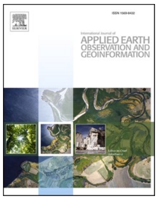
Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



A deep learning framework based on generative adversarial networks and vision transformer for complex wetland classification using limited training samples

Ali Jamali[a], Masoud Mahdianpari[b,c], Fariba Mohammadimanesh[c,*], Saeid Homayouni[d]

[a] *Civil Engineering Department, Faculty of Engineering, University of Karabük, Karabük, Turkey*
[b] *Department of Electrical and Computer Engineering, Memorial University of Newfoundland, St. John's, NL A1B3X5, Canada*
[c] *C-CORE, 1 Morrissey Rd, St. John's, NL A1B 3X5, Canada*
[d] *Centre Eau Terre Environnement, Institut National de la Recherche Scientifique, Quebec, QC G1K 9A9, Canada*

ARTICLE INFO

Keywords:
Generative adversarial network
Convolutional neural network
Wetland classification
New Brunswick
Vision Transformer (ViT)
Deep learning

ABSTRACT

Wetlands have long been recognized among the most critical ecosystems globally, yet their numbers quickly diminish due to human activities and climate change. Thus, large-scale wetland monitoring is essential to provide efficient spatial and temporal insights for resource management and conservation plans. However, the main challenge is the lack of enough reference data for accurate large-scale wetland mapping. As such, the main objective of this study was to investigate the efficient deep-learning models for generating high-resolution and temporally rich training datasets for wetland mapping. The Sentinel-1 and Sentinel-2 satellites from the European Copernicus program deliver radar and optical data at a high temporal and spatial resolution. These Earth observations provide a unique source of information for more precise wetland mapping from space. The second objective was to investigate the efficiency of vision transformers for complex landscape mapping. As such, we proposed a 3D Generative Adversarial Network (3D GAN) to best achieve these two objectives of synthesizing training data and a Vision Transformer model for large-scale wetland classification. The proposed approach was tested in three different study areas of Saint John, Sussex, and Fredericton, New Brunswick, Canada. The results showed the ability of the 3D GAN to stimulate and increase the number of training data and, as a result, increase the accuracy of wetland classification. The quantitative results also demonstrated the capability of jointly using data augmentation, 3D GAN, and Vision Transformer models with overall accuracy, average accuracy, and Kappa index of 75.61%, 73.4%, and 71.87%, respectively, using a disjoint data sampling strategy. Therefore, the proposed deep learning method opens a new window for large-scale remote sensing wetland classification.

1. Introduction

Wetlands are ecosystems found at the intersection of land and fresh- or salt-water environments and are defined by hydric soils that are regularly flooded (Cowardin et al., 1979). Emergent, shrub, and woodland vegetation dominate these habitats. They provide nutrients, flood prevention, erosion reduction, recreation, and aesthetics (Davidson, 2016). Wetlands also have significant economic value for commercial and recreational fisheries as they are home to various fish and wildlife (Ozesmi and Bauer, 2002). Nonetheless, the twentieth century's wetlands have been substantially degraded due to industrialization, climate change, and pollution. Due to the significance of these areas and the challenges they face, an accurate map of the distribution and structure of wetland vegetation is critical. These maps are an essential source of information for assessing the consequences and direct anthropogenic use on wetlands and protecting and managing them. Satellite-based Earth observations from optical and radar systems have become a valuable source of observation for wetland mapping (B. Hosseiny et al., 2022; Bansal et al., 2017). Remote sensing offers a cost-efficient and timely approach to wetland mapping. However, accurately mapping of wetlands may be a challenging task since these regions are often not portrayed by a particular vegetation cover but rather by the presence of water at the surface, underneath of the vegetation canopy, or within the soil, which makes them difficult to classify utilizing spectral or backscattering information with coarse spatial resolutions (Gallant, 2015). On the other hand, high spatial and temporal resolution Earth

[*] Corresponding author.
E-mail address: m.mahdianpari@mun.ca (F. Mohammadimanesh).

https://doi.org/10.1016/j.jag.2022.103095
Received 10 July 2022; Received in revised form 21 October 2022; Accepted 30 October 2022
Available online 8 November 2022
1569-8432/© 2022 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

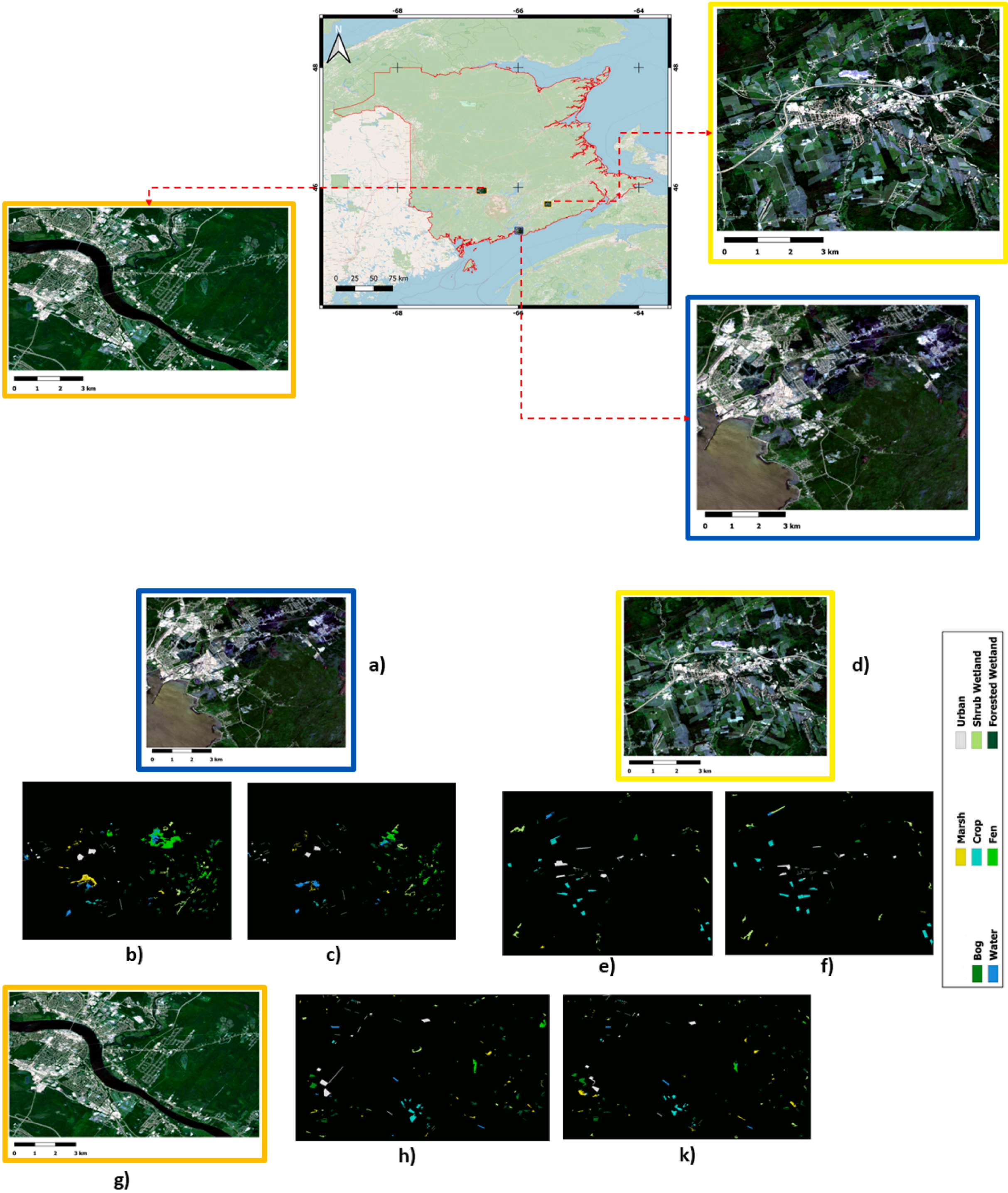


Fig. 1. The pilot sites and their reference data of a) Saint John, b) Saint John (DDSTr), and c) Saint John (DDSTs), d) Sussex, e) Sussex (DDSTr), and f) Sussex (DDSTs), g) Fredericton, h) Fredericton (DDSTr), and k) Fredericton (DDSTs), New Brunswick, Canada (DDSTr = training data for disjoint data sampling, DDSTs = test data for disjoint data sampling).

observations, such as Sentinel-1 and Sentinel-2 data, provide a perfect opportunity for detailed and accurate wetland classification from space (Slagter et al., 2020).

