DGSSC: A Deep Generative Spectral-Spatial Classifier for Imbalanced Hyperspectral Imagery

Bobo Xi, *Member, IEEE,* Jiaojiao Li, *Member, IEEE,* Yan Diao, Yunsong Li, *Member, IEEE,*
Zan Li, *Senior Member, IEEE,* Yan Huang, *Member, IEEE,* and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—In recent years, hyperspectral image classification (HSIC) has achieved impressive progress with emerging studies on deep learning models. However, the classification performance downgrades due to the limited number of annotated samples, especially for minority classes. Notably, the imbalanced data dilemma is familiar in remote sensing hyperspectral image because the ground objects are commonly distributed without evenness. Therefore, this paper proposes a novel deep generative spectral-spatial classifier (DGSSC) for addressing the issues of imbalanced HSIC. Specifically, the DGSSC comprises three components, a two-stage encoder, a decoder, and a classifier, which are trained in an end-to-end manner. In particular, to exploit the abundant spectral-spatial features with relatively low computational complexity, the first stage of the encoder comprises successive three-dimensional (3D) and two-dimensional (2D) convolutions, exploring the spectral-spatial and deep spatial information. In addition, the second stage involves the deep latent variable model to achieve minority-class data augmentation. Furthermore, a patch distance-based reconstruction loss function is designed to facilitate the outputs of the decoder being more similar to the input 3D patch samples. The proposed DGSSC can outperform the state-of-the-art methods on three benchmark datasets, especially with its more robust prediction results. For instance, the DGSSC achieves a remarkable 97.85% mean overall accuracy with 0.24% standard deviation over ten independent runs with randomly selected imbalanced 1% training samples on the University of Pavia dataset.**

***Index Terms*—Imbalanced data, deep latent variable model, spectral-spatial features, hyperspectral image classification.**

I. Introduction

HYperspectral image classification (HSIC) assigns an individual label to each pixel in a captured hyperspectral imagery (HSI) scene. Due to fine-grained semantic parsing, HSIC has been applied to many practical applications, such as land-cover and land-use investigation, precision agriculture, and urban planning [1]–[3]. Thus, in the last few decades, researchers have devoted great efforts to HSIC to promote classification accuracy [4]–[8].

Remarkably, deep learning (DL)-based methods [9]–[11], such as convolutional neural networks (CNNs), have achieved significant breakthroughs compared to traditional methods such as the K-nearest neighbors (KNN) [12] method, support vector machines (SVMs) [13], sparse representation-based methods [14], low-rank-based models [15], [16]. This is mainly because the methods based on DL can extract the deep spectral-spatial features contained in HSI more effectively. For instance, based on the assumptions that adjacent pixels have the tendency to be the same land-cover types, Li *et al.* proposed a 3D-CNN [17] by taking the neighboring square region as input samples. It is able to simultaneously excavate the spectral and spatial features, thus acquiring excellent performance. Zhong *et al*. exploited the skip-connection architecture and presented a spectral-spatial residual network (SSRN) [18] for HSIC, which improves the generalization capability of the frameworks. Furthermore, to reduce the calculation burden of the 3D convolution (3D-Conv), Roy *et al*. proposed a hybrid spectral CNN (HybridSN) [19] comprising 3D-Conv and 2D convolution (2D-Conv), where the 2D-Conv can reduce the number of computations needed in the spectral dimension for the 3D-Conv. In addition, Xie *et al*. presented a multiscale densely connected fusion network, which adopts dense blocks to exploit the information contained in the multiscale neighboring patches, achieving results superior to the performances of several well-known HSIC methods [20].

However, the classification performance downgrades when small sample size (SSS) is used for training the abovementioned networks, especially for the minority classes that have fewer annotated samples than other categories [21]–[23]. This is because the classifier tends to be biased toward the majority classes, which harness more prior information in the training process [24]. In some applications, the minority classes in one scene are more valuable and require a higher-quality identification rate than the majority classes. For instance, forest fires are expected to be classified in the early stage when they occupy only a small region. Another example is that the valuable rare species that humans pay more attention to generally have small portions in vegetation-covered areas.

Taking the University of Pavia (UP) benchmark dataset[1] as

This work was supported in part by the National Nature Science Foundation of China 61901343 and the Youth Innovation Team of Shaanxi Universities. The project was also partially supported by the Fundamental Research Funds for the Central Universities JB190107 and the 111 project B08038. (Corresponding author: Jiaojiao Li)

B. Xi is with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China, and also with the National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China. (e-mail: xibobo1301@foxmail.com).

J. Li, Y. Diao, Y. Li, and Z. Li are with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China. (e-mail: jjli@xidian.edu.cn; ydiao@stu.xidian.edu.cn; ysli@mail.xidian.edu.cn; zanli@xidian.edu.cn).

Y. Huang is with the State Key Laboratory of Millimeter Waves, School of Information Science and Engineering, Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratory, Nanjing 211100, China. (e-mail: yellowstone0636@hotmail.com).

J. Chanussot is with the Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France, also with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China. (e-mail: jocelyn@hi.is).

[1] https://rslab.ut.ac.ir/data

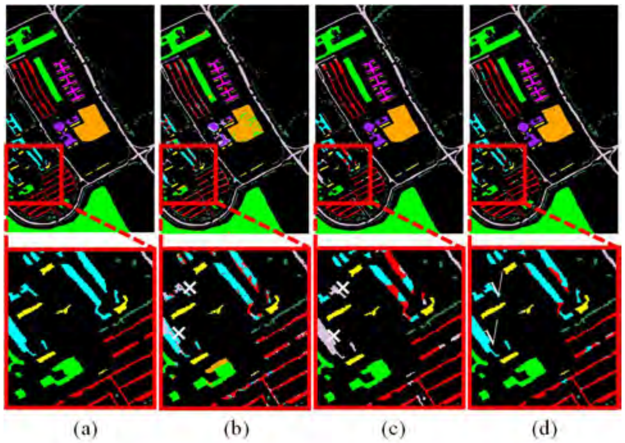


Fig. 1: The (a) groundtruth of the UP dataset, and classification maps obtained by (b) the 3D-CNN [17], (c) the SSRN [18], and (d) our proposed DGSSC, respectively, with only 1% training samples. Different colors indicate different classes. We highlight the most significant two small areas by × and ✓ in white for convenient comparison.

an example, Fig. 1 demonstrates the limitations of the baseline methods on imbalanced data problems, in which different colors indicate different classes. Specifically, Fig. 1 (a) depicts the groundtruth and only 1% of the labeled samples from each category are randomly selected as the training set, which fits the hard SSS and imbalanced conditions. Figs. 1 (b)-1 (d) illustrate the classification maps obtained by the benchmark 3D-CNN, the SSRN, and our proposed DGSSC, respectively. Fig. 1 (b) and (c) shows that more misclassifications occur for the minority class in blue (20 training samples and 2079 test samples), while the majority class in green (186 training samples and 18463 test samples) is much better recognized. In contrast, as shown in Fig. 1 (d), our proposed method can successfully recall more samples of the blue minority class.

With the emerging surge of DL research, deep generative models (DGMs) have been proposed to deal with complex imbalanced data [25]–[27]. One of the most representative algorithms is the conditional variational autoencoder (CVAE) [28], [29], which can be applied to the imbalanced data to capture the dimensional dependencies via the underlying variable, and then generate new samples from the learned latent variable without sophisticated adversarial learning [27]. However, if we directly employed the CVAE to produce synthetic samples, data generation and classification are isolated processes, leading to suboptimal classification performance. Additionally, the normal CVAE is designed for generating the simple one-dimensional (1D) signal and 2D natural images but not HSI with higher dimensionality. Consequently, the wealth of spectral-spatial information cannot be fully exploited. Moreover, the label information in the CVAE generally incorporates the training process by combining itself with the input image. Due to the high dimensionality of HSI, the labeled information may be overwhelmed in this manner. To overcome these drawbacks and inspired by the above instructive works, we proposed a new deep generative spectral-spatial classifier (DGSSC) to cope with the imbalanced HSIC by combining the minority-class data generation and classifier as a unified model.

In brief, the main contributions of this paper can be summarized as follows.

1) A novel DGSSC is presented for imbalanced HSIC. Specifically, the proposed DGSSC comprises a two-stage 3D encoder, a 3D decoder, and a classifier, which are trained in an end-to-end manner.

2) To better exploit the spectral-spatial information contained in HSI, the first stage of the encoder successively adopts 3D-Conv and 2D-Conv to investigate the spectral-spatial and deep spatial features, respectively. Moreover, we devise a patch distance-based reconstruction loss function to enhance the consistency of the decoded 3D samples with the inputs.

3) The framework involves the deep latent variable model and utilizes the label information in the low-dimensional feature space to achieve data augmentation, which makes the classifier more robust considering the imbalanced data dilemma. Experimental results on three benchmark datasets indicate that the proposed DGSSC can achieve better performance than the state-of-the-art methods, with higher mean accuracy and lower standard deviation over ten independent runs.

The remainder of the article is organized as follows. Section II briefly introduces the related works. Section III describes our proposed method in detail. Section IV shows the experimental results and analysis. Section V draws the conclusions.

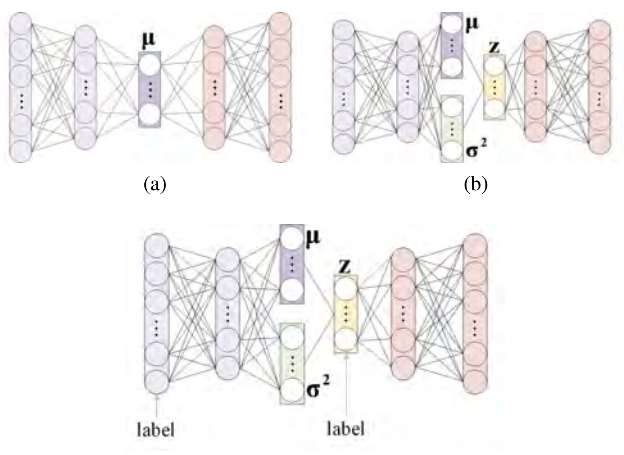
II. Related works

A. Imbalanced HSIC

Some prior works have tackled the imbalanced data problems for HSIC [30]–[32]. The most common approach is direct resampling (oversampling or undersampling) in the original data space. For example, Li *et al*. proposed an orthogonal complement subspace projection (OCSP) [33] method to oversample the small classes and undersample the large classes to balance all categories. Validation shows that it is more effective than the linear combination of the synthetic minority oversampling technique (SMOTE) [34], [35] on several conventional classifiers. However, the method is designed for a comparatively low-dimensional feature space (i.e., spectral curves) and has difficulty handling high-dimensional data such as 3D patch samples.

