ins}}(W) \geq 0, \qquad \mathcal{H}_{\text{glo}}(\bar{W}) \leq \log K, \tag{16}$$

the objective is bounded as $-\log K \leq \mathcal{L}_{\text{act}} \leq 0$. This bounded property indicates that the proposed constraint has a well-defined optimization range.

A.3 Experiment Details

All experiments are conducted on the Pytorch platform, whose version is 2.1.0 and the computing hardware is an NVIDIA RTX 4090D GPU. To prevent randomly initialized codewords from destabilizing the model training, the whole training procedure is divided into warm-up training and joint training. Details are presented in Algorithm 1. The maximum training epoch E is set to 400, and the number of warm-up epoch E_{w} is set to 40 for all experiments.

A.4 More Experiments

A.4.1 Hyperparameter Tuning Study

We conduct a tuning study on the warm-up epochs E_{w} to evaluate the impact of the warm-up stage. Experimental dataset is employed as Houston, whose results are shown in Fig. 7. When directly

Algorithm 1 Training Procedure of the Proposed Method

Require: Source dataset $\mathcal{D}_{\mathrm{s}}$, warm-up epoch E_{w}, maximum epoch E, analytical model Θ composed of an encoder θ_{enc}, a domain memory $\mathcal{C}$, and a decoder θ_{dec}

Randomly initialize Θ
Stage 1. Warm-up training
for $epoch = 1$ to E_{w} **do**
  for each mini-batch (X^{clean}, y) from $\mathcal{D}_s$ **do**
    Generate X^{pert} and X^{inter} by Algorithm 2
    Generate predictions via $\theta_{\mathrm{dec}}(\theta_{\mathrm{enc}}(\cdot))$ and freeze parameter of $\mathcal{C}$
    Update θ_{enc} and θ_{dec} through the task-oriented constraint $\mathcal{L}_{\mathrm{task}}$
    Freeze θ_{enc} and θ_{dec}, activate the parameter of $\mathcal{C}$
    Obtain semantic features F_{S} via $\theta_{\mathrm{enc}}(\cdot)$ and obtain domain features F_{D} following Equ. 1
    Generate predictions via $\theta_{\mathrm{dec}}(\cdot)$ based on F_{D}
    Update $\mathcal{C}$ by minimizing $\mathcal{L}$
  end for
end for
Activate the parameter of Θ
Stage 2. Joint training
for $epoch = E_{\mathrm{w}}$ to E **do**
  for each mini-batch (X^{clean}, y) from $\mathcal{D}_s$ **do**
    Generate X^{pert} and X^{inter} by Algorithm 2
    Obtain F_{S} via $\theta_{\mathrm{enc}}(\cdot)$ and obtain F_{D} following Equ. 1
    Generate predictions via $\theta_{\mathrm{dec}}(\cdot)$ based on F_{D}
    Update Θ by minimizing $\mathcal{L}$
  end for
end for

Algorithm 2 Procedure of injecting perturbation

Require: Original HSI mini-batch X^{clean}, maximum number of Gaussian distributions $M_{\max}$, filtering range Δb, standard deviation range $[\sigma_{\min}, \sigma_{\max}]$

Randomly sample the number of Gaussian distributions $M \sim \mathcal{U}\{1, \ldots, M_{\max}\}$
Set the center of primary peak $\mu_1 \leftarrow 0$
Randomly sample the center of auxiliary interference peaks $\mu_i \sim \mathcal{U}\{-B, \ldots, B\}, \quad i \in [2, M]$
Randomly sample the decay factor $\gamma \sim \mathcal{U}(0, 1)$, set the amplitudes as $\rho_m = \gamma^{m-1}, \quad m \in [1, M]$
Randomly sample the standard deviation for each Gaussian distribution $\sigma_m \sim \mathcal{U}(\sigma_{\min}, \sigma_{\max})$
Compute the filter response as Equ. 10
Obtain the perturbed HSI X^{pert} via Equ. 11
Obtain the intermediate HSI X^{inter} via $\beta X^{\mathrm{clean}} + (1 - \beta) X^{\mathrm{pert}}$, where β is randomly sample by $\mathcal{U}(0, 1)$
return X^{pert} and X^{inter}

performing joint training (i.e., $E_{\mathrm{w}} = 0$), the randomly initialized domain memory $\mathcal{C}$ generates noisy gradients that mislead the feature projection, causing a significant performance drop. Conversely, an excessively long warm-up period (e.g., $E_{\mathrm{w}} \geq 60$) would compress the joint optimization resource, making it difficult for the memory space to represent the semantic structures of the source domain. Therefore, we set E_{w} as 40 to balance the proper initialization and the subsequent joint training.