Moreover, different machine-learning techniques have been developed for accurate wetland identification in recent years (Jean Elizabeth Granger et al., 2021; Mahdavi et al., 2018). In particular, traditional supervised machine learning models, including decision trees (Jamali et al., 2021a), Support Vector Machines (Huang et al., 2014), and Random Forest (Berhane et al., 2018), are commonly utilized for wetland classification. However, deep learning approaches have become more relevant as computational capabilities have increased (DeLancey et al., 2020). Due to the general hand-crafted feature engineering (i.e., information extraction), the execution of the conventional classifiers significantly relies on the quality of the feature selection procedure. However, deep learning models learn through representation rather than empirical feature engineering. Internal feature representations are automatically learned, making these approaches extremely effective for image classification (Martins et al., 2020). While deep learning models

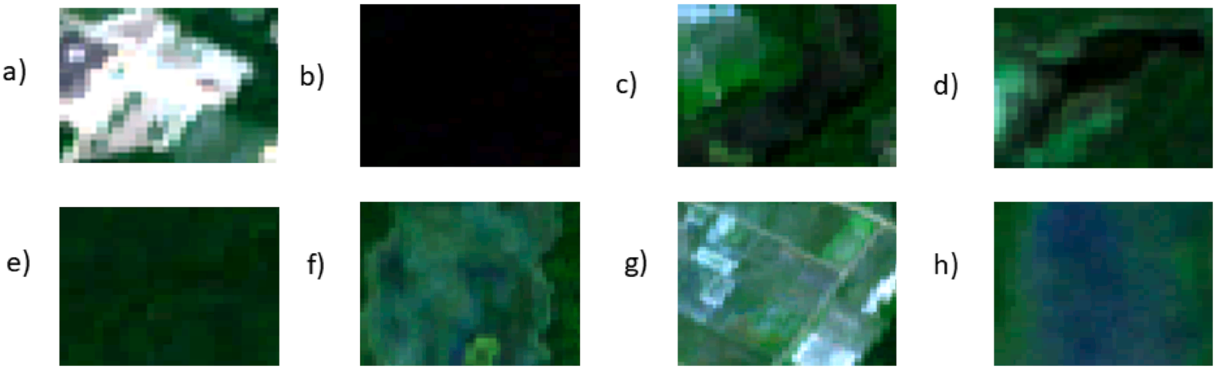


Fig. 2. Examples of wetland and non-wetland samples of a) urban, b) water, c) shrub wetland, d) marsh, e) forested wetland, f) fen, g) crop, and d) bog.

have been used in different applications, these models often rely on massive training datasets, extensive domain expertise, and computational resources (Martins et al., 2020).

CNNs are among the most well-known and utilized deep learning models in remote sensing image classification due to their great success in obtaining a higher classification accuracy than conventional methods such as RF (DeLancey et al., 2020). Generally, CNNs have driven computer vision modeling, primarily image classification. However, transformers are currently the most utilized architectures in natural language processing (NLP). The transformers (Vaswani et al., 2017) use attention mechanisms to explore long-range patterns in data. Its resounding success in the NLP field has encouraged scientists to explore its use in image classification, that have already shown promising results in a variety of research, including several remote sensing tasks (Bazi et al., 2021; D. Hong et al., 2021).

Moreover, the acquisition of training and validation samples is a hurdle in wetland mapping due to the high cost of field wetland data acquisitions. While a wide range of research has been undertaken in Canada on wetland classification (Amani et al., 2018; Asselen et al., 2013), there is no research on the possibility of producing synthetic reference data for large-scale mapping of wetlands. The transformers overperform the CNN networks thanks to their more generalizable characteristics. Additionally, transformers consider the connectivity between different characteristics of an image (i.e., attention component or positional encoding). Although, transformers necessitate more training information than that of CNNs to attain full image classification capabilities, which can be a considerable issue in some remote sensing tasks. Transfer learning (Jiao et al., 2021; Khan et al., 2021) and Generative Adversarial Networks (GANs) (Jamali et al., 2021b) are two techniques that can address the issue of scarcity of remote sensing reference data, specifically in wetland mapping. Therefore, this paper proposes a deep learning methodology based on GANs and the vision transformers for complex wetland classification to address the main limitations of deep learning models for large-scale wetland monitoring, including the scarcity of high-quality reference data. As such, the main contributions of this study can be defined as:

(1) To address the main limitation of deep learning models, i.e., limited training data, a 3D GAN was developed and proposed for the generation of high-resolution satellite reference data.

(2) To investigate the capability of vision transformers for the large-scale classification of complex wetlands.

2. Materials

2.1. Study area

As shown in Fig. 1, in this research, three pilot sites were selected in and around the towns of Saint John, Sussex, and Fredericton, located in New Brunswick, Canada. We identified five types of wetlands, including bog, marsh, fen, and forested and shrub wetlands. The reference data were collected from New Brunswick's wetland inventory, as seen in Fig. 2. It is worth mentioning that the reference data was collected as shape files and were extracted for the study areas in QGIS software. To minimize the over-classification of wetland regions, we used Google Earth's high-resolution imagery to visually identify three additional non-wetland classes, including water bodies, urban areas, and agricultural lands.

Table 1
Using a disjoint data sampling strategy, the number of references, test, and train pixels in Saint John, Fredericton, and Sussex pilot sites.

Class	Saint John (pixels)	Fredericton (pixels)	Sussex (pixels)	Training 3D GAN/Vision Transformer	Test Data
Bog	2145	3249	314	673/7845	5708
Marsh	1610	4202	272	841/9363	6084
Fen	6206	1245	–	1792/10918	7451
Forested Wetland	4980	3809	933	1251/13052	9722
Shrub Wetland	3346	3073	4021	1277/13290	10,440
Water	5055	1526	349	760/8628	6930
Urban	3511	4453	3389	1545/15940	11,353
Crop	624	3589	6918	1060/11335	11,131

For accuracy assessment, we used a disjoint data sampling strategy to significantly reduce the correlation between the training and test data, as seen in Table 1 (N. Audebert et al., 2019). The classification accuracy may considerably decrease; however, better results comparison from different algorithms can be obtained and discussed. We used a lower number of training data in the 3D GAN, as training the GAN model with a high number of training data is significantly costly in terms of time and hardware computation cost.

2.2. Remote sensing data

Various features from Sentinel-1/2 data, including spectral bands and normalized backscattering coefficients, were extracted in the Google Earth Engine (GEE) platform. The median Sentinel-1/2 image data were utilized. We should note that we utilized *maskS2clouds* and *median* algorithms to only include Sentinel-2 images with less than 10 % cloud coverage dated from June 1st to September 1st, 2020. The GEE platform provides preprocessed Sentinel-2 Level 2A images utilizing the *sen2cor* algorithm (Louis et al., 2016). Given that the spatial resolutions of the Sentinel-2 bands are different (10–60 m), utilized bands were converted into 10 m spatial resolutions. We utilized Sentile-2 bands of B2, B3, B4, B5, B6, B7, B8, B8A, B11, and B12 (Jamali et al., 2021a). Moreover, to improve the wetland classification accuracy, the normalized difference built-up index (NDBI), the modified normalized difference water index (MDNWI), the normalized difference vegetation index (NDVI), and the bare soil index (BSI) were employed as well. Sentinel-1 ground range