DL methods can effectively map high-dimensional information into a discriminative low-dimensional deep latent space driven by the original data. Inspired by this concept, researchers have introduced the resampling strategy into the DGMs, which captures the data distribution in the latent space by training a deep network and then synthesizes new samples by using the embedding space as a bridge. In particular, generative adversarial networks (GANs) are one of the most successful DGMs, which replace the complexity of resampling by searching for a Nash equilibrium of the generator and discriminator through adversarial training [36], [37]. For instance, Zhan *et al*. presented a 1D-GAN to exploit the spectral feature, which achieves promising results with a small

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3215513



(a)

(b)

(c)

Fig. 2: The diagrams of the (a) AE, (b) VAE, and (c) CVAE

sample size [38]. Zhu *et al*. proposed a 3D-GAN [39] with an auxiliary classifier, which is capable of investigating the spectral-spatial information and producing samples of specific classes leveraging the conditional label information. Zhong *et al*. employed conditional random fields after the trained discriminator to postprocess the classification maps, which obtains outperformed results [40]. However, relevant studies show that 3D-GAN fails to generate practicable samples for the minority class and collapses toward learning the basic mode of the realistic samples [41]. Recently, Roy *et al*. proposed a 3D generative adversarial minority oversampling method for imbalanced HSIC (HyperGAMO) [42], which devises a 3D convex generator based on the conditional GAN and produces new samples within the convex hull of the realistic minority-class samples. Notably, HyperGAMO is practical for producing synthetic samples and realizing classification in one stage rather than in two separated phases. Nonetheless, the convex hull of a minority class of HyperGAMO would be far from the true data distribution, which may generate less informative instances.

B. Conditional Variational Autoencoder

As a typical unsupervised model, an autoencoder (AE) has been verified to be effective in feature extraction and has been applied to many applications, such as object detection and fault diagnosis [43]–[45]. As shown in Fig. 2 (a), the AE aims to learn abstract features μ of input samples by approximating the expected output of the network to the inputs. To make the AE more robust to data noise, Kingma *et al*. proposed a variational autoencoder (VAE) [28], as depicted in Fig. 2 (b), which adds Gaussian noise $\sigma^{\mathbf{2}}$ to the encoded results and reconstructs the inputs by using this disturbed feature $\mathbf{z}$. Because high-dimensional HSI data generally have complex noise, a VAE is more suitable for processing HSI over the standard AE.

Given one sample $\mathbf{x}$, the VAE is supposed to generate $\hat{\mathbf{x}}$ that is similar to $\mathbf{x}$ from the latent space variable $\mathbf{z}$. Generally, $\mathbf{z}$ follows the prior distribution $q_\varphi(\mathbf{z})$ and $\hat{\mathbf{x}}$ is produced by the conditional distribution $q_\varphi(\mathbf{x}|\mathbf{z})$, where φ corresponds to the trainable parameters of the decoder. Because the posterior probability $p_\theta(\mathbf{z}|\mathbf{x})$ is difficult to estimate, it is commonly assumed to be a Gaussian distribution, where θ represents the learnable parameters of the encoder. Additionally, the mean vector μ and the deviation $\sigma^{\mathbf{2}}$ can be adaptively learned by the deep networks. Specifically, the Evidence Lower Bound (ELBO) is formulated as the object function of the VAE:

$$\begin{aligned} L_{VAE}(\mathbf{x};\theta,\varphi) = \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x})} q_\varphi(\mathbf{x}|\mathbf{z}) - \\ KL(p_\theta(\mathbf{z}|\mathbf{x})||q_\varphi(\mathbf{z})), \end{aligned} \tag{1}$$

in which $\mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x})} q_\varphi(\mathbf{x}|\mathbf{z})$ is the expectation of the generation probability of $\hat{\mathbf{x}}$, which is related to the reconstruction loss normally calculated by the mean square error (MSE). KL represents the KL-divergence [46] between $p_\theta(\mathbf{z}|\mathbf{x})$ and $q_\varphi(\mathbf{z})$. In practice, we assume that $p_\theta(\mathbf{z}|\mathbf{x})$ follows a multivariant Gaussian distribution $\mathcal{N}(\mu,\sigma^{\mathbf{2}})$, which is captured by the reparameterization strategy as

$$\mathbf{z} = \mu + \varepsilon\sigma, \tag{2}$$

where $\varepsilon \sim \mathcal{N}(0,1)$. In addition, $q_\varphi(\mathbf{z})$ is assigned to be a standard normal distribution to ensure its similarity with $p_\theta(\mathbf{z}|\mathbf{x})$.

In the VAE, μ can be regarded as the essential representation of the original data, and $\sigma^{\mathbf{2}}$ is considered the added Gaussian noise. In this manner, the extracted latent features can tolerate noise inference, which is beneficial for HSI data processing. Most importantly, the VAE can be used to generate samples by sampling noise that follows the distribution of $\mathbf{z}$. However, since the VAE is a totally unsupervised model, it cannot generate class-specific samples, which is not appropriate in imbalanced data problems.

Considering the VAE, the conditional VAE (CVAE) [29] is proposed by employing label information to generate class-specific samples. The diagram of the CVAE is demonstrated in Fig. 2 (c), and the loss function can be formulated as follows:

$$\begin{aligned} L_{CVAE}(\mathbf{x},\mathbf{y};\theta,\varphi) = \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y})} q_\varphi(\mathbf{x}|\mathbf{z},\mathbf{y}) - \\ KL(p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y})||q_\varphi(\mathbf{z}|\mathbf{y})), \end{aligned} \tag{3}$$

where $\mathbf{y}$ is the label of $\mathbf{x}$ that is in a one-hot style. Benefiting from the concrete label information, the CVAE can be utilized to compensate for the lack of samples in minority categories. Nevertheless, data generation and classification are completed in two separate phases, which may lead to suboptimal performance. In addition, as shown in Fig. 2 (c), the labeled information is involved in the training process at the beginning, which is normally concatenated with the image data. For high-dimensional HSI samples, the labeled information may be overwhelmed, and the samples generated for different classes may be entangled. Furthermore, it is necessary to build a model to thoroughly exploit the spectral-spatial information contained in the 3D patch samples but not the vector format shown in Fig. 2. To enhance the similarity between the generated samples and the input, a more effective distance metric for measuring 3D samples can be utilized to deploy new reconstruction loss functions.

III. Proposed method

In this section, a novel DGSSC is proposed for imbalanced HSIC and the architecture is illustrated in Fig. 3. To better describe the proposed method, we first introduce the necessary data preprocessing and some related notations.

A. Data Preprocessing

An HSI dataset $\mathbf{D} \in R^{H\times W\times L}$ is taken, where $H\times W$ represents the spatial dimension and L is the number of spectral bands. First, to reduce the computational burden, we conduct principal component analysis (PCA) on the original data and preserve *K* principal components (PCs) for the subsequent calculation. Noticeably, PCA can decrease the dimensionality and enhance the discrimination of the input feature, which is beneficial to both data generation and classification tasks [47]–[49]. Then, to exploit the spectral and spatial information contained in HSI, we select $S\times S$ sized neighborhoods around the target pixels to construct 3D patch samples. Namely, the training set can be represented as $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and each sample $\mathbf{x}_i \in \mathbb{R}^{S\times S\times K}$. Correspondingly, the labels of $\mathbf{X}$ can be denoted $\mathbf{Y} = \{y_1, ..., y_N\}$, where $y_i \in \{1, 2, ..., C\}$ and C is the number of classes. The total number of training samples can be calculated as $\sum_{i=1}^{C} n_i$, where n_i is the size of the *i*th class. Notably, there may be a large difference among $\{n_1, ..., n_C\}$ in our focused imbalanced classification. In such situations, it is more challenging for a classifier to predict correctly when a new sample comes, especially for the minority classes.

B. Frameworks of the DGSSC

As shown in Fig. 3, the proposed DGSSC mainly comprises a two-stage 3D encoder denoted $f_\theta = \{f_{\theta_1}, f_{\theta_2}\}$, a 3D decoder referred to as f_φ, and a classifier represented as g_η, where θ, φ, and η are trainable parameters of the networks. Among them, the encoder is supposed to simultaneously achieve two goals in an adversary-free manner: matching the encoded distribution of training examples to the prior distribution as measured by a specified divergence, while ensuring that the latent variables provided to the decoder are informative enough to reconstruct the encoded training examples. The decoder can not only reconstruct the input realistic samples but also generate artificial samples from the randomly sampled noise that follows the distribution of the latent variable of the real samples. Suppose we have an arbitrary sample denoted $\mathbf{x}_i \in \mathbb{R}^{S\times S\times K}$ and the corresponding label is y_i. In this subsection, we take $S = 13$ and $K = 20$ as an example to detail the proposed network, which is displayed in Table I. The corresponding structures of the three portions are described as follows:

1) Two-stage 3D Encoder: The first stage of the encoder is denoted f_{θ_1}, where θ_1 is the trainable kernel parameter. To exploit the spectral-spatial information and the complementary deep spatial features, f_{θ_1} is mainly composed of 3D and 2D convolutional operations. In the formula, the 2D convolution can be expressed as follows:

$$m_{ij}^{xy} = b_{ij} + \sum_c \sum_{\tau=0}^{h_i-1} \sum_{\sigma=0}^{w_i-1} k_{ijc}^{\tau\sigma} \times m_{(i-1)c}^{(x+\tau)(y+\sigma)}. \quad (4)$$

The 3D convolution can be calculated as

$$m_{ij}^{xyz} = b_{ij} + \sum_c \sum_{\tau=0}^{h_i-1} \sum_{\sigma=0}^{w_i-1} \sum_{\delta=0}^{d_i-1} k_{ijc}^{\tau\sigma\delta} \times m_{(i-1)c}^{(x+\tau)(y+\sigma)(z+\delta)}, \quad (5)$$

where m refers to the value of the feature map and (x, y, z) are the location indices for the *j*th feature map of the *i*th layer. (τ, σ, δ) are indices of the kernels. c is the *c*th feature map of the $(l-1)$th layer. b is the bias parameter. (h, w, c_{in}, c_{out}) and $(h, w, d, c_{in}, c_{out})$ denote the learnable 2D and 3D convolution kernels, where (h, w, d) is the kernel size in the height, width, and depth dimensions, respectively. c_{in} and c_{out} are the numbers of input and output channels, respectively.