A.4.2 Ablation for scalar coefficients in the training objective

We further investigate the sensitivity of the scalar coefficients λ_1, λ_2, and λ_3 in Eq. 8 on Houston. During each experiment, only one coefficient is varied while the remaining coefficients are fixed to the default value of 1. As shown in Tab. 4, the scalar coefficients provide effective control over the relative importance of different objective terms. Moderate weights generally lead to better target-domain performance, while overly large weights do not further improve the results and may even reduce performance due to over-regularization or optimization imbalance.

Table 4: Hyperparameter sensitivity analysis of λ_1, λ_2, and λ_3.

Value	λ_1			λ_2			λ_3		
	OA ↑	F1 ↑	Kappa ↑	OA ↑	F1 ↑	Kappa ↑	OA ↑	F1 ↑	Kappa ↑
0.1	<u>80.98</u>	<u>0.7979</u>	<u>0.6597</u>	77.37	0.7775	0.6279	79.86	0.7783	0.6303
0.5	**81.63**	**0.8061**	**0.6737**	<u>80.08</u>	<u>0.7913</u>	<u>0.6421</u>	**80.49**	<u>0.7910</u>	<u>0.6483</u>
1.0	80.37	0.7927	0.6532	**80.37**	**0.7927**	**0.6532**	<u>80.37</u>	**0.7927**	**0.6532**
1.5	79.71	0.7822	0.6308	79.16	0.7703	0.6207	79.49	0.7758	0.6281
2.0	80.15	0.7855	0.6388	79.75	0.7808	0.6246	79.46	0.7757	0.6251
5.0	78.81	0.7765	0.6221	78.70	0.7725	0.6128	78.53	0.7638	0.6046

A.4.3 Efficiency Comparison

To evaluate the computational efficiency of the proposed method, we conduct a comparative experiment on inference time. The computing hardware is an NVIDIA RTX 4090D GPU. All methods are tested under the same batch setting (the size is $256 \times 48 \times 13 \times 13$). For each method, we conduct 50 warm-up iterations to stabilize the execution, followed by 200 inference iterations. This procedure is repeated five times, and the final results are reported as the mean inference time with standard deviation. The augmentation-based methods are collectively regarded as baselines, as they introduce no modifications to the classifier. Results are presented in Tab. 5. Since TF-Cal performs style transfer by replacing sample amplitudes in the frequency domain via Fast Fourier Transform, it introduces the highest computational overhead. In contrast, the remaining methods incorporate lightweight operations that exert a significantly smaller impact on efficiency than TF-Cal.

Table 5: Comparison of inference time.

Method	Time (ms/batch)
Baseline	2.0922 ± 0.0000
TF-Cal	3.0124 ± 0.0261
SPCNet	2.2541 ± 0.0002
ConstStyle	2.1304 ± 0.0073
DDG	2.1396 ± 0.0029
Ours	2.1431 ± 0.0115

A.4.4 More Visual Comparisons

Visual comparisons of classification results on Pavia and HyRnk are depicted in Fig. 8 and Fig. 9, respectively. Similarly, we enlarge a complex region marked by a white rectangle to facilitate a clearer comparison, with the corresponding sub-figure shown below the dashed line.