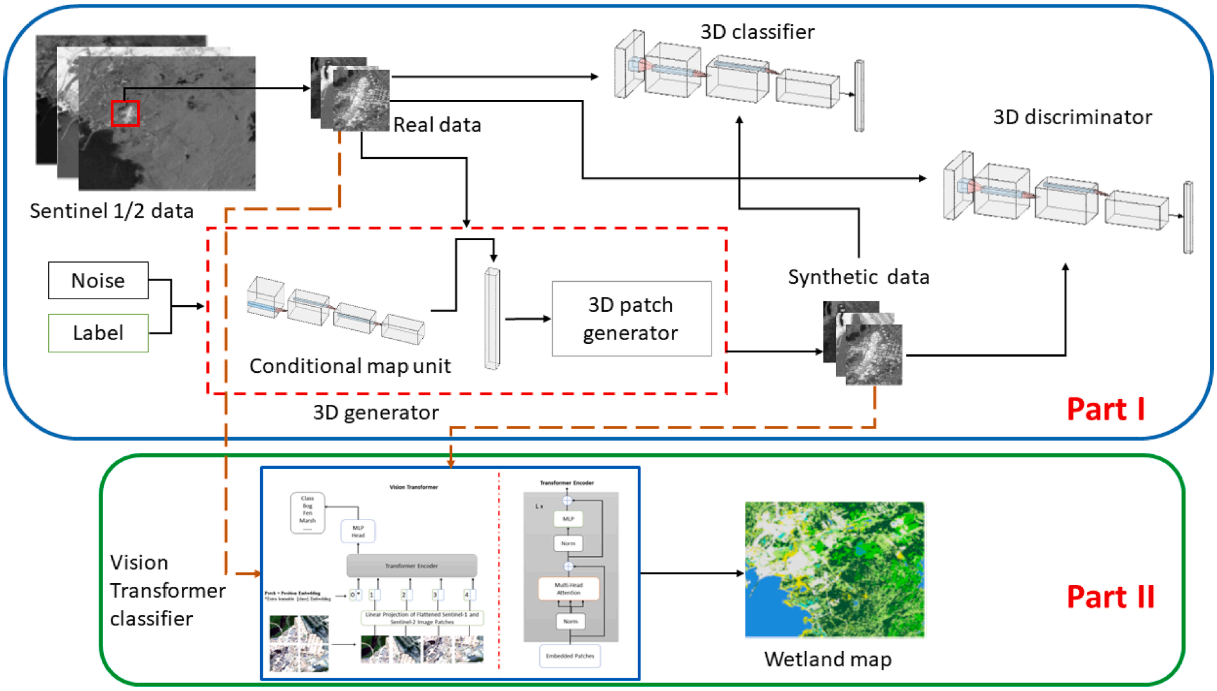


Fig. 3. The overall architecture of the developed wetland classifier.

detected (GRD) data, comprising σ^0_{VV}, σ^0_{VH}, σ^0_{HH}, and σ^0_{HV} that is log scaled at 10 m ground sampling distance are also available in the GEE platform. In the GEE platform, the Sentinel-1 toolbox pre-processes the delivered Sentinel-1 data, including terrain correction, thermal noise removal, and radiometric calibration.

3. Proposed deep learning method

A complete overview of the proposed methodology diagram is presented in Fig. 3. First, synthetic Sentinel-1 and Sentinel-2 data for classes with a few training samples are generated using a 3D GAN to increase wetland classification accuracy and address the common issue of limited reference wetland data (Part I). Then, as shown in Fig. 3, the real and synthetic data from three pilot sites are combined. Finally, synthetic and real data are fed to the Vision Transformer classifier for large-scale wetland mapping in New Brunswick (Part II). In Part I of the framework, the 3D GAN model only uses 10 % of the reference data as training and 90 % of the reference data as test data. The 3D generator produces high-quality Sentinel-1/2 for different wetland and non-wetland classes with a minority of training samples, while the 3D discriminator's task is to recognize the real samples from produced synthetic data. After reaching an acceptable overall classification accuracy by the 3D classifier of 3D GAN in Part I, the synthetic data and real data are utilized in the proposed Vision Transformer for large-scale complex wetland mapping in Part II, as described in the following sections. We employed 70 % of real and synthetic data and 30 % as test data in Part II by the Vision Transformer.

3.1. *3D generative adversarial network*

As seen in Equation (1), based on the min–max objective function of the 3D GAN, the generator tries to create more realistic Sentinel-1 and Sentinel-2 wetland data utilizing noises ($z\ p_z(z)$), while the discriminator tries to distinguish the real Sentinel-1 and Sentinel-2 data from simulated ones.

$$\min_G \max_D U(D,G) = E_{x\ p_{data}}[\ln(D(x))] + E_{z\ p_z}[\ln(1-D(G(z)))] \tag{1}$$

where the expected function and the objective function are formulated by E and $U(D,G)$. The distribution of real wetland samples is presented by p_{data}. We used a conditional map unit in the proposed 3D GAN to generate synthetic data from a random noise vector, similar to the GAMO (Subhra Mullick et al., 2019) and 3D-HyperGAMO (S. K. Roy et al., 2021) models, only for classes with fewer reference data. The advantage is solving the unbalanced data issue, which is common in wetland mapping. The 3D patch generator utilizes seven units, one for each class with a lower amount of training data. Consequently, the unit U_i generates γ^g_i samples as shown in Equation (2).

$$\gamma^g_i = \gamma_m - \gamma_i \tag{2}$$

where γ_m denotes the training data number in the class with the highest number of training samples, and γ_i denotes the number of training data in the classes with the lower training samples. Each of U_i uses data from 3D patches of Sentinel-1 and Sentinel-2 imagery with a dimension of $\gamma_i \times S \times S \times B$, as well as the output of the conditional map unit. The output of the conditional unit map is the intermediate feature expressed by I_f. Then it is transformed into a feature with the length of γ_i using a dense layer followed by a softmax layer. The feature vector is repeated by $n = S \times S \times B$ times for the generation of the class-specific random feature (I_m) with the dimension of $n \times \gamma$. The $\gamma_i \times S \times S \times B$ dimension of the 3D Sentinel-1 and Sentinel-2 data samples is then translated into a matrix (P_m) of dimension $n \times \gamma$. The class-specific feature matrix (F_m) for Sentinel-1 and Sentinel-2 image patches is then computed by Equation (3).

$$F_m = I_m.(F_m)^T \tag{3}$$

Then, a flattened vector (F_v) with a dimension of n is calculated by adding the column-wise sums of each row belonging to the F_m matrix. Finally, the synthetic Sentinel-1 and Sentinel-2 image samples are generated by converting F_v into the dimension of $S \times S \times B$ where S and B are equal to 8 and 16, respectively.

3.2. *Discriminator and classifier*

The architecture of the 3D classifier of the proposed GAN network is

Table 2
Configuration of the 3D classifier of the proposed GAN network.

Input dimension	8 × 8 × 16 × 1
Layer type	Size
3D Convolutional layer	1 × 1 × 7 × 16
3D Convolutional layer	3 × 3 × 5 × 32
3D Convolutional layer	5 × 5 × 7 × 32
Reshape layer	–
2D Convolutional layer	3 × 3 × 64
2D Convolutional layer	3 × 3 × 64
2D Global Average Pooling layer	–
Flatten layer	–
Dropout layer	0.5
Dense layer	20
Dense layer	10
Dense layer	8

Table 3
Configuration of the 3D Discriminator of the proposed GAN network.

Input dimension	8 × 8 × 16 × 1
Layer type	Size
3D Convolutional layer	1 × 1 × 7 × 32
3D Convolutional layer	3 × 3 × 3 × 64
Reshape layer	–
2D Convolutional layer	1 × 1 × 64
2D Convolutional layer	3 × 3 × 64
2D Global Average Pooling layer	–
Flatten layer	–
Dropout layer	0.5
Dense layer	20
Dense layer	10
Dense layer	1

demonstrated in Table 2, which includes three 3D convolutional layers and two 2D convolutional layers, followed by a 2D global average pooling layer and three dense layers with sizes of 20, 10, and 8.

As seen in Table 3, the proposed 3D discriminator network consists of two 3D convolutional layers and two 2D convolutional layers, followed by a 2D global average pooling layer and three dense layers with sizes of 20, 10, and 1.

The 3D GAN can be seen as a regularization technique that significantly reduces the overfitting issue. The 3D generator and 3D discriminator are trained in a competitive approach. The generator produces as realistic Sentinel-1 and Sentinel-2 synthetic data as possible, and the 3D discriminator attempts to classify the actual and synthetic samples. In this competitive rivalry, 3D networks (i.e., 3D generator and 3D discriminator) desire the best results. The 3D discriminator wants to achieve the best classification results, and the 3D generator tries to make simulated samples with the most matching distribution to the real Sentinel-1 and Sentinel-2 image patch data.