According to Ref [50], a convolutional filter with kernel size $(3, 3, 3)$ and stride $(1, 1, 1)$ can capture promising features for action recognition. Thus, we employ three 3D convolutional layers with a kernel size of $(3, 3, d)$, where $d = 7, 5$, and 3, and the numbers of the three kernels are M, $2M$, and $4M$, respectively. The variant depth dimension can facilitate network learning of multiscale channel features. After that, we reshape the feature into a cube and further explore the deep spatial information by a 2D convolution. The kernel is also assigned to $(3, 3)$ with a $(1, 1)$ stride, and the kernel number is experimentally set to 128 to reduce the feature dimension. Note that batch normalization [51] and an ReLU [52] nonlinear function follow each convolutional operation to accelerate the training process and alleviate the overfitting problem. Then, we flatten the feature into a vector denoted $f_{\theta_1}(\mathbf{x}_i)$ as the output of the first-stage encoder.

In the second stage of the encoder, the label y_i is first encoded in a one-hot fashion as $\mathbf{y}_i$ and then concatenated with $f_{\theta_1}(\mathbf{x}_i)$, which can be expressed as

$$\mathbf{H}_1 = f_{\theta_1}(\mathbf{x}_i)||\mathbf{y}_i, \quad (6)$$

where || is the concatenate operation. Then, a fully connected (FC) layer is employed to summarize the information in $\mathbf{H}_1$, which can be represented as

$$\mathbf{H}_2 = \mathbf{H}_1\mathbf{W}_1 + b_1, \quad (7)$$

in which $\mathbf{W}_1$ is the trainable parameter of the FC layer and b_1 is the bias. Afterward, two independent FCs are adopted to fit the mean and variance vector of the multivariate Gaussian distribution for the latent variable, which is formulated as

$$\mu(\mathbf{x}_i, \mathbf{y}_i) = \mathbf{H}_2\mathbf{W}_2 + b_2, \quad (8)$$

$$\sigma(\mathbf{x}_i, \mathbf{y}_i) = \mathbf{H}_2\mathbf{W}_3 + b_3, \quad (9)$$

where $\mathbf{W}_2$ and $\mathbf{W}_3$ are the trainable parameters of the two FC layers, and b_2 and b_3 are the biases. Finally, we can obtain the latent variable by using the reparameterization technique:

$$\mathbf{z}_i = \mu + \varepsilon\sigma, \quad (10)$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ and $\mathbf{z}_i$ represents the deep latent variable.

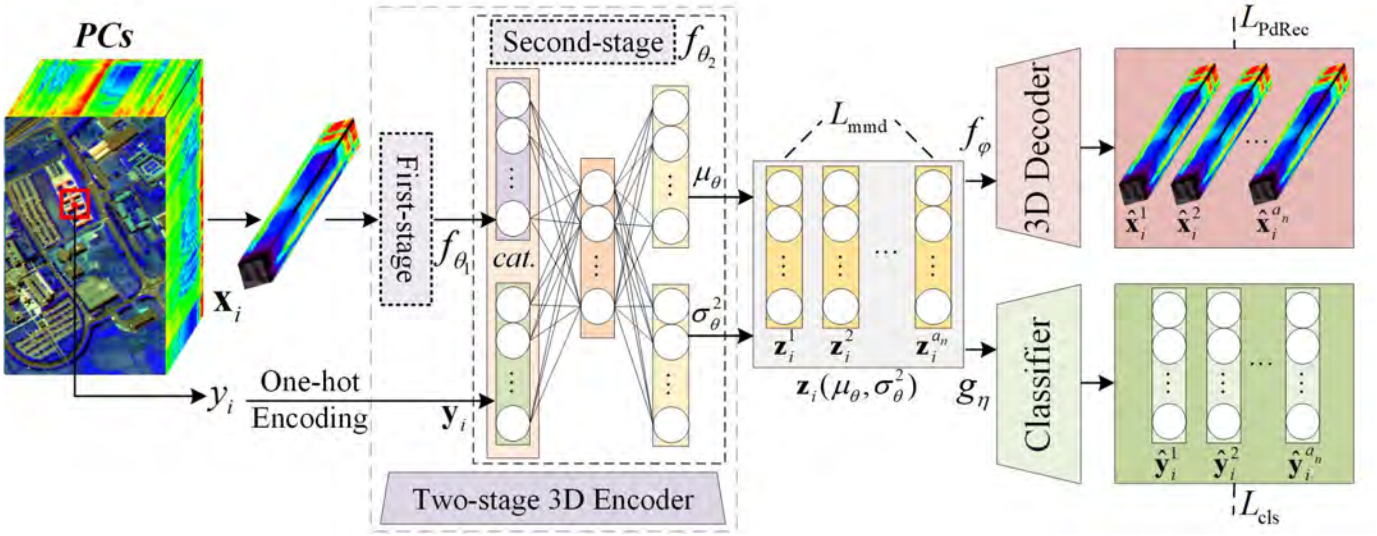


Fig. 3: The architecture of the proposed DGSSC for imbalanced hyperspectral imagery. The encoder, decoder, and classifier are trained in an end-to-end fashion, driven by the integrated loss of L_{PdRec}, L_{mmd}, and L_{cls}.

TABLE I: THE DETAILED STRUCTURE AND PARAMETERS OF THE PROPOSED DGSSC

Layer_name	Input shape	Kernel size	Stride	Output shape
Two-stage 3D Encoder				
Conv3D_1	(13,13,20,1)	(3,3,7,M)	(1,1,1)	(11,11,14,M)
Conv3D_2	(11,11,14,M)	(3,3,5,2M)	(1,1,1)	(9,9,10,2M)
Conv3D_3	(9,9,10,2M)	(3,3,3,4M)	(1,1,1)	(7,7,8,4M)
Reshape_1	(7,7,8,4M)	-	-	(7,7,8*4M)
Conv2D_1	(7,7,8*4M)	(3,3,128)	(1,1)	(5,5,128)
Flatten	(5,5,128)	-	-	(3200,1)
Concat	(3200,1)$\vert\vert$(C,1)	-	-	(3200+C,1)
FC_1	(3200+C,1)	-	-	(3200,1)
FC_2	(3200,1)	-	-	(64,1)
FC_3	(3200,1)	-	-	(64,1)
3D Decoder				
FC_1	(64,1)	-	-	(3×3×7×4N,1)
Reshape_1	(3×3×7×4N,1)	-	-	(3,3,7,4N,1)
DeConv3D_1	(3,3,7,4N)	(3,3,3,2N)	(2,2,2)	(6,6,10,2N)
DeConv3D_2	(6,6,10,2N)	(3,3,3,N)	(2,2,2)	(12,12,20,N)
DeConv3D_3	(12,12,20,N)	(2,2,1,1)	(1,1,1)	(13,13,20,1)
Classifier				
FC_1	(64,1)	-	-	(C,1)

2) 3D Decoder and Classifier: The decoder f_φ aims to reconstruct the 3D patch samples by using the informative latent variable $\mathbf{z}_i$. As shown in Table I, we first employ an FC layer to transform the latent feature into a long vector as $(3\times3\times7\times4N, 1)$ and then reshape it to a set of feature cubes as $(3, 3, 7, 4N, 1)$. Next, three 3D deconvolutional operations with $2N$, N, and 1 kernels are adopted to generate the reconstructed result with the same size of the input, i.e., $(13, 13, 20, 1)$. The detailed convolutional kernel size and stride are shown in Table I. In terms of classifier g_η, we employ an FC layer to comprehensively parse the semantic information contained in $\mathbf{z}_i$, and the output dimension equals the number of categories C.

C. Training and Testing Processes

In the training process, considering the imbalanced data, we augment the minority classes in the latent space. For better understanding, the adopted deep latent variable model is illustrated in Fig. 4, which is the schematic of Fig. 3. The dotted circles represent the augmented information. Specifically, if sample $\mathbf{x}_i$ belongs to any other class except the largest class, we will sample a set of latent codes $\{\mathbf{z}_i^j\}_{j=1}^{a_n}$ by using (10) to extend the latent feature $\mathbf{z}_i$ of $\mathbf{x}_i$, where $a_n = \max\{1, r \times \frac{n_{\max}}{n_c}\}$ is the number of oversampling latent codes for each instance of the cth class. r is a sampling rate in the scope of $[0, 1]$, and $n_{\max} = \max\{n_1, ..., n_C\}$ is the size of the largest class. The oversampling variables $\mathbf{z}_i^j$ will be used to generate synthetic data $\hat{\mathbf{x}}_i^j$ in the original feature space via the decoder f_φ and to infer the predicted label $\hat{\mathbf{y}}_i^j$ by the classifier g_η for more robust results.

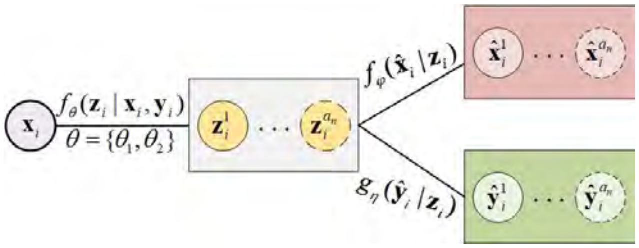

$\mathbf{x}_i$
$f_\theta(\mathbf{z}_i \mid \mathbf{x}_i, \mathbf{y}_i)$
$\theta = \{\theta_1, \theta_2\}$
$\mathbf{z}_i^1$
$\mathbf{z}_i^{a_n}$
$f_\varphi(\hat{\mathbf{x}}_i \mid \mathbf{z}_i)$
$g_\eta(\hat{\mathbf{y}}_i \mid \mathbf{z}_i)$
$\hat{\mathbf{x}}_i^1$
$\hat{\mathbf{x}}_i^{a_n}$
$\hat{\mathbf{y}}_i^1$
$\hat{\mathbf{y}}_i^{a_n}$

Fig. 4: Data augmentation for the minority classes by adopting the deep latent variable model.

To ensure the stability of the model, the maximum mean discrepancy (MMD) (L_{mmd}) [53] is employed to enhance the similarity between the conditional distribution of the latent variable $\mathbf{z}_i$ and the prior Gaussian distribution. Because the MMD can form an unbiased U-estimator, it is combined with the subsequent stochastic gradient descent (SGD) method [53]. Specifically, both the realistic and synthetic latent variables are fed into the 3D decoder to produce the reconstructed samples, which approximate the original instance by the patch distance-based reconstruction loss (L_{PdRec}). Moreover, the encoded and sampled latent variables are classified in the label space with the cross-entropy-based classification loss function (L_{cls}). In summary, the total objective function in the training process can be expressed as

$$L_{\text{cls}} = -\sum_{c=1}^{C} \mathbf{y}_{ic} \log \widehat{\mathbf{y}}_{ic}, \tag{11}$$

$$L_{\text{mmd}} = D_{\mathbf{z}}(p_{\theta}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i)||q_{\varphi}(\mathbf{z})), \tag{12}$$

$$L_{\text{total}} = L_{\text{cls}} + \lambda \cdot L_{\text{mmd}} + L_{\text{PdRec}}, \tag{13}$$

where $\mathbf{y}_{ic}$ and $\hat{\mathbf{y}}_{ic}$ represent the true and predicted probabilities of the cth class, respectively. λ is a weight factor of L_{mmd}. L_{PdRec} is detailed in the next subsection.