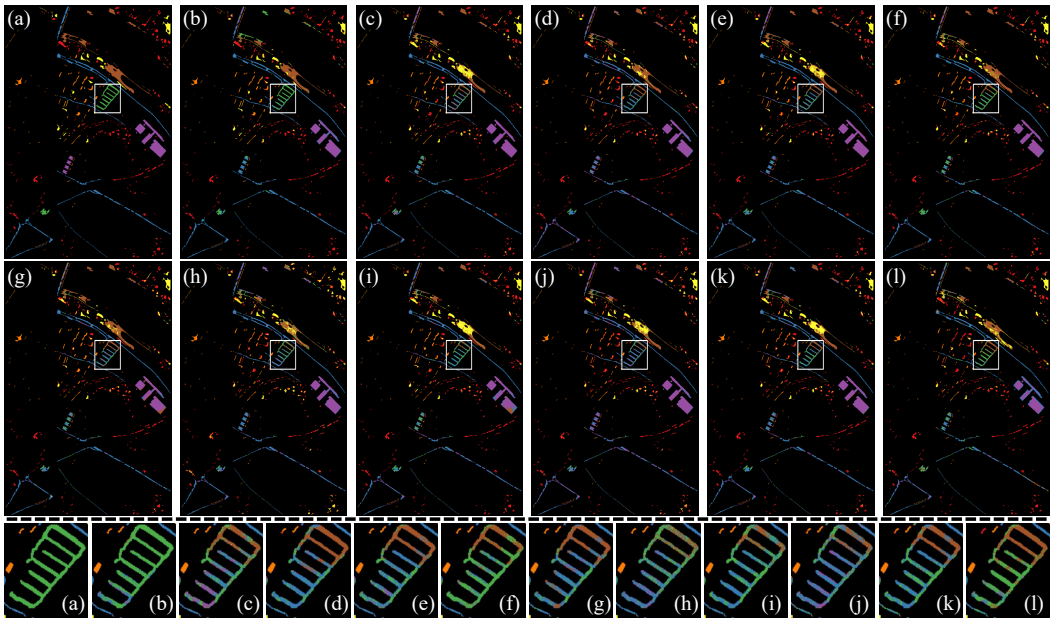


Figure 8: Visual comparison of classification results on Pavia. (a) Ground Truth. (b) Ours. (c) FDGNet. (d) S2AMSNet. (e) S2ECNet. (f) SPDDA. (g) ProRandConv. (h) StyDeSty. (i) TF-Cal. (j) SPCNet. (k) ConstStyle. (l) DDG.

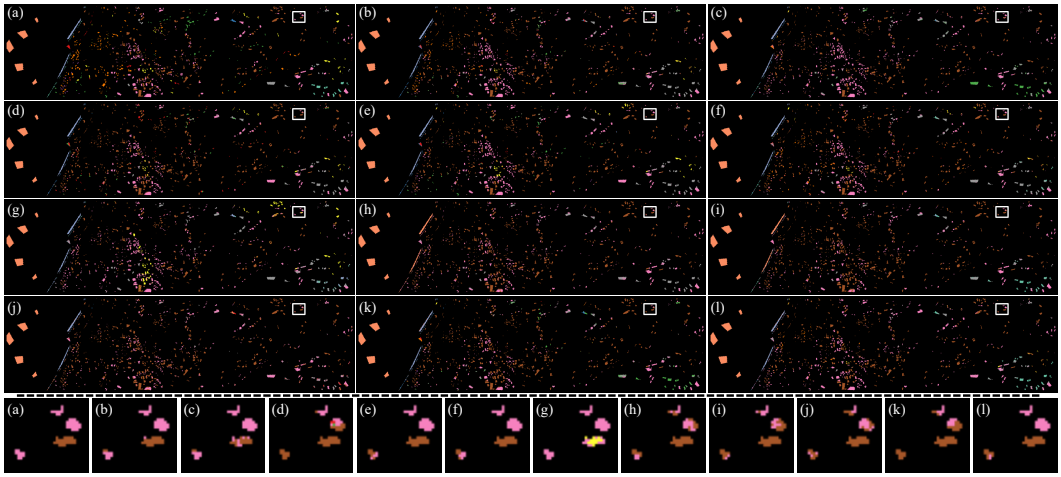


Figure 9: Visual comparison of classification results on HyRank. (a) Ground Truth. (b) Ours. (c) FDGNet. (d) S2AMSNet. (e) S2ECNet. (f) SPDDA. (g) ProRandConv. (h) StyDeSty. (i) TF-Cal. (j) SPCNet. (k) ConstStyle. (l) DDG.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract succinctly outlines the key contributions and scope of the paper, providing a clear overview of the research objectives and findings. The introduction effectively sets the stage by presenting the context, motivation, and objectives of the study, aligning with the subsequent content and outcomes presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are shown in the section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Experiments and Appendix A.3 sections, we fully disclose all necessary information to reproduce the main experimental results, ensuring transparency and clarity in the methodology employed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our code will be made publicly available upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In the Experiments and Appendix A.3 sections, we fully specify all necessary training and test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No result is accompanied by error bars, confidence intervals, or statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information of computer resources is shown in Appendix A.3 and A.4.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper complies with NeuroIPS ethical standards in all aspects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of the assets used in this paper are properly credited. The datasets, codebases, models, and other resources adopted or referenced in our work are cited accordingly, and their licenses and terms of use are explicitly mentioned and respected where applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code will be released later for the assets. The experimental datasets used in this paper are publicly available.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.