3.3. *Vision transformer*

To evaluate the efficiency of transformers for ecological mapping, we used the Vision Transformer (Alexander et al., 2021) for complex wetland classification. The Vision Transformer classifier receives real and synthetic Sentinel-1 and Sentinel-2 data generated by the proposed 3D GAN (see Fig. 4). The standard transformer takes a 1D series of token embeddings as input data. To use 3D Sentinel-1 and Sentinel-2 image patches, the satellites' image patches ($x \in R^{H\times W\times B}$) are reshaped into a sequence of flattened 2D patches ($x_p \in R^{N\times(P^2.B)}$), where B presents the band's number, $(H \times W)$ is the primary resolution of images, (P, P) denotes each image patch's resolution and resulted number of patches is presented by $N = HW/P^2$ which also presents as the transformer's effective input sequence size. With a trainable linear projection, the image patches are flattened and transferred into D dimensions as the transformer employs a constant latent vector with a size of D in all layers (see Eq. (4)). It should be noted the names used for the output of the projection are called patch embeddings.

$$z_0 = \left[x_{class}; x_p^1E; x_p^2E; \cdots; x_p^NE\right] + E_{pos}E \in R^{(P^2.B)\times D}, E_{pos} \in R^{(N+1)\times D} \tag{4}$$

A learnable embedding to the series of embedded Sentinel-1/2 image patches ($z_0^0 = x_{class}$) is appended, the state of which serves as the image representation (y) at the transformer encoder output (z_L^0) (see Eq. (5)).

$$y = LN(z_L^0) \tag{5}$$

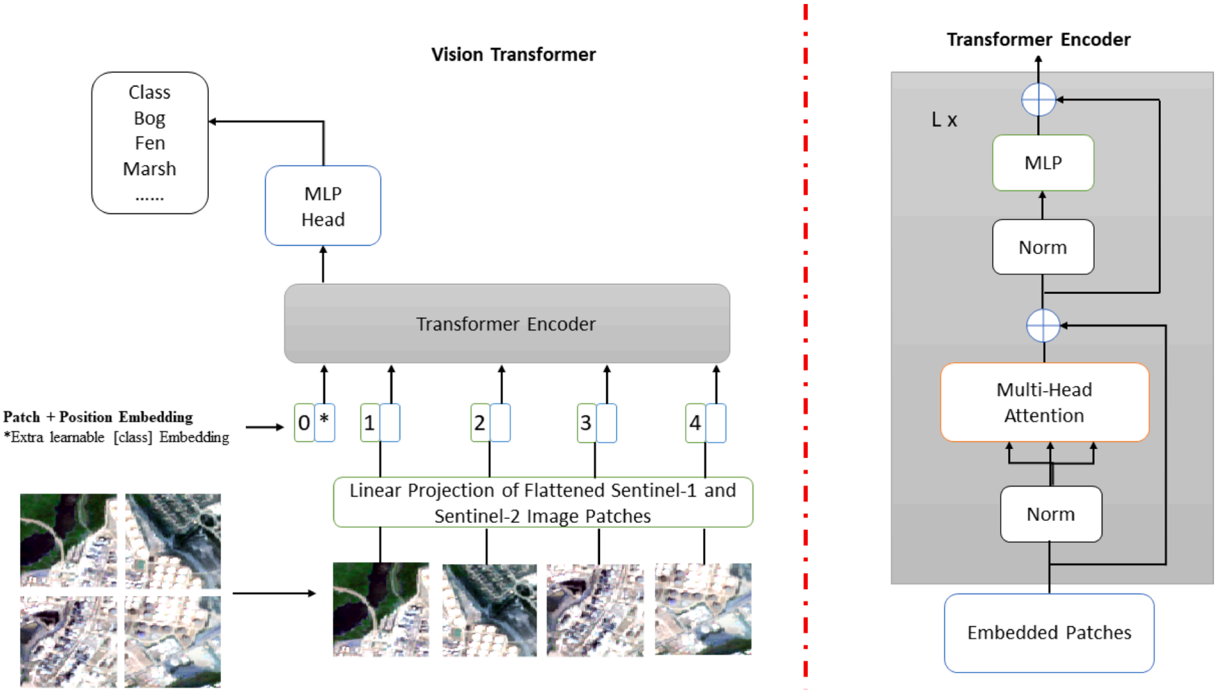
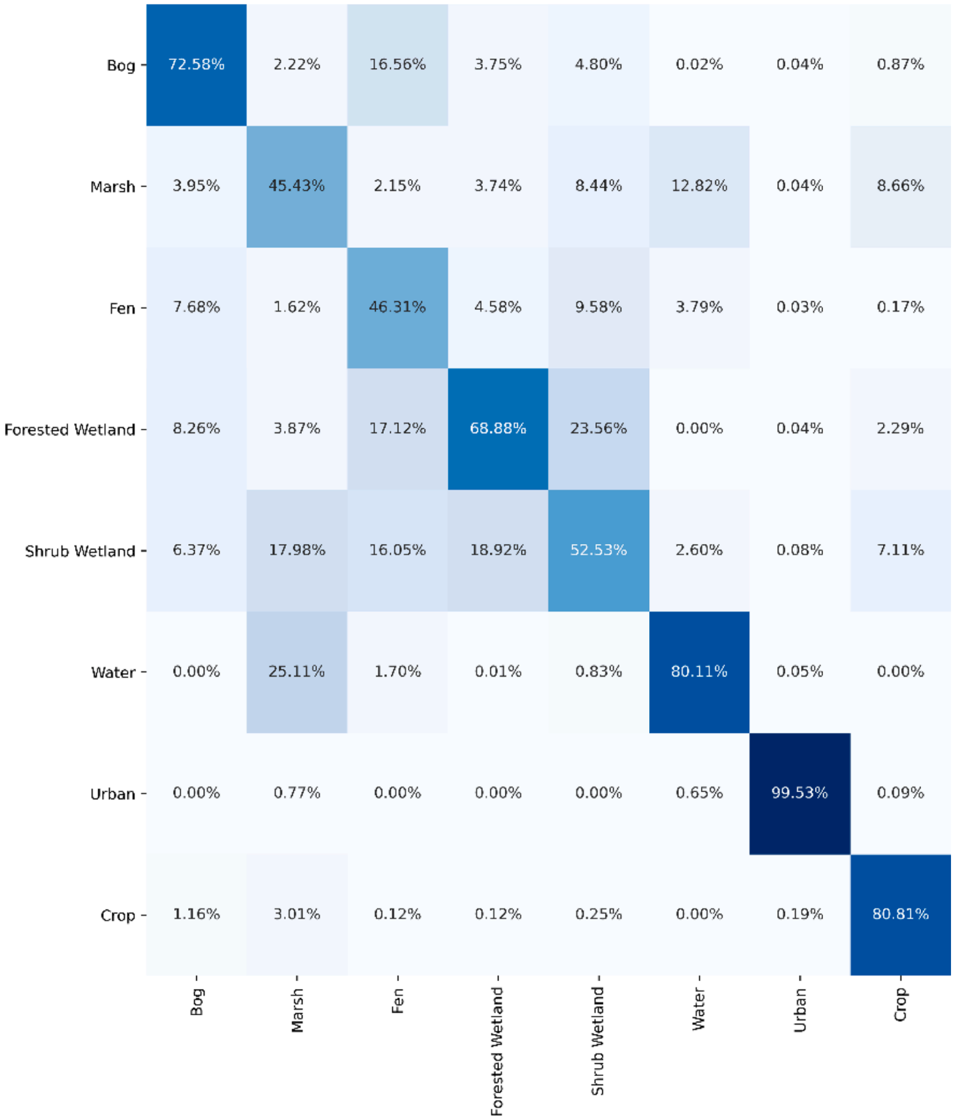


Fig. 4. Overview of the Vision Transformer (ViT).