After obtaining the well-trained parameters $\{\theta, \varphi, \eta\}$, an optimal deep generative model is built, where the conditional distribution $p(\mathbf{y}_c|\mathbf{x})$ for the cth category lies in a low-dimension manifold. Then, when a new sample $\bar{\mathbf{x}}$ comes during the testing process, the predicted label $\bar{\mathbf{y}}$ can be acquired by approximating Bayes' rule with importance sampling $\bar{\mathbf{z}}_c^s \sim p(\bar{\mathbf{z}}|\bar{\mathbf{x}}, \mathbf{y}_c)$:

$$p(\bar{\mathbf{y}}|\bar{\mathbf{x}}) \approx \text{softmax}_{c=1}^{C}\{\log \frac{1}{S} L_{\text{total}}(\bar{\mathbf{x}}, \mathbf{y}_c, \bar{\mathbf{z}}_c^s; \theta, \varphi, \eta)\}, \tag{14}$$

where S is the sampling size.

D. Patch Distance-Based Reconstruction Loss

As mentioned above, the reconstruction loss can force the reconstructed/generated samples to be close to the original data, which plays a significant role in ensuring the stability of the generative model. The typical approaches employ the MSE as the discrepancy constraint, which is represented as

$$L_{\text{Rec}} = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2. \tag{15}$$

For HSI, the input $\mathbf{x}_i$ and the reconstructed/generated $\hat{\mathbf{x}}_i$ are both 3D patches, and measuring the similarity of such high-dimensional data by Euclidean distance may not work well. Additionally, the spatial correlations are not exploited by the primitive point-by-point calculations. Thus, we introduce the image patch distance [54] into the reconstruction loss.

Assume that $\mathbf{p}_{ik}$ and $\hat{\mathbf{p}}_{ik}$ are arbitrary pixel vectors in the patch $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$, respectively. Then, the distance between pixel $\mathbf{p}_{ik}$ and patch $\hat{\mathbf{x}}_i$ can be calculated as

$$d(\mathbf{p}_{ik}, \hat{\mathbf{x}}_i) = \min_{\hat{\mathbf{p}}_{ik} \in \hat{\mathbf{x}}_i} d(\mathbf{p}_{ik}, \hat{\mathbf{p}}_{ik}), \tag{16}$$

where $d(\cdot)$ is the Euclidean distance used to measure the spectral discrepancy. Symmetrically, the distance between pixel $\hat{\mathbf{p}}_{ik}$ and patch $\mathbf{x}_i$ can be represented as

$$d(\hat{\mathbf{p}}_{ik}, \mathbf{x}_i) = \min_{\mathbf{p}_{ik} \in \mathbf{x}_i} d(\hat{\mathbf{p}}_{ik}, \mathbf{p}_{ik}). \tag{17}$$

Then, the distance between $\mathbf{p}_{ik}$ and $\hat{\mathbf{p}}_{ik}$ is defined as

$$d(\mathbf{p}_{ik}, \hat{\mathbf{p}}_{ik}) = \max(d(\mathbf{p}_{ik}, \hat{\mathbf{x}}_i), d(\hat{\mathbf{p}}_{ik}, \mathbf{x}_i)). \tag{18}$$

Finally, we can obtain the patch distance-based reconstruction loss as

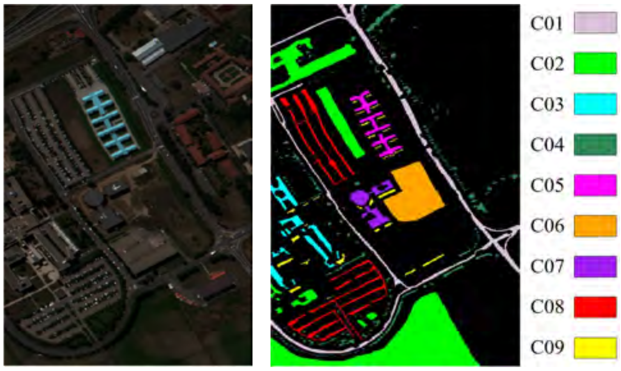
$$L_{\text{PdRec}} = d(\mathbf{x}_i, \hat{\mathbf{x}}_i) = \sum_{k=1}^{S \times S} d(\mathbf{p}_{ik}, \hat{\mathbf{p}}_{ik}). \tag{19}$$

IV. Experimental results and analysis

In this section, the classification performance of the proposed DGSSC is evaluated on three widely used hyperspectral datasets, including the UP, Chikusei, and Loukia datasets. Furthermore, we adopt average accuracy (AA), overall accuracy (OA), and kappa coefficient (k) as criteria metrics.

A. Experimental Datasets

1) University of Pavia: The first UP dataset, captured by the Reflective Optics Spectrometer (ROSIS) sensor over Pavia University, consists of 610×340 pixels, and the geometric resolution is 1.3 m/pixel. After discarding the noisy bands, 103 channels remained, covering the range from 0.43 to 0.86 μm. Fig. 5 (a) and (b) show the composite false color image and groundtruth. From a total of 42776 samples, the available ground reference map includes nine classes. Table II lists the numbers of randomly selected 1% training portions from each class.

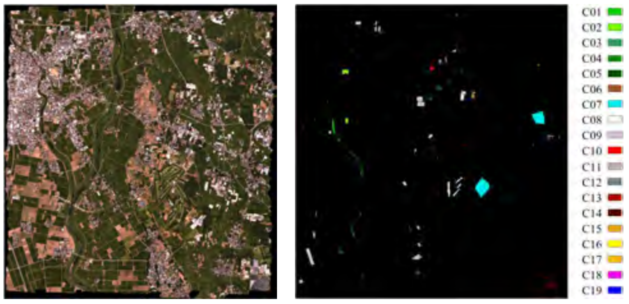


(a) (b)

Fig. 5: UP dataset. (a) False-color composite image (RGB-R: 50, G: 30, B: 20), (b) groundtruth.

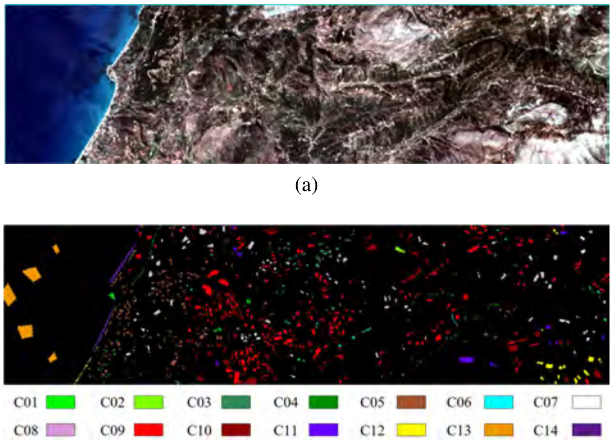
2) Chikusei: The second dataset, Chikusei, is gathered by a Headwall Hyperspec-VNIR-C imaging sensor over Chikusei, Ibaraki, Japan. This image has a large scenario comprising 2517×2335 pixels, and mainly includes agricultural and urban areas with a spatial resolution of 2.5 m/pixel. A total of 128 bands were derived, ranging from 0.363 to 1.018 μm, after spectral binning was performed to increase the SNR. Referring to Fig. 6 (a) and (b), there are nineteen labeled land-cover categories in the image, which are listed in Table III. In addition, the training samples are also randomly selected as 1% from each class.

3) Loukia: The last dataset, Loukia, was obtained from the Hyperion sensor carried by the National Aeronautics and Space Administration (NASA) Earth Observing 1 (EO-1) satellite. Due to the lack of illumination, low sensitivity, water absorption and overlap between the visible and near-infrared (VNIR) and shortwave infrared (SWIR) spectrometers, a total of 176 bands were derived after the removal process. Notably, the benchmark HyRANK dataset contains two HSI datasets, Dioni and Loukia, with twelve and fourteen labeled categories, respectively. In Fig. 7, our experiment chooses the Loukia



(a) (b)

Fig. 6: Chikusei dataset. (a) False-color composite image (RGB-R: 49, G: 33, B: 19), (b) groundtruth.



(a)

(b)

Fig. 7: Loukia dataset. (a) False-color composite image (RGB-R: 23, G: 11, B: 7), (b) groundtruth.

dataset because it has more land-cover types, and 5% of each category is randomly selected as the training set, as shown in Table IV.

TABLE II: LABELED LAND-COVER TYPES AND NUMBERS OF TRAINING, TEST, AND TOTAL SAMPLES FOR UP DATASET

No.	Land-cover	Training	Test	Total
01	Asphalt	66	6565	6631
02	Meadows	186	18463	18649
03	Gravel	20	2079	2099
04	Trees	30	3034	3064
05	Painted metal Sheets	13	1332	1345
06	Bare Soil	50	4979	5029
07	Bitumen	13	1317	1330
08	Self-Blocking Bricks	36	3646	3682
09	Shadows	9	938	947
	Total	423	42353	42776

B. Parameter Tuning

To maximize the utility of the proposed method, we adjust the parameters of the framework in this section.

TABLE III: LABELED LAND-COVER TYPES AND NUMBERS OF TRAINING, TEST, AND TOTAL SAMPLES FOR CHIKUSEI DATASET

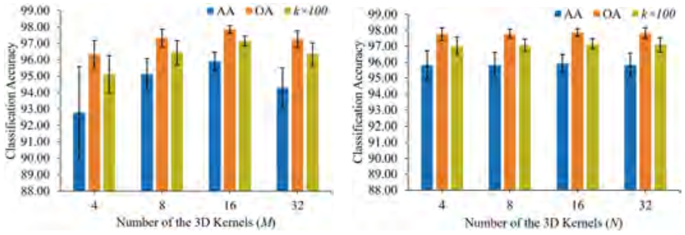
No.	Land-cover	Training	Test	Total
01	Water	28	2817	2845
02	Bare Soil (school)	28	2831	2859
03	Bare Soil (park)	2	284	286
04	Bare Soil (farmland)	48	4804	4852
05	Natural Plants	42	4255	4297
06	Weeds in Farmland	11	1097	1108
07	Forest	205	20311	20516
08	Grass	65	6450	6515
09	Rice Field (grown)	133	13236	13369
10	Rice Field (first stage)	12	1256	1268
11	Row Crops	59	5902	5961
12	Plastic House	21	2172	2193
13	Manmade (non-dark)	12	1208	1220
14	Manmade (dark)	76	7588	7664
15	Manmade (blue)	4	427	431
16	Manmade (red)	2	220	222
17	Manmade Grass	10	1030	1040
18	Asphalt	8	793	801
19	Paved Ground	1	144	145
	Total	767	76825	77592

TABLE IV: LABELED LAND-COVER TYPES AND NUMBERS OF TRAINING, TEST, AND TOTAL SAMPLES FOR LOUKIA DATASET

No.	Land-cover	Training	Test	Total
01	Dense Urban Fabric	14	274	288
02	Mineral Extraction Sites	3	64	67
03	Non Irrigated Arable Land	27	515	542
04	Fruit Trees	4	75	79
05	Olive Groves	70	1331	1401
06	Broad_leaved Forest	11	212	223
07	Coniferous Forest	25	475	500
08	Mixed Forest	54	1018	1072
09	Dense Scleorophyllous Vegetation	190	3603	3793
10	Sparce Sclerophyllous Vegetation	140	2663	2803
11	Sparcely Vegetated Areas	20	384	404
12	Rocks and Sand	24	463	487
13	Water	70	1323	1393
14	Coastal Water	23	428	451
	Total	675	12828	13503

As described in Section III-A, we first arrange the pixel-level samples into cubes to involve the spectral-spatial information contained in HSI. According to the classification performance, we experimentally assign the number of PCs as 20 for the UP and Chikusei scenes and 50 for the Loukia dataset. In addition, the spatial neighborhood size for the UP and Loukia datasets is set to 13, and for the Chikusei dataset, it is set to 15. The experiments and analysis of these two parameters are detailed in Section IV-F.

In addition, the numbers of kernels M in the two-stage 3D encoder and N in the 3D decoder are crucial since they influence the representative capability of the embedding features. Thus, we measure the classification performance with different M and N values to explore the most appropriate network architecture. In detail, the AA, OA, and k with various M and N values of $\{4, 8, 16, 32\}$ for the UP dataset are delineated in Fig. 8. The optimal performance appears when M and N are both 16; thus, they are set to 16 in the



(a) (b)

Fig. 8: The classification accuracy on UP dataset with different number of 3D kernels in (a) the two-stage 3D encoder and (b) the 3D decoder.

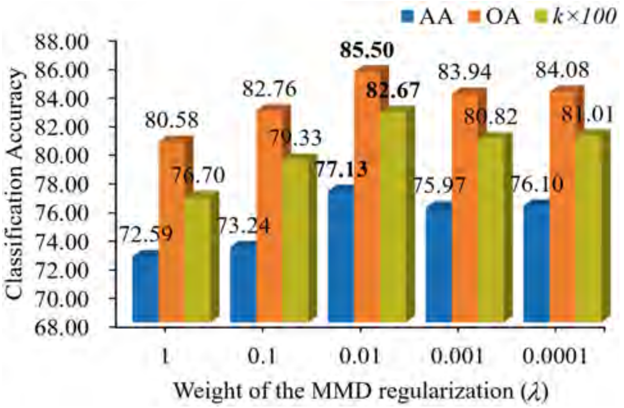


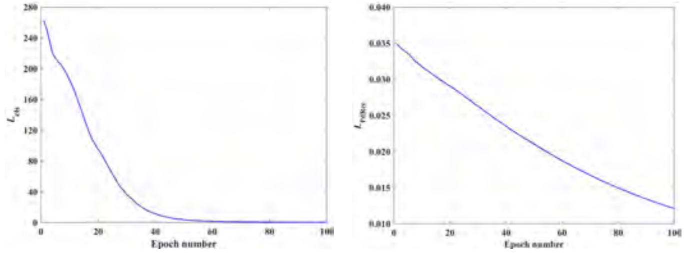
Fig. 9: The classification accuracy on Loukia dataset with different weight of the MMD regularization.

experiments. Analogously, the M and N for the Chikusei and Loukia datasets are also 16. Moreover, the augmentation rate working in the latent space is experimentally set to 0.4 for the three datasets, which is analyzed in Section IV-C.

Furthermore, we validate the performance with different weights of the MMD loss, which is denoted λ in (13). Specifically, λ varies in $\{1,0.1,0.01,0.001,0.0001\}$, and the corresponding AA, OA, and k for the Loukia dataset are depicted in Fig. 9. The classification performance is first promoted and then decreases, and it achieves the best performance when λ equals 0.01. This implies that moderate regularization between the distribution of the latent variable and the prior Gaussian is significant to ensure a stable learning process. Thus, we experimentally set it to 0.01 for the three datasets. Regarding the optimization of the network, we adopt the Adam optimizer to iteratively train the kernel parameters and the bias of the hidden neurons in each layer. Specifically, the learning rate is experimentally set to 1e-5 and the epoch number is validated as 100 for the UP and Chikusei scenes, and 200 for the Loukia dataset. Taking the UP dataset as an example, we display the classification loss and reconstruction loss over iterations in Fig. 10. The network can converge smoothly.

C. Ablation Study

In the proposed frameworks, the most significant innovation is the data augmentation in the latent variable space used to conquer the imbalanced data problem and the devised patch distance-based reconstruction loss used to enhance the consistency of the original and reconstructed samples. Thus, an ablation study is carried out to demonstrate the superiority of these two components. First, the parameter influencing the data augmentation process is the augmentation rate r working in the latent space. Thus, we evaluate the proposed DGSSC with varying r values in the range $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and the classification accuracy on the Chikusei dataset is delineated in Fig. 11. The classification results are better when r is larger than 0, especially in terms of the AA index, which indicates the positive effect of the presented augmentation strategy embedded in the networks. Additionally, the classification performance reaches the peak when r equals 0.4. This implies that an appropriate augmentation rate is required to ensure an adequate increment is achieved to remedy the data imbalance issues and avoid excessive data disturbance at the same time.



(a) (b)

Fig. 10: The classification loss and reconstruction loss over iterations on UP dataset. (a) L_{cls}, (b) L_{PdRec}.

TABLE V: CLASSIFICATION ACCURACY OF MSE-BASED AND PD-BASED RECONSTRUCTION LOSS ON THE EXPERIMENTAL DATASETS

		UP	Chikusei	Loukia
AA (%)	MSE	95.23±0.98	93.06±1.77	72.78±1.70
	PD	**95.90±0.57**	**94.00±1.58**	**77.13±1.21**
OA (%)	MSE	97.58±0.40	99.17±0.21	81.07±0.79
	PD	**97.85±0.24**	**99.23±0.16**	**85.50±0.55**
k × 100	MSE	96.79±0.53	99.04±0.24	77.43±0.90
	PD	**97.14±0.31**	**99.11±0.19**	**82.67±0.64**

Moreover, the classification performance on the three datasets with different reconstruction losses is displayed in Table V. The patch distance (PD)-based reconstruction loss achieves superior classification performance with higher mean accuracy and lower standard deviation than the MSE-based loss on the experimental datasets. In particular, the AA, OA, and k of the Loukia dataset obtain substantial improvements of 4.35%, 4.43%, and 5.24%, respectively. Considering the size of the input sample and the experimental results of different datasets, it can be speculated that the PD-based loss function can play a more important role in the reconstruction of higher-dimensional data.

D. Comparison with Other Methods

Some experiments are conducted to comprehensively compare the proposed DGSSC with other state-of-the-art algorithms. Among them are traditional approaches such as classic

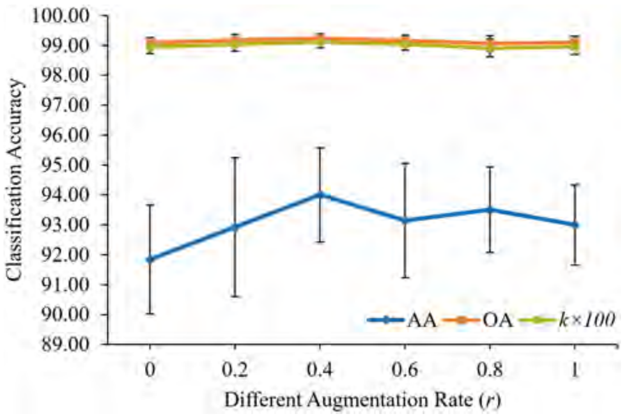


Fig. 11: The classification accuracy on Chikusei dataset with different augmentation rate.

machine learning-based methods, namely, the SVM with radial basis function kernel (SVM-RBF) [13] and 3D-CNN [17] method, which only adopts 3D-Convs to extract spectral-spatial features. In addition, the residual network SSRN [18] and the HybridSN [19], which combines 2D-Conv and 3D-Conv for HSI feature extraction, are also involved. Meanwhile, we compare the proposed framework with highly related algorithms, i.e., the 3D-GAN [39] and the HyperGAMO [42], which can extract information from imbalanced HSIs. The hyperparameters of the competitors are assigned as the corresponding citations. To ensure the reliability of the experimental results, all methods are repeated ten times with randomly selected training samples, and the final accuracy values are the mean results along with standard deviations, which are reported in Tables VI to VIII. To better illustrate the advantages of the proposed method in classifying minority classes, we define the category with a training sample size less than the average sample size as the comparative minority class (CMC) and highlight it in bold. Additionally, for the convenience of comparison, the highest and suboptimal accuracies of each row are shown in bold font and underlined, respectively.

Table VI presents the accuracy for each land-cover class of the UP dataset, and the last three rows calculate the AA, OA, and k values for all compared methods. The proposed DGSSC achieves the highest AA, OA, and k, with the lowest standard deviations, verifying the superiority and stability of the DGSSC. Regarding the class-specific accuracy, the DGSSC obtains optimal accuracy on the 3rd, 5th, and 7th CMCs, which demonstrates that the augmentation strategy and appropriate loss function working together can extract the features of the minority class more effectively. In contrast, HyperGAMO acquires the second-best results on the 3rd, 8th, and 9th CMCs and AA due to the convex 3D patch generator oversampling the data points from the minority classes. Additionally, the SSRN achieves competitive OA and k, which is attributed to the skip connections of the networks. However, the obtained AA is not that satisfying owing to the relatively poor performance on the minority classes.

Table VII reports the accuracy of the Chikusei dataset. Our proposed DGSSC outperforms the compared methods in terms of AA, OA, and k. Specifically, it achieves the optimal results on the 1st, 10th CMCs and suboptimal results on the 2nd, 12th, 17th, 18th, and 19th CMCs, which confirms the effectiveness of the DGSSC in identifying the minority classes. In addition, some observations similar to the UP dataset can be found on the SSRN method. The obtained OA and k are favorable, but the AA is unsatisfactory because of the poor classification performance on the minority classes, especially for the 3rd and 19th categories, with 0% accuracy. In addition, the classic SVM-RBF performs well compared to the other DL-based approaches, which indicates the scalability of this traditional milestone method.