	Bog	Marsh	Fen	Forested Wetland	Shrub Wetland	Water	Urban	Crop
Bog	72.58%	2.22%	16.56%	3.75%	4.80%	0.02%	0.04%	0.87%
Marsh	3.95%	45.43%	2.15%	3.74%	8.44%	12.82%	0.04%	8.66%
Fen	7.68%	1.62%	46.31%	4.58%	9.58%	3.79%	0.03%	0.17%
Forested Wetland	8.26%	3.87%	17.12%	68.88%	23.56%	0.00%	0.04%	2.29%
Shrub Wetland	6.37%	17.98%	16.05%	18.92%	52.53%	2.60%	0.08%	7.11%
Water	0.00%	25.11%	1.70%	0.01%	0.83%	80.11%	0.05%	0.00%
Urban	0.00%	0.77%	0.00%	0.00%	0.00%	0.65%	99.53%	0.09%
Crop	1.16%	3.01%	0.12%	0.12%	0.25%	0.00%	0.19%	80.81%

Fig. 5. The confusion matrix of the proposed 3D GAN model for generating synthetic Sentinel-1 and Sentinel-2 data using a) random sampling strategy b) disjoint data sampling strategy.

In fine-tuning and pre-training steps, the z_L^0 is connected to a classification head. In the pre-training step, a multi-layer perceptron (MLP) that has one hidden layer is employed as the classification head, while a single linear layer is utilized at the fine-tuning step. It is worth noting that to keep positional information, position embeddings are attached to the patch embeddings. Standard learnable 1D position embeddings in the Vision Transformer are employed as more complex 2D-aware position embeddings do not significantly improve performance (Alexander et al., 2021). The encoder receives the generated sequence of embedding vectors as its input data. The transformer encoder includes two layers of multiheaded self-attention (MSA) and MLP blocks formulated in Eq. (6) and Eq. (7).

Before each block, layer normalization (LN) is applied, and residual connections are added after each block, as seen in Fig. 4. Moreover, the Gaussian Error Linear Unit (GELU) non-linearity is utilized in the two layers of the MLP.

$$z_L^{'} = MSA(LN(z_{l-1})) + z_{l-1} l = 1, \cdots, L \quad (6)$$

$$z_L = MLP(LN(z_l^{'})) + z_l^{'} l = 1, \cdots, L \quad (7)$$

3.4. *Experimental settings*

In this study, 8 by 8 patch sizes were chosen experimentally from preprocessed Sentinel-1 and Sentinel-2 images. We also used Adam optimizer to train the 3D GAN and the Vision Transformer with a 0.0004 learning rate. The maximum training iteration for the 3D GAN was set to 30,000 epochs, and for the Vision Transformer classifier, we set it to 100 epochs. In addition, the noise dimension and training batch size in the 3D GAN were 100 and 32, respectively. The experiments were run with an i7-10750H Intel processor, 16 GB Random Access Memory (RAM), and an RTX 2070 NVIDIA GeForce Graphical Processing Unit (GPU) running on 64-bit Windows 10. It is worth mentioning that the loss function in the 3D GAN is mean squared error, while we utilized sparse categorical cross entropy as the loss function for the proposed Vision Transformer.

4. Results and discussion

To assess the classification capability of the proposed classifier, its result is compared with several state-of-the-art CNN models, including HybridSN (S. K. Roy et al., 2020), a Multi-model CNN classifier (Jamali and Mahdianpari, 2022a), and the Swin Transformer (Liu et al., 2021)

Table 4
Accuracy parameters of the proposed 3D GAN model: Kappa Index (KI), Average Accuracy (AA), Overall Accuracy (OA), and F1-score (Ag = data augmentation, ViT = Vision Transformer, ST = Swin Transformer). (**Using Disjoint Data Sampling**).

Class	ViT	ST (Jamali and Mahdianpari, 2022b)	HybridSN (S. K. Roy et al., 2020)	Multi-Model (Jamali and Mahdianpari, 2022a)	3D GAN (ours)	ViT + Ag (ours)	GAN + ViT (ours)	GAN + ViT + Ag (ours)
Bog	0.64	0.69	0.71	0.7	0.58	**0.77**	0.69	0.76
Marsh	0.43	0.34	0.44	0.45	0.45	0.45	0.43	**0.53**
Fen	0.61	0.56	**0.65**	0.61	0.57	0.62	**0.65**	**0.65**
Forested Wetland	0.61	0.57	0.63	0.61	0.56	0.62	0.59	**0.64**
Shrub Wetland	0.54	0.49	0.61	**0.62**	0.49	**0.62**	0.56	0.61
Water	0.79	0.72	0.78	**0.79**	0.77	0.76	**0.79**	0.77
Urban	0.99	0.99	**1**	0.99	0.99	**1**	0.99	**1**
Crop	0.87	0.48	0.94	**0.95**	0.88	0.9	0.92	0.94
KI (%)	66.46	55.24	70.89	70.06	64.66	69.97	68.91	**71.87**
OA (%)	70.92	60.95	74.79	74.07	69.37	73.95	73.06	**75.61**
AA (%)	68.41	60.69	71.93	71.25	66.52	71.84	70.3	**73.4**

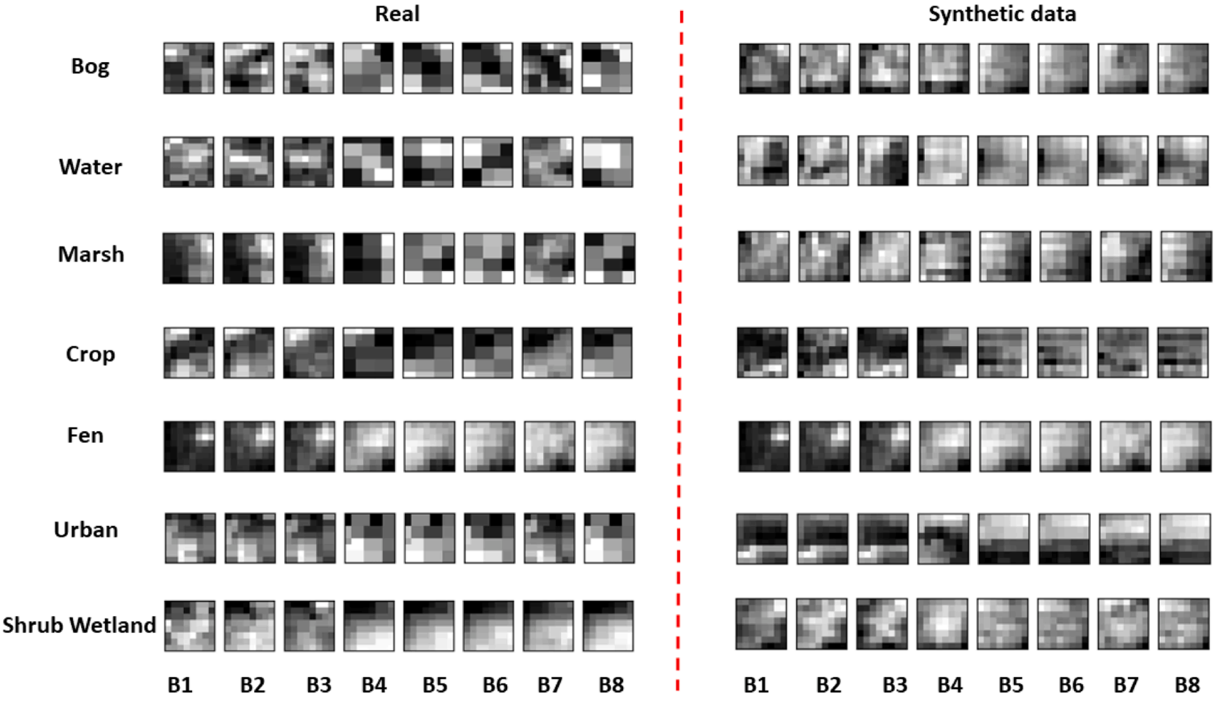


Fig. 6. Random examples of the real and synthetic data generated by the 3D GAN generator for different wetland and non-wetland classes.

that was developed and proposed for complex wetland mapping in New Brunswick (Jamali and Mahdianpari, 2022b). For accuracy assessment, the wetland classification results are compared in terms of the overall accuracy, average accuracy, kappa index, F-1 score, precision, and recall (Jamali et al., 2021a; Stehman and Foody, 2019).