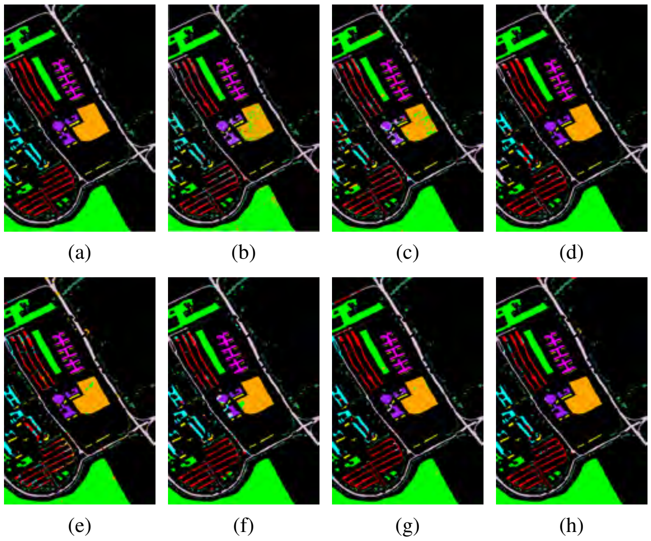


Fig. 12: Classification maps obtained via various algorithms for UP dataset. (a) Groundtruth. (b) SVM-RBF. (c) 3D-CNN. (d) SSRN. (e) HybridSN. (f) 3D-GAN. (g) HyperGAMO. (h) Proposed DGSSC. Best zoomed-in view.

Table VIII displays the classification accuracy of various methods for the Loukia dataset. By comparison, the SVM-RBF cannot produce acceptable classification results in such training and test splits on the Loukia dataset. The accuracies of all methods are comparatively lower than those on the other two datasets. This result is perhaps due to the resolution of this Spaceborne data being lower than the resolutions of the Airborne UP and Chikusei datasets. Even in this case, the AA, OA, and k of the DGSSC again achieve better values than the other algorithms, and it obtains the best accuracy on the 2nd and 3rd CMCs. Moreover, although HyperGAMO achieves encouraging classification results on the CMCs, the DGSSC outperforms it in AA, OA, and k, with values of 1.15%. 1.50%, and 1.84%, respectively. The stable performance on different datasets indicates that the deep latent variable model involved in the framework is effective for data augmentation and feature extraction.

Considering subjective evaluation, we depict the classification maps produced by different methods of each dataset from Figs. 12 to 14, which are consistent with Tables VI to VIII, respectively. To facilitate comparison, the corresponding groundtruths are also placed together. As seen in Figs. 12-

TABLE VI: CLASSIFICATION ACCURACY OF DIFFERENT ALGORITHMS FOR UP DATASET

No.	SVM-RBF [13]	3D-CNN [17]	SSRN [18]	HybridSN [19]	3D-GAN [39]	HyperGAMO [42]	DGSSC
01	86.91±2.15	92.45±1.83	**98.87±2.31**	95.09±2.49	94.49±2.01	95.06±2.43	98.45±0.51
02	96.08±1.87	98.50±0.68	99.75±0.46	99.28±1.32	98.93±0.66	**99.96±0.12**	99.94±0.07
03	65.95±7.42	53.13±6.36	76.31±12.01	79.46±6.43	78.35±9.82	80.39±7.57	**84.69±4.49**
04	86.64±4.07	**95.74±1.14**	93.72±4.72	90.61±4.68	86.33±3.58	85.55±5.45	93.90±2.16
05	98.88±0.68	95.86±3.57	89.54±29.80	99.27±1.54	85.15±2.21	96.06±0.95	**99.78±0.35**
06	76.66±4.82	86.84±3.93	**100.00±0.00**	93.71±2.56	98.30±1.12	92.04±2.47	98.94±0.50
07	77.17±8.62	59.51±8.24	**99.23±1.14**	96.11±3.15	89.79±5.06	95.94±2.20	94.93±1.89
08	83.43±5.10	81.91±4.31	90.00±24.69	80.81±9.11	85.79±2.81	95.96±1.45	**96.25±2.66**
09	**99.65±0.28**	92.10±6.55	87.44±11.06	89.53±11.27	78.18±7.51	98.94±1.81	96.19±2.49
AA (%)	85.71±1.71	84.01±1.79	92.76±3.58	91.54±1.95	88.37±0.81	93.32±1.38	**95.90±0.57**
OA (%)	88.71±1.40	90.90±0.80	96.61±1.88	94.48±1.33	93.95±0.41	96.04±1.27	**97.85±0.24**
$k \times 100$	84.92±1.86	87.87±1.08	95.50±2.48	92.66±1.74	91.97±0.54	94.74±1.67	**97.14±0.31**

TABLE VII: CLASSIFICATION ACCURACY OF DIFFERENT ALGORITHMS FOR CHIKUSEI DATASET

No.	SVM-RBF [13]	3D-CNN [17]	SSRN [18]	HybridSN [19]	3D-GAN [39]	HyperGAMO [42]	DGSSC
01	99.10±0.67	96.52±3.67	98.87±2.26	98.42±3.14	98.70±1.72	45.14±25.21	**99.22±1.64**
02	96.35±1.90	97.67±2.34	**99.12±1.75**	98.33±1.68	98.36±2.41	37.01±20.66	98.65±1.74
03	6.51±13.54	12.39±6.03	0.00±0.00	44.86±22.99	82.54±16.44	**85.95±9.18**	49.51±25.80
04	98.68±3.65	98.95±0.74	**100.00±0.00**	99.34±1.64	96.63±3.30	94.36±8.28	99.61±0.75
05	98.48±0.67	99.36±0.38	99.95±0.16	99.88±0.19	**100.00±0.00**	91.68±7.38	99.95±0.16
06	92.19±3.33	89.24±4.98	95.09±2.88	86.30±11.76	86.01±8.04	**96.07±8.82**	94.42±2.80
07	99.36±0.30	**100.00±0.00**	**100.00±0.00**	99.92±0.08	99.94±0.07	91.84±8.12	**100.00±0.00**
08	98.80±0.42	98.95±0.76	99.36±1.87	98.50±2.02	98.95±0.53	96.19±6.76	**99.76±0.22**
09	99.59±0.25	99.32±0.63	99.99±0.04	99.02±1.19	99.96±0.13	97.10±2.06	**100.00±0.00**
10	99.62±0.30	92.22±3.23	99.82±0.52	99.50±0.73	93.36±4.81	92.27±3.26	**99.89±0.26**
11	99.12±0.76	98.00±1.89	**100.00±0.00**	99.06±1.72	68.11±9.36	96.29±2.89	99.81±0.29
12	93.34±3.44	97.42±1.91	**99.45±0.80**	85.66±10.15	68.19±9.33	96.11±3.43	98.27±1.95
13	93.24±4.36	89.56±3.19	97.28±2.10	91.77±6.23	93.55±4.47	**99.90±0.19**	94.09±2.91
14	99.01±0.26	98.21±0.99	99.92±0.11	99.49±0.46	**100.00±0.00**	97.00±3.40	99.34±0.38
15	**98.78±1.37**	70.07±16.34	94.10±9.86	89.91±20.63	61.73±9.98	98.61±2.04	94.36±12.54
16	85.41±6.37	57.86±25.70	9.04±27.13	54.95±44.97	52.36±12.92	**98.78±1.82**	83.55±8.58
17	98.10±1.51	81.16±6.13	**99.99±0.03**	99.22±1.57	63.76±6.71	96.91±4.02	99.78±0.27
18	90.47±7.40	72.79±11.52	88.83±13.76	75.47±11.74	60.28±11.71	**99.44±0.90**	96.36±7.49
19	8.61±9.48	6.11±3.31	0.00±0.00	77.71±27.32	5.97±5.34	**99.98±0.04**	79.38±13.36
AA (%)	87.09±1.36	81.89±2.57	83.20±1.59	89.33±4.63	80.44±1.75	90.03±3.74	**94.00±1.58**
OA (%)	98.06±0.27	97.31±0.53	98.76±0.25	97.96±0.82	94.32±0.47	97.92±0.85	**99.23±0.16**
$k \times 100$	97.76±0.31	96.89±0.62	98.57±0.29	97.64±0.95	93.44±0.54	97.60±0.98	**99.11±0.19**

TABLE VIII: CLASSIFICATION ACCURACY OF DIFFERENT ALGORITHMS FOR LOUKIA DATASET

No.	SVM-RBF [13]	3D-CNN [17]	SSRN [18]	HybridSN [19]	3D-GAN [39]	HyperGAMO [42]	DGSSC
01	1.79±0.78	42.34±7.80	60.84±10.99	51.35±12.74	51.17±12.89	**98.24±2.95**	66.17±5.09
02	60.63±22.02	71.87±15.94	58.59±41.03	66.41±21.83	78.44±18.76	28.67±8.90	**90.31±10.27**
03	41.46±4.29	78.33±5.05	81.32±8.00	75.48±11.68	44.95±7.80	44.58±8.50	**83.77±5.03**
04	6.32±4.63	10.00±6.43	12.53±12.97	28.67±10.40	18.29±6.15	**46.62±12.24**	26.80±6.78
05	14.27±3.75	74.51±6.62	87.44±6.41	88.48±3.34	82.56±6.03	68.59±6.03	**93.48±1.70**
06	35.14±11.54	29.57±7.99	31.65±7.57	30.57±9.06	20.66±4.47	**99.65±0.39**	30.90±5.48
07	0.40±0.57	45.62±8.51	63.79±11.48	54.86±5.73	44.61±8.74	**93.68±2.97**	69.33±4.60
08	3.92±1.37	69.86±4.75	**79.64±14.46**	68.20±15.14	78.83±7.75	64.42±12.37	78.82±5.52
09	**89.50±2.66**	83.65±2.94	87.87±4.49	79.13±6.26	73.96±5.63	87.84±3.29	87.15±1.80
10	40.34±2.23	81.27±2.16	79.69±4.60	75.90±5.89	79.47±2.86	72.47±3.38	**83.54±2.84**
11	36.07±8.11	62.87±4.34	80.97±6.91	72.24±4.98	72.84±5.73	**89.00±3.24**	81.02±6.01
12	76.65±3.82	91.88±2.91	80.15±19.64	90.11±5.43	89.37±3.58	**100.00±0.00**	89.52±3.01
13	0.00±0.00	**100.00±0.00**	90.00±30.00	**100.00±0.00**	**100.00±0.00**	83.23±5.13	99.87±0.39
14	0.00±0.00	99.70±0.38	97.95±3.00	**99.93±0.21**	99.91±0.29	86.72±1.94	99.16±0.57
AA (%)	29.03±2.25	67.25±2.34	70.89±3.88	70.10±3.40	66.79±2.73	75.98±2.41	**77.13±1.21**
OA (%)	41.78±0.59	79.12±1.39	82.29±3.67	78.80±2.47	76.53±0.64	84.00±1.38	**85.50±0.55**
$k \times 100$	23.93±0.72	74.92±1.68	78.78±4.51	74.71±2.91	72.08±0.78	80.83±1.69	**82.67±0.64**

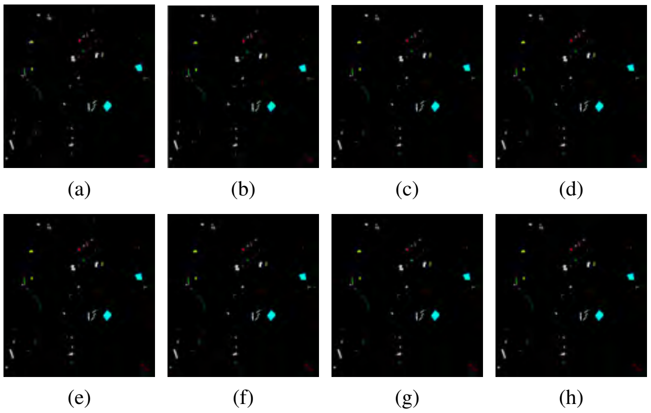


Fig. 13: Classification maps obtained via various algorithms for Chikusei dataset. (a) Groundtruth. (b) SVM-RBF. (c) 3D-CNN. (d) SSRN. (e) HybridSN. (f) 3D-GAN. (g) HyperGAMO. (h) Proposed DGSSC. Best zoomed-in view.