4.1. 3D GAN

We evaluated the results of the developed 3D GAN model to generate synthetic Sentinel-1/2 data. We merged the reference data from the three study sites of Saint John, Sussex, and Fredericton to assess the efficiency of the proposed 3D GAN model for large-scale wetland reference data generation. The general results demonstrated a relatively high agreement between the reference and predicted non-wetland and wetland classes (see Fig. 5). Moreover, using the disjoint data sampling, the 3D GAN classifier obtained a kappa index, average accuracy, and overall accuracy of 64.66 %, 66.52 %, and 69.37 %, respectively, as seen in Table 4 and Fig. 5.

Table 4 and Fig. 5 reveal that the 3D GAN classifier demonstrated relatively lower performance on wetland classes with high similarity in vegetation structure. A high confusion was observed between the shrub and forest wetland classes. This confusion can be due to a significant similarity between these two wetlands regarding vegetation structure and patterns, which results in their comparable spectral reflectance, notably in optical satellite imaging of Sentinel-2. Examples of real and generated synthetic data by the 3D GAN model for different wetland and non-wetland classes are illustrated in Fig. 6. The presented examples are random samples created by the 3D generator of the developed 3D GAN. Using thousands of training samples of wetland and non-wetland classes, the 3D generator produced high-quality synthetic data for the classes with a minority of training samples. It should be noted that the shown samples are random and are not the exact synthetic data for the presented real data.

4.2. Accuracy assessment

We evaluated the wetland classification results utilizing different settings, including data augmentation and synthetic data with the Vision

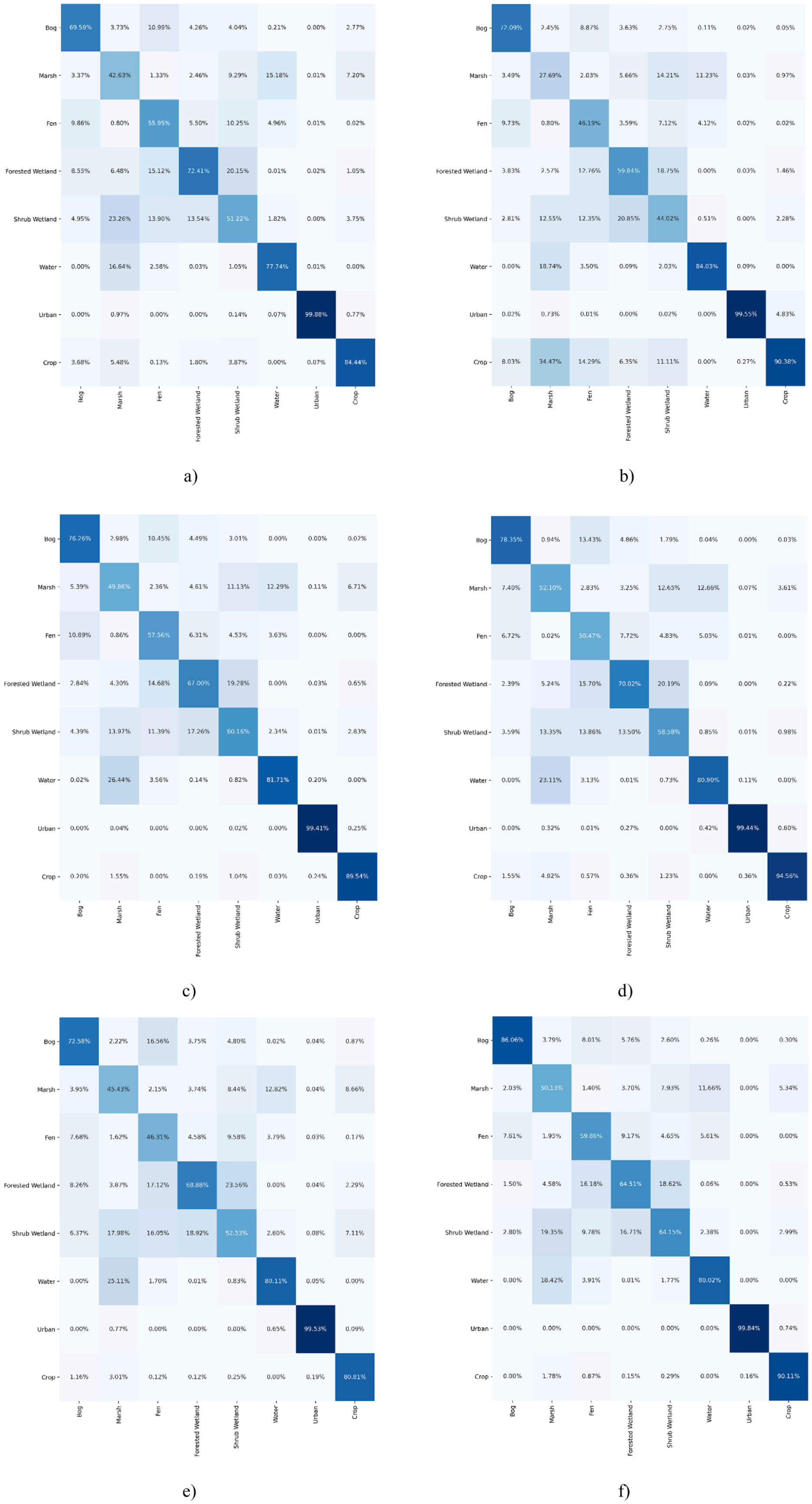


Fig. 7. The confusion matrix of a) Vision Transformer, b) Swin Transformer, c) HybridSN, d) Multi-Model, e) 3D GAN, and f) the proposed method (**Using Disjoint data sampling**).

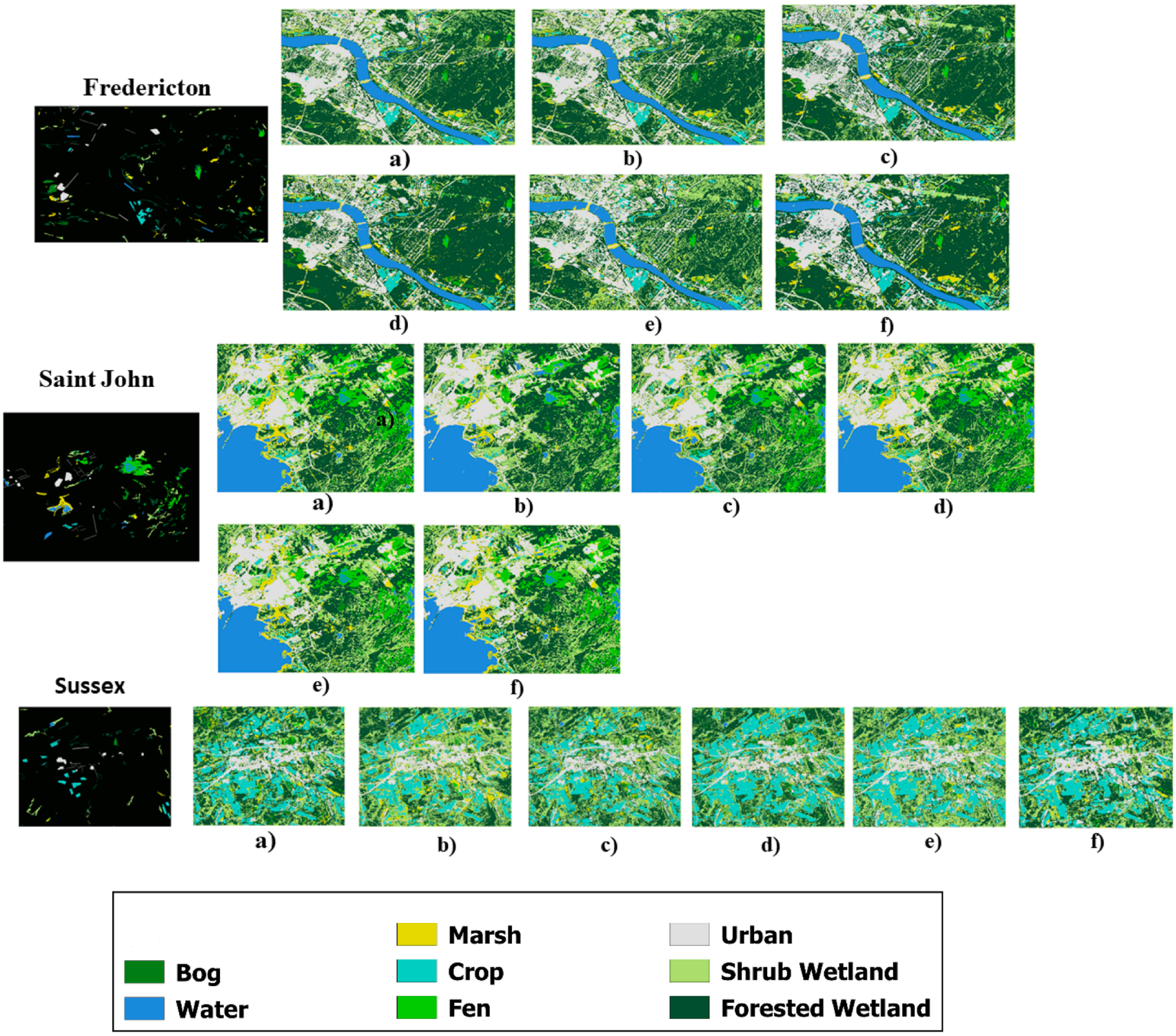


Fig. 8. Wetland maps of three pilot sites using of a) Vision Transformer, b) Swin Transformer, c) HybridSN, d) Multi-Model, e) 3D GAN, and f) the proposed method.

Transformer classifier. It is worth noting that data augmentation includes random crop and random rotation of image patches to increase the number of training data fed to the Vision Transformer classifier to increase the accuracy of wetland classification. It is worth mentioning that both Swin Transformer and Multi-Model algorithms were developed and proposed for complex wetland classification in New Brunswick. The Vision Transformer illustrated much better wetland classification accuracy in terms of average accuracy outperforming the state-of-the-art vision transformer of the Swin Transformer by approximately 8 % (see Table 4 and Fig. 7). The reason can be explained as the better generalization capability of the Vision Transformer over the Swin Transformer for large-scale wetland mapping. Moreover, based on the results of the disjoint data sampling, the proposed classifier utilizing real, synthetic, and augmented data achieved the highest classification accuracy compared to the other settings in the kappa index, average accuracy, and overall accuracy of 71.87 %,73.4 %, and 75.61 %, respectively, as seen in Table 4. The Vision Transformer obtained an average accuracy of 68.41 %; however, adding augmented data improved the average accuracy of the Vision Transformer by 3.43 %. Moreover, the inclusion of both synthetic and augmented data improved the classification accuracy of the Vision Transformer by approximately 5 % (see Table 4 and Fig. 7). In addition, the HybridSN (71.93 %) and the Multi-Model (71.25 %) algorithms demonstrated better classification results over the Swin Transformer (60.69 %) and the Vision Transformer (68.41 %) in terms of average accuracy, as seen in Table 4.

The uncertainty of the results originates from two main sources of the error and uncertainty created by the possible wrong-labeled synthetic wetland and non-wetland samples produced by the 3D generator of the 3D GAN, as well as the ability of the vision transformer to fit and train by the existent limited training data. The synthetic data may increase the uncertainty of the classification accuracy; however, as the results demonstrated, the inclusion of the synthetic data improved the classification accuracy of wetlands by the proposed vision transformer, as seen in Table 4. As such, the significance and positive effect of using synthetic data to improve the accuracy of wetland classification covered the increase of the result's uncertainty. It is worth noting that the most important objective of producing synthetic data is to improve the vision transformer's classification accuracy. As discussed in the introduction section, utilizing synthetic data may increase the existent complexity of the wetlands; however, this study was based on the idea that the synthetic data generated by cutting-edge algorithms (GANs) can improve the classification accuracy of wetlands even with their possible imperfections and increase of uncertainties.

4.3. Wetland classification

Fig. 8 presents the wetland and non-wetland regions obtained by the different CNN classifiers in the study sites. The proposed method

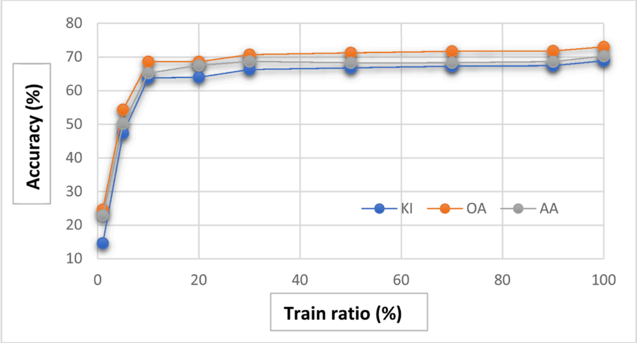


Fig. 9. Classification accuracies of the proposed classifier (GAN + ViT) based on different training ratios (%).

Table 5
Results of the proposed Vision Transformer classifier in terms of F-1 score (ws = window size, Ag = data augmentation, ViT = Vision Transformer).

Class	ViT + Ag ws = 4	ViT + Ag ws = 8	ViT + Ag ws = 12
Bog	0.65	**0.77**	**0.77**
Water	0.39	0.45	**0.52**
Marsh	0.59	0.62	**0.65**
Crop	0.66	**0.62**	0.61
Fen	0.58	0.62	**0.65**
Urban	0.85	**0.76**	0.7
Shrub Wetland	0.99	**1**	**1**
Forested Wetland	0.89	0.9	**0.96**
KI (%)	68.92	69.97	**72.11**
OA (%)	73.15	73.95	**75.81**
AA (%)	69.66	71.84	**73.52**

successfully differentiated various non-wetland and wetland regions based on our wetland experts' visual interpretation of wetland maps. Based on the visual interpretation of wetland experts, the proposed algorithm for large-scale wetland mapping demonstrated better visual maps compared to other classifiers, including the HybridSN, Swin Transformer, and Vision Transformer. For example, HybridSN, Vision Transformer, and Multi-Model classifier showed over-classification of shrub wetlands in the Fredericton study site. Moreover, in Sussex, there was an over-classification of shrub wetlands by the HybridSN and Swin Transformer algorithms. Additionally, the Multi-Model and HybridSN classifiers resulted in an over-classification of marsh wetlands in Saint John.

4.4. *The effect of training ratio and window size*

Fig. 9 demonstrates the classification accuracies of the proposed Vision Transformer for different training sample rates and image patch sizes. Results showed that the training sample ratio substantially affected the wetland classification accuracies (1 % (i.e., 697 samples) to 5 % (i.e., 3488 samples)). In other words, the average accuracy is significantly improved by approximately 27 %. Compared to the 5 % to 10 % (i.e., 6977 samples) training sample ratio, average accuracy increased by around 15 %. After that, the classification accuracy improvement became steady. Based on the results, from 10 % to 100 % (i.e., 69,775 samples) training sample rates, the average accuracy increased by around 5 %, as seen in Fig. 9. Results demonstrated that there is a limit where increasing the number of training data would substantially improve the classification accuracy of wetlands. Although increasing the training data may increase the classification accuracy of wetlands by a few percent, the computation cost in terms of time will significantly increase from 50 % to 100 % training ratio. As such, there should be a trade-off between the number of training data to reach acceptable classification accuracy and the required computation cost that is dependent on the available hardware in a remote sensing project.

Additionally, when the image patch size was set to 12, the proposed classifier obtained the highest average accuracy 73.52 %), as seen in Table 5, while the least average accuracy was obtained by setting the image patch size to 4 (69.66 %). When the size is too small, it is evident that the effect of spatial information is minimal. Moreover, when the image patch size is too large, the classification becomes slow, and the processing complexity rises significantly.

5. Conclusion

Due to the intrinsic complexity of wetlands, characterizing these valuable and threatened ecosystems using Earth observations is challenging. Moreover, the scarcity of reference wetland data for precise large-scale wetland mapping is one of the most encountered challenges. Consequently, this study investigated the use and efficiency of cutting-edge CNNs (i.e., GANs) and transformers (i.e., the Vision Transformer) for wetland generation and classification. To improve the wetland classification accuracy, we used a 3D GAN to generate synthetic Sentinel-1 and Sentinel-2 data with classification accuracies in the kappa index, average accuracy, and overall accuracy with values of 64.66 %, 66.52 %, and 69.37 %, respectively, by the 3D GAN classifier. Moreover, using both synthetic and augmented data, the developed classifier obtained wetland classification accuracies in the kappa index, average accuracy, and overall accuracy of 71.87 %, 73.4 %, and 75.61 %, respectively. The Vision Transformer obtained an average accuracy of 68.41 %; however, adding augmented data improved the average accuracy of the Vision Transformer by 3.43 %. Moreover, the inclusion of both synthetic and augmented data improved the classification accuracy of the Vision Transformer by approximately 5 %. One of the main significances of the developed model is the capability of the developed 3D GAN to produce high-quality synthetic wetland data. In this way, the need for costly and time-consuming field data acquisition by wetland experts will be significantly reduced.

Availability of data and materials

Data are available upon reasonable request.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

References

Alexander, K., Alexey, D., Dirk, W., Georg, H., Jakob, U., Lucas, B., Matthias, M., Mostafa, D., Neil, H., Sylvain, G., Thomas, U., Xiaohua, Z., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Presented at the International Conference on Learning Representations (ICLR).

Amani, M., Salehi, B., Mahdavi, S., Brisco, B., 2018. Spectral analysis of wetlands using multi-source optical satellite imagery. ISPRS J. Photogramm. Remote Sens. 144, 119–136.

Audebert, N., Le Saux, B., Lefevre, S., 2019. Deep learning for classification of hyperspectral data: a comparative review. IEEE Geosci. Remote Sens. Mag. 7, 159–173. https://doi.org/10.1109/MGRS.2019.2912563.

Bansal, S., Katyal, D., Garg, J.K., 2017. A novel strategy for wetland area extraction using multispectral MODIS data. Remote Sens. Environ. 200, 183–205.

Bazi, Y., Bashmal, L., Rahhal, M.M.A., Dayil, R.A., Ajlan, N.A., 2021. Vision transformers for remote sensing image classification. Remote Sens. 13 https://doi.org/10.3390/rs13030516.

Berhane, T.M., Lane, C.R., Wu, Q., Autrey, B.C., Anenkhonov, O.A., Chepinoga, V.V., Liu, H., 2018. Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory. Remote Sens. 10.

Cowardin, L., Carter, V., Golet, F., LaRoe, E., 1979. Classification of wetlands and deepwater habitats of the United States. Fish and Wildlife Service, Washington.

Davidson, N.C., 2016. The Ramsar Convention on Wetlands, in: The Wetland Book I: Structure and Function, Management and Methods. Springer Publishers, Dordrecht.

DeLancey, E.R., Simms, J.F., Mahdianpari, M., Brisco, B., Mahoney, C., Kariyeva, J., 2020. Comparing deep learning and shallow learning for large-scale wetland classification in Alberta, Canada. Remote Sens. 12 https://doi.org/10.3390/rs12010002.

Gallant, A.L., 2015. The challenges of remote monitoring of wetlands. Remote Sens. 7 https://doi.org/10.3390/rs70810938.

Granger, J.E., Mahdianpari, M., Puestow, T., Warren, S., Mohammadimanesh, F., Salehi, B., Brisco, B., 2021. Object-based random forest wetland mapping in Conne River, Newfoundland, Canada. J. Appl. Remote Sens. 15, 1–10. https://doi.org/10.1117/1.JRS.15.038506.

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J., 2021. SpectralFormer: rethinking hyperspectral image classification with transformers. IEEE Trans. Geosci. Remote Sens. 1–1 https://doi.org/10.1109/TGRS.2021.3130716.

Hosseiny, B., Mahdianpari, M., Brisco, B., Mohammadimanesh, F., Salehi, B., 2022. WetNet: A spatial-temporal ensemble deep learning model for wetland classification using sentinel-1 and sentinel-2. IEEE Trans. Geosci. Remote Sens. 60, 1–14. https://doi.org/10.1109/TGRS.2021.3113856.

Huang, C., Peng, Y., Lang, M., Yeo, I.-Y., McCarty, G., 2014. Wetland inundation mapping and change monitoring using Landsat and airborne LiDAR data. Remote Sens. Environ. 141, 231–242. https://doi.org/10.1016/j.rse.2013.10.020.

Jamali, A., Mahdianpari, M., Brisco, B., Granger, J., Mohammadimanesh, F., Salehi, B., 2021a. Deep Forest classifier for wetland mapping using the combination of Sentinel-1 and Sentinel-2 data. GIScience & Remote Sens. 1–18 https://doi.org/10.1080/15481603.2021.1965399.

Jamali, A., Mahdianpari, M., Mohammadimanesh, F., Brisco, B., Salehi, B., 2021b. A synergic use of sentinel-1 and sentinel-2 imagery for complex wetland classification using generative adversarial network (GAN) scheme. Water 13. https://doi.org/10.3390/w13243601.

Jamali, A., Mahdianpari, M., 2022a. Swin transformer and deep convolutional neural networks for coastal wetland classification using sentinel-1, sentinel-2, and LiDAR data. Remote Sens. 14 https://doi.org/10.3390/rs14020359.

Jamali, A., Mahdianpari, M., 2022b. Swin transformer for complex coastal wetland classification using the integration of sentinel-1 and sentinel-2 imagery. Water 14. https://doi.org/10.3390/w14020178.

Jiao, W., Wang, Q., Cheng, Y., Zhang, Y., 2021. End-to-end prediction of weld penetration: A deep learning and transfer learning based method. J. Manuf. Processes 63, 191–197. https://doi.org/10.1016/j.jmapro.2020.01.044.

Khan, M.A., Akram, T., Zhang, Y.-D., Sharif, M., 2021. Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework. Pattern Recogn. Lett. 143, 58–66. https://doi.org/10.1016/j.patrec.2020.12.015.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau, E., Gascon, F., 2016. Sentinel-2 Sen2Cor: L2A processor for users. Presented at the Proceedings Living Planet Symposium 2016, Spacebooks Online, pp. 1–8.

Mahdavi, S., Salehi, B., Granger, J., Amani, M., Brisco, B., Huang, W., 2018. Remote sensing for wetland classification: A comprehensive review. GIScience & Remote Sens. 55, 623–658.

Martins, V.S., Kaleita, A.L., Gelder, B.K., Nagel, G.W., Maciel, D.A., 2020. Deep neural network for complex open-water wetland mapping using high-resolution WorldView-3 and airborne LiDAR data. Int. J. Appl. Earth Obs. Geoinf. 93, 102215 https://doi.org/10.1016/j.jag.2020.102215.

Ozesmi, S.L., Bauer, M.E., 2002. Satellite remote sensing of wetlands. Wetlands Ecol. Manage. 10, 381–402. https://doi.org/10.1023/A:1020908432489.

Roy, S.K., Krishna, G., Dubey, S.R., Chaudhuri, B.B., 2020. HybridSN: exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 17, 277–281. https://doi.org/10.1109/LGRS.2019.2918719.

Roy, S.K., Haut, J.M., Paoletti, M.E., Dubey, S.R., Plaza, A., 2021. Generative adversarial minority oversampling for spectral-spatial hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 1–15 https://doi.org/10.1109/TGRS.2021.3052048.

Slagter, B., Tsendbazar, N.E., Vollrath, A., Reiche, J., 2020. Mapping wetland characteristics using temporally dense Sentinel-1 and Sentinel-2 data: A case study in the St. Lucia wetlands, South Africa. Int. J. Appl. Earth Obs. Geoinf. 86.

Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sens. Environ. 231, 111199 https://doi.org/10.1016/j.rse.2019.05.018.

Subhra Mullick, S., Datta, S., Das, S., 2019. Generative Adversarial Minority Oversampling, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1695–1704.

van Asselen, S., Verburg, P.H., Vermaat, J.E., Janse, J.H., 2013. Drivers of wetland conversion: a global meta-analysis. PLoS ONE 8, e81292.