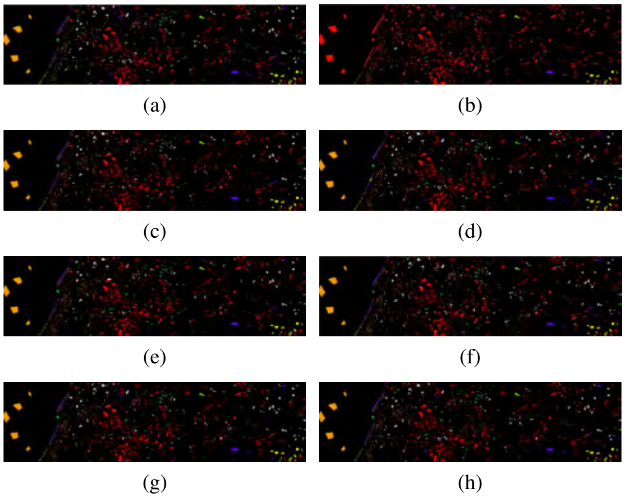


Fig. 14: Classification maps obtained via various algorithms for Loukia dataset. (a) Groundtruth. (b) SVM-RBF. (c) 3D-CNN. (d) SSRN. (e) HybridSN. (f) 3D-GAN. (g) HyperGAMO. (h) Proposed DGSSC. Best zoomed-in view.

14, the proposed DGSSC yields the fewest misclassifications on the generated maps compared to the other methods when referring to the groundtruth. Exclusively investigating the spectral features without the spatial information, the classification maps of the SVM-RBF have more salt-and-pepper noise interference than the other spectral-spatial classifier. In particular, the relatively small regions of the minority land-cover classes are accurately distinguished in the classification maps of the DGSSC. Furthermore, for the boundary region that is prone to misclassification in HSI, the maps of DGSSC are much clearer than those of other methods, especially on the UP dataset. In summary, these high-quality visualized results demonstrate that the proposed DGSSC can better serve realistic applications under imbalanced data conditions.

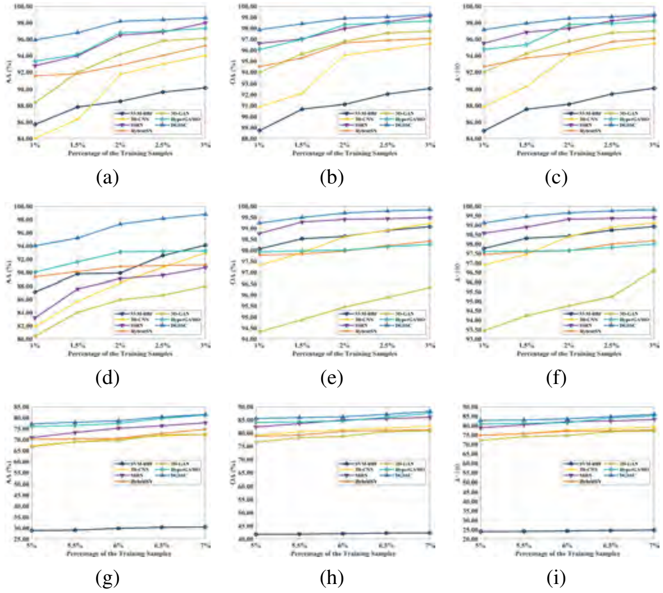


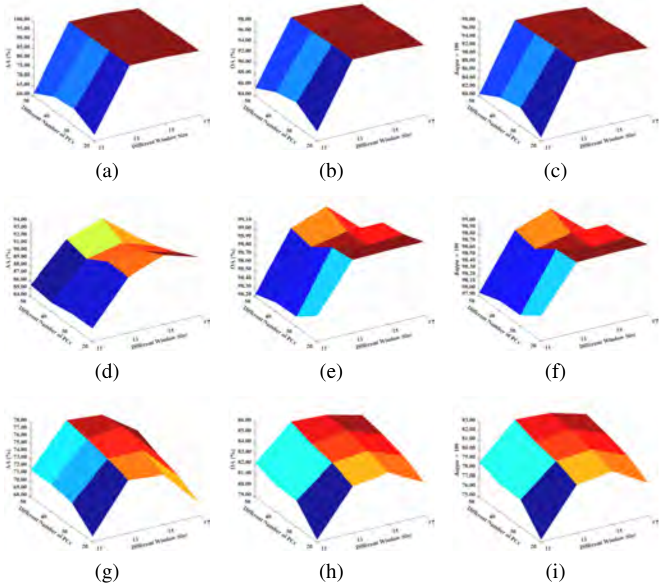
Fig. 15: Classification performance with different percentages of training samples. (a) UP-AA. (b) UP-OA. (c) UP-*k*. (d) Chikusei-AA. (e) Chikusei-OA. (f) Chikusei-*k*. (g) Loukia-AA. (h) Loukia-OA. (i) Loukia-*k*. Best zoomed-in view.

TABLE IX: COMPUTATIONAL COMPLEXITY OF DIFFERENT DEEP LEARNING-BASED METHODS ON THE EXPERIMENTAL DATASETS

	Dataset	3D-CNN	SSRN	HybridSN	3D-GAN	HyperGAMO	DGSSC
Training time	UP	131.94	205.01	28.91	194.59	3662.38	82.87
	Chikusei	267.55	421.46	52.30	281.10	6079.68	207.40
	Loukia	251.73	507.67	45.33	399.06	7048.61	471.50
Test time	UP	15.29	24.88	11.31	2.74	19.78	159.63
	Chikusei	32.23	47.76	20.81	0.80	42.12	620.08
	Loukia	5.67	8.82	3.10	5.04	9.33	141.19
FLOPs ($\times 10^6$)	UP	0.24	4.17	11.94	1066.15	582.24	24.96
	Chikusei	0.30	1.18	11.95	1066.15	1352.52	82.48
	Loukia	0.41	1.55	11.96	1066.15	969.45	28.28
Model size (MB)	UP	0.45	1.52	9.11	28.17	353.73	47.44
	Chikusei	0.57	1.80	9.12	28.18	807.31	157.14
	Loukia	0.78	2.36	9.12	28.17	585.43	53.78

E. Influence of Different Percentages of Training Samples

To further evaluate the generalization capability of our proposed framework, we investigate the classification performance of the DGSSC algorithm with different proportions of training samples for the three datasets. Because the algorithm is designed for class-imbalanced data, the training data are still randomly selected in a certain percentage from each class. Specifically, the training samples of the UP and Chikusei datasets are 1%-3% of the whole with an interval of 0.5%, while the training samples of the Loukia dataset are set to 5%-7%, and the interval is also 0.5%. Meanwhile, other experimental settings remain the same as those in Tables VI-VIII. Here, the experiments are repeated 10 times and the resulting average OA, AA, and k are delineated in Fig. 15. The classification accuracy grows with increasing amounts of labeled data for the three datasets. Specifically, the DGSSC always ranks first among all algorithms, which demonstrates the stability of the proposed method.


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 16: Classification performance with different number of PCs and window sizes. (a) UP-AA. (b) UP-OA. (c) UP-*k*. (d) Chikusei-AA. (e) Chikusei-OA. (f) Chikusei-*k*. (g) Loukia-AA. (h) Loukia-OA. (i) Loukia-*k*. Best zoomed-in view.

F. Impacts of the Number of PCs and Window Size

Next, we investigate the impacts of different numbers of PCs and window sizes. Specifically, keeping the number of training samples the same as Tables II to IV, the number of PCs and the window size are varied in the grid of $\{20, 30, 40, 50\}$ and $\{11, 13, 15, 17\}$, respectively, and the classification results are depicted in Fig. 16. Fig. 16 shows that the optimal number of PCs for the UP and Chikusei datasets is 20, and the best number for the Loukia dataset is 50. In addition, the most appropriate neighborhood size for the UP and Loukia datasets is 13, and the most suitable neighborhood size for the Chikusei dataset is 15. The middle size indicates that the spatial context information cannot be fully utilized when the window size is small, while a large neighborhood surrounding the center pixel may interfere with the classifier.

G. Analysis of Computational Complexity

In this section, the computational complexity of different DL-based algorithms is considered. All the networks are executed on the same server, which comprises an Intel Xeon E5-1620 V4 processor and a GeForce GTX 1080 Ti GPU. For a comprehensive view, the training time, test time, floating-point operations (FLOPs), and model size of each algorithm on the three datasets are calculated and summarized in Table IX. HybridSN consumes the least training time due to the lightweight 3D-2D combined architectures. Additionally, the 3D-GAN is relatively fast because only three principal components are utilized as in the original paper, but it takes a comparatively large number of FLOPs due to the large spatial size of the patch samples (i.e., 64×64). Although HyperGAMO obtains competitive classification performance, it consumes the longest training time, takes a relatively large number of FLOPs, and gains the largest model size. Remarkably, the training time, FLOPs, and model size of the proposed DGSSC network are moderate compared to those of other approaches. However, the test time is the longest due to the tenfold expansion of the test data during the inference phase, which is sacrificed to exchange for improved accuracy.

V. Conclusion

In this paper, we propose a DGSSC model for imbalanced HSIC. The framework consists of a two-stage 3D encoder, a 3D decoder, and a classifier that are trained in an end-to-end manner. By integrating the deep latent variable model into the networks, the minority classes are augmented so that their classification accuracies are improved. Meanwhile, the overall classification performance of the DGSSC is also superior to that of the state-of-the-art methods. Moreover, we introduce the patch distance-based reconstruction loss function to our model, which is significant when measuring the consistency of high-dimensional 3D samples. Furthermore, the proposed classifier is validated as robust and stable since the standard deviations of ten independent runs with different locations of the training samples are relatively small, and the classification performance is always best with different percentages of training samples. In the future, we will develop a more lightweight network to save inference time while maintaining high-quality classification results.

References

[1] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, Dec 2017.

[2] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2021.

[3] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, 2021.

[4] B. Xi, J. Li, Y. Li, R. Song, Y. Xiao, Q. Du, and J. Chanussot, "Semisupervised cross-scale graph prototypical network for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[5] C. Zhao, W. Zhu, and S. Feng, "Superpixel guided deformable convolution network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 3838–3851, 2022.

[6] B. Xi, J. Li, Y. Li, R. Song, Y. Xiao, Y. Shi, and Q. Du, "Multidirection networks with attentional spectral prior for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[7] B. Xi, J. Li, Y. Li, R. Song, D. Hong, and J. Chanussot, "Few-shot learning with class-covariance metric for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 5079–5092, 2022.

[8] J. Wang, J. Li, Y. Shi, J. Lai, and X. Tan, "Am3net: Adaptive mutual-learning-based multimodal data fusion network," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[9] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 159–173, June 2019.

[10] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[11] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[12] B. Tu, J. Wang, X. Kang, G. Zhang, X. Ou, and L. Guo, "KNN-based representation of superpixels for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4032–4047, 2018.

[13] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, Aug 2004.

[14] J. Fan, T. Chen, and S. Lu, "Superpixel guided deep-sparse-representation learning for hyperspectral image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3163–3173, 2018.

[15] H. Liu, Y. Jia, J. Hou, and Q. Zhang, "Global-local balanced low-rank approximation of hyperspectral images for classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2013–2024, 2022.

[16] L. Sun, C. Ma, Y. Chen, Y. Zheng, H. J. Shim, Z. Wu, and B. Jeon, "Low rank component induced spatial-spectral kernel method for hyperspectral image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3829–3842, 2020.

[17] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sensing*, vol. 9, no. 1, 2017.

[18] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 847–858, Feb 2018.

[19] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2020.

[20] J. Xie, N. He, L. Fang, and P. Ghamisi, "Multiscale densely-connected fusion networks for hyperspectral images classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 246–259, 2021.

[21] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1579–1597, March 2018.

[22] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, and S. Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, vol. 448, pp. 179–204, 2021.

[23] W. Feng, W. Huang, and W. Bao, "Imbalanced hyperspectral image classification with an adaptive ensemble method based on smote and rotation forest with differentiated sampling rates," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 12, pp. 1879–1883, 2019.

[24] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[25] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *arXiv preprint arXiv:2103.04922*, 2021.

[26] X. Wang, Y. Lyu, and L. Jing, "Deep generative model for robust imbalance classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 124–14 133.

[27] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "CVA2E: A conditional variational autoencoder with an adversarial training process for hyperspectral imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5676–5692, 2020.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[29] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, pp. 3483–3491, 2015.

[30] Z. Lv, G. Li, Z. Jin, J. A. Benediktsson, and G. M. Foody, "Iterative training sample expansion to increase and balance the accuracy of land classification from vhr imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 139–150, 2020.

[31] S. Jia, S. Jiang, Z. Lin, M. Xu, W. Sun, Q. Huang, J. Zhu, and X. Jia, "A semisupervised siamese network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[32] X. Zhang, Y. Wang, N. Zhang, D. Xu, H. Luo, B. Chen, and G. Ben, "Spectral–spatial fractal residual convolutional neural network with data balance augmentation for hyperspectral classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 473–10 487, 2021.

[33] J. Li, Q. Du, Y. Li, and W. Li, "Hyperspectral image classification with imbalanced data based on orthogonal complement subspace projection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 3838–3851, 2018.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[35] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, 2019.

[36] Z. He, H. Liu, Y. Wang, and J. Hu, "Generative adversarial networks-based semi-supervised learning for hyperspectral image classification," *Remote Sensing*, vol. 9, no. 10, p. 1042, 2017.

[37] Z. Chen, L. Tong, B. Qian, J. Yu, and C. Xiao, "Self-attention-based conditional variational auto-encoder generative adversarial networks for hyperspectral classification," *Remote Sensing*, vol. 13, no. 16, 2021.

[38] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 212–216, 2018.

[39] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5046–5063, 2018.

[40] Z. Zhong, J. Li, D. A. Clausi, and A. Wong, "Generative adversarial networks and conditional random fields for hyperspectral image classification," *IEEE transactions on cybernetics*, vol. 50, no. 7, pp. 3318–3329, 2019.

[41] J. Wang, F. Gao, J. Dong, and Q. Du, "Adaptive dropblock-enhanced generative adversarial networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5040–5053, 2020.

[42] S. K. Roy, J. M. Haut, M. E. Paoletti, S. R. Dubey, and A. Plaza, "Generative adversarial minority oversampling for spectral–spatial hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[43] L. Windrim, R. Ramakrishnan, A. Melkumyan, and R. J. Murphy, "A physics-based deep learning approach to shadow invariant representations of hyperspectral images," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 665–677, 2018.

[44] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial–spectral feature learning by 3d convolutional autoencoder for hyperspectral classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6808–6820, 2019.

[45] C. Shi and C.-M. Pun, "Multiscale superpixel-based hyperspectral image classification using recurrent neural networks with stacked autoencoders," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 487–501, 2020.

[46] J. Goldberger, S. Gordon, H. Greenspan *et al.*, "An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures." in *Proceedings Ninth IEEE International Conference on Computer Vision*, vol. 3, 2003, pp. 487–493.

[47] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, "PCA-based edge-preserving features for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7140–7151, Dec 2017.

[48] Q. Hao, S. Li, and X. Kang, "Multilabel sample augmentation-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4263–4278, 2020.

[49] X. Li, M. Ding, and A. Pižurica, "Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2615–2629, 2020.

[50] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[51] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[52] F. Wang, M. Jiang, C. Qian, S. Yang, C. C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," *2017*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450–6458, 2017.

[53] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, "Training generative neural networks via maximum mean discrepancy optimization," *arXiv preprint arXiv:1505.03906*, 2015.

[54] A. Mohan, G. Sapiro, and E. Bosch, "Spatially coherent nonlinear dimensionality reduction and segmentation of hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 2, pp. 206–210, 2007.

Bobo Xi (Member, IEEE) received the B.E. degree in information engineering and Ph.D. degree in information and communication engineering from Xidian University, Xi'an, China, in 2017 and 2022, respectively.

He is currently a Lecturer with the State Key Laboratory of Integrated Services Networks, School of Telecommunications, Xidian University. He has published over fifteen papers in refereed journals, including the IEEE Transactions on Image Processing, the IEEE Transactions on Neural Networks and Learning Systems, and the IEEE Transactions on Geoscience and Remote Sensing. His research interests include hyperspectral image processing, machine learning, and deep learning.

Jiaojiao Li (Member, IEEE) received the B.E. degree in computer science and technology, the M.S. degree in software engineering, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2009, 2012, and 2016, respectively.

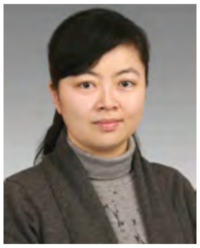
She was an exchange Ph.D. Student of Mississippi State University, supervised by Dr. Qian Du. She is currently an Associate Professor and doctoral supervisor with the State Key Laboratory of Integrated Service Networks, School of Telecommunications, Xidian University. Her research interests include hyperspectral remote sensing image analysis and processing, pattern recognition, and data compression.

Yan Diao received the B.E. degree in communication engineering in 2021 from Harbin Engineering University, Heilongjiang, China. She is currently studying for MA.Eng degree in the State Key Laboratory of Integrated Service Networks, School of Telecommunications, Xidian University, Xi'an, China.

Her research interests include hyperspectral image processing and deep learning.

Yunsong Li (Member, IEEE) received the M.S. degree in telecommunication and information systems and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 1999 and 2002, respectively.

In 1999, he joined the School of Telecommunications Engineering, Xidian University, where he is currently a Professor. He is also the Director of the State Key Laboratory of Integrated Service Networks, Image Coding and Processing Center. His research interests include image and video processing, hyperspectral image processing, and high-performance computing.



Zan Li (Senior Member, IEEE) received the B.S. degree in communications engineering and the M.S. and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 1998, 2001, and 2006, respectively.

She is a Professor with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University.

Prof. Li was awarded the National Science Fund for Distinguished Young Scholars. She is the fellow of the Institution of Engineering and Technology (IET), China Institute of Electronics (CIE), and China Institute of Communications (CIC). She serves as an Associate Editor for the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING and China Communications. Her research interests include topics on wireless communications and signal processing, such as covert communication, spectrum sensing, and cooperative communications.

Yan Huang received the B.S. degree in electrical engineering and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 2013 and 2018, respectively. He was a Visiting Ph.D. Student with the Electrical and Computer Engineering Department, University of Florida, Gainesville, FL, USA, from September 2016 to July 2017, and with the Electrical and Systems Engineering Department, Washington University in St. Louis, St. Louis, MO, USA, from July 2017 to August 2018.

He is currently an Associate Professor with the State Key Laboratory of Millimeter Waves, Southeast University, Nanjing, China. His research interests include machine learning, synthetic aperture radar, image processing, and remote sensing.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree in electrical engineering from the Université de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP, Grenoble, France, where he is currently a Professor of signal and image processing. He was a Visiting Scholar at Stanford University, Stanford, CA, USA, KTH Royal Institute of Technology, Stockholm, Sweden, and National University of Singapore, Singapore. Since 2013, he has been an Adjunct Professor of the University of Iceland, Reykjavk, Iceland, and the Chinese Academy of Sciences, Aerospace Information research Institute, Beijing, China. In 2015-2017, he was a Visiting Professor at the University of California, Los Angeles (UCLA), Los Angeles, CA, USA. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Prof. Chanussot holds the AXA Chair in remote sensing with the Chinese Academy of Sciences, Aerospace Information research Institute. He is the founding President of IEEE Geoscience and Remote Sensing French chapter (2007-2010), which received the 2010 IEEE GRSS Chapter Excellence Award. He has received multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia (2017-2019). He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS). He was the Chair (2009-2011) and Cochair of the GRS Data Fusion Technical Committee (2005-2008). He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006-2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011-2015). In 2014 he served as a Guest Editor for the IEEE Signal Processing Magazine. He is a member of the Institut Universitaire de France (2012-2017) and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters).