Journal of *Marine Science and Engineering*

MDPI

Article

Low-Resource Generation Method for Few-Shot Dolphin Whistle Signal Based on Generative Adversarial Network

Huiyuan Wang [1,2], Xiaojun Wu [2], Zirui Wang [1], Yukun Hao [1,2], Chengpeng Hao [3], Xinyi He [4] and Qiao Hu [1,5,*]

1 School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China
2 School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
3 Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China
4 Naval Academy of Armament, Beijing 100161, China
5 Shaanxi Key Laboratory of Intelligent Robots, Xi'an Jiaotong University, Xi'an 710049, China
* Correspondence: hqxjtu@xjtu.edu.cn

Abstract: Dolphin signals are effective carriers for underwater covert detection and communication. However, the environmental and cost constraints terribly limit the amount of data available in dolphin signal datasets are often limited. Meanwhile, due to the low computational power and resource sensitivity of Unmanned Underwater Vehicles (UUVs), current methods for real-time generation of dolphin signals with favorable results are still subject to several challenges. To this end, a Masked AutoEncoder Generative Adversarial Network (MAE-GAN) model is hereby proposed. First, considering the few-shot condition, the dataset is extended by using data augmentation techniques. Then, to meet the low arithmetic constraint, a denoising autoencoder with a mask is used to obtain latent codes through self-supervised learning. These latent codes are then utilized in Conditional Wasserstein Generative Adversarial Network-Gradient Penalty (CWGAN-GP) to generate a whistle signal model for the target dataset, fully demonstrating the effectiveness of the proposed method for enhancing dolphin signal generation in data-limited scenarios. The whistle signals generated by the MAE-GAN and baseline models are compared with actual dolphin signals, and the findings indicate that the proposed approach achieves a discriminative score of 0.074, which is 28.8% higher than that of the current state-of-the-art techniques. Furthermore, it requires only 30.2% of the computational resources of the baseline model. Overall, this paper presents a novel approach to generating high-quality dolphin signals in data-limited situations, which can also be deployed on low-resource devices. The proposed MAE-GAN methods provide a promising solution to address the challenges of limited data and computational power in generating dolphin signals.

Keywords: bionic signal generation; few-shot learning; generative adversarial network; data augmentation

Citation: Wang, H.; Wu, X.; Wang, Z.; Hao, Y.; Hao, C.; He, X.; Hu, Q. Low-Resource Generation Method for Few-Shot Dolphin Whistle Signal Based on Generative Adversarial Network. *J. Mar. Sci. Eng.* **2023**, *11*, 1086. https://doi.org/10.3390/jmse11051086

Academic Editors: Sergey Pereselkov, Matthias Ehrhardt, Pavel Petrov and Yassine Amirat

Received: 15 March 2023
Revised: 27 April 2023
Accepted: 17 May 2023
Published: 22 May 2023

1. Introduction

Unmanned Underwater Vehicles (UUVs) are widely used for various underwater tasks. Moreover, some tasks require covert operations, such as detection and communication, making it challenging for UUVs to remain undetected. Dolphin signals, due to their frequency and duration, are well-suited for these covert tasks [1,2] and are gaining attention in underwater sonar research. While the academic community has made progress in detecting [3], classifying [4], and extracting dolphin whistle signals [5] and click signals [6], practical scientific research faces difficulties in acquiring sufficient underwater samples. This is attributed to factors such as cost, environmental constraints, and time limitations, which can result in insufficient sample size or an under-representation of the actual data. Additionally, covert missions amplify the risk of identification due to the limited number of samples available, making it crucial to develop methods that can effectively outfit UUVs under limited data conditions. Such methods can enable UUVs to execute covert tasks without risking detection, which has significant implications for underwater sonar research.

There are three main methods of covert underwater communication and detection: low probability of detection [7] (LPD), low probability of recognition [8] (LPR), and low probability of interception [9] (LPI). Among these methods, LPD provides the highest level of secrecy by ensuring that the signal is not detectable, but it conflicts with the key indicators of communication and detection. Increasing transmission power to improve distance, reliability, and effectiveness will inevitably expose underwater platforms. LPR, on the other hand, enables communication or detection signal detection without identification. This technology utilizes bionic camouflage to communicate and detect through the ocean's inherent biological calls, making it challenging for sonar operators to distinguish biological noise and detection signals. In addition, LPI facilitates signal detection, which cannot be accurately deciphered, and is primarily used in the field of underwater acoustic communication. However, this method also poses the problem of exposing communication platforms, as well as the lack of marine environment friendliness. In other words, the LPR signal is a more extensively applicable method for covert underwater communication and detection, capable of striking a balance between secrecy and practicality.

Realizing the LPR signal relies on biological noise, but it is difficult to collect the signal, thereby leading to small datasets. Moreover, the continued use of these datasets expands potential detection and communication exposure. To extract and reconstruct dolphin signals and obtain new biological signals, modifications can be made during reconstruction. However, these methods fail to fundamentally solve the small data problem. A simple solution is to directly expand the number of data samples. Data augmentation is a popular solution to expanding the number of data samples, commonly used in scenarios featuring an insufficient amount of data or significant model parameters, particularly in the image field. However, data augmentation can also be applied to sequencing data generation, which can be considered a simplified image field data expansion. Ekin D. Cubuk et al. [10] utilized a search algorithm to seek an optimal image augmentation strategy for a specific dataset, but this approach was computationally intensive. Regarding RandAugment [11], the author proposed a random augmentation method, which selected all sub-strategies with equal probability, and thus made the method more suitable to be migrated to other datasets. Experiments showed that RandAugment had a good effect, even in the case of training large models. Cutout and Mixup [12,13] are other data augmentation techniques used in image classification. The former could simulate the classification scene when the subject was partially occluded, thereby promoting the model to make full use of more image content to classify and prevent overfitting, whereas the latter was a simple and easy-to-implement image aliasing augmentation scheme that achieved favorable good results in both image classification and target detection. Cutmix [14], a similar approach to Mixup, randomly cropped an ROI (Range Of Interest) from an image and covered the corresponding area in the current image. However, not taking into account the actual underwater conditions and inter-population variation, all these above-described data augmentation methods are not applicable to small-sample dolphin whistle signal datasets.

Data expansion methods may increase the amount of data and partially integrate data distribution. However, it is challenging to demonstrate the validity and reliability of the expanded data, and data expansion methods that incorporate constraints often produce a limited amount of expanded data. To address this issue, a method known as Generative Adversarial Network (GAN) [15] has been introduced in the field of deep learning in recent years, which offers a solution to the challenge of small data and enhances the accuracy and reliability of LPR signals. In the field of few-shot generation, Yijun Li et al. proposed a technique called few-shot image generation [16], which aimed to generate more data for a given domain even when there were only a few training examples available. Kai Li et al. [17] proposed an Adversarial Feature Hallucination Network (AFHN), a novel framework for Few-Shot Learning (FSL) based on conditional Wasserstein Generative Adversarial Networks (CWGAN). AFHN generated diverse and discriminative features by conditioning a small number of labeled samples, with two innovative regularizers included to promote the discriminability and diversity of the synthesized features. Data Augmented Generative

Adversarial Network (f-DAGAN) proposed by Bharat Subedi et al. [18] was motivated by a DAGAN that learned data distributions from two real datasets. A dual discriminator was used to process the generated data as well as the resulting feature space to better learn the given data. Jiayu Xiao et al. [19] adapted a GAN well-trained on a large-scale source domain to the target domain with a few samples. Abhishek Sinha et al. proposed Diffusion-Decoding models with contrastive representations (D2C) [20], a method for training unconditional Variational AutoEncoders (VAE) for few-shot conditional image generation. For time-series data generation, Synthetic biomedical Signals GAN used bidirectional grid long short-term memory (BiGridLSTM) as the generator and Convolutional Neural Networks (CNN) as the discriminator [21]. Different features associated with each of the different signal types could be captured. Lue Zhang et al. [22] employed deep WAVEGAN and confirmed its effectiveness in generating realistic dolphin signals from both time and frequency domains. However, all these methods often require large network parameters and fail to account for the distance between different dolphin samples; thus, they are considered unsuitable for limited data and computing resources.

Based on the above-mentioned problems, an improved GAN method, Masked Autoencoder Generative Adversarial Network (MAE-GAN), was hereby proposed under limited data conditions. First, the data augmentation scheme was used to effectively expand the training set of the original signal. After that, an asymmetric autoencoder based on convolution was constructed by using an innovative mask mechanism to achieve an encoder model with high reconstruction capability, based on which the Conditional Wasserstein Generative Adversarial Network-Gradient Penalty (CWGAN-GP) learned the latent codes of MAE to achieve a whistle signal generation model with low computational resource requirements. The time-frequency contours, t-SNE plots, discriminant scores, and time and space complexity are subsequently used as the evaluation metrics. Experimental results showed that the network can generate dolphin signals better on low-power devices. In addition, this method works efficiently in follow-up tasks such as target recognition, which is of great significance in solving the problem of insufficient samples and provides new ideas for some data-driven methods.

In the following sections, the second part describes the signal extraction and synthesis method, data extension method, and model structure; the third section presents the details of the experimental setup; the fourth section introduces the experiments and analysis of the results; and the final section summarizes the experimental results and future perspectives.

2. Theory and Method

2.1. Dolphin Whistle Signal Modeling and Synthesis

To deal with the low signal-to-noise ratio found in collected dolphin sound signals, the time-frequency spectrum obtained through the Short-Time Fourier Transform (STFT) is hereby utilized, which enables the signal to be processed and analyzed to produce a distinct profile of its time-frequency spectrum.

2.1.1. Peak Detection

Assuming a whistle signal represented by $s[t]$ and a sampling rate of $1/T$, a sampled signal $s[n] = s[t/T]$ can be obtained. A time-frequency map can be drawn through STFT, where the window length is chosen to be point W. It is assumed that the whistle signal is smooth during the duration of point D. The signal, applied STFT with a window length of point W, is divided into M data blocks, and each block is assigned a data block number m to obtain the result $X_m[k]$. Then, the fundamental frequency can be expressed as follows:

$$f_1(m) = \text{argmax}|X_m\lfloor k \rfloor| \tag{1}$$

In Formula (1), the data block's ordinal number is m and the ordinal number of the result of STFT is k. The fundamental frequency of the mth data block is denoted as $f_1(m)$ and can be expressed using the peak extraction method.

Typically, the fundamental frequency of a signal has the strongest energy, with the energy intensity decreasing as the number of harmonics increases. In real samples, the energy of the second or third harmonic at certain frequency points may exceed that of the fundamental frequency signal, resulting in abnormal time-frequency profiles of the extracted whistle signal. To address this, a judgment condition must be added to the peak extraction method, which involves obtaining the frequency of $e(m)$ where the maximum energy occurs in the time-frequency signal, along with the corresponding energy value $e(m)'$ at one-half of this frequency. This can be expressed as follows:

$$\begin{cases} [e(m), index(m)] = \max|X_m(K)| \\ e(m)' = X_m(index(m)/2) \end{cases} \quad (2)$$

In Formula (2), $index(m)$ denotes the sample point value at the frequency corresponding to the maximum energy value $e(m)$ after a Fourier transform, i.e., the sample point corresponding to the fundamental frequency, and $e(m)'$ represents the energy value at $index(m)/2$.

Furthermore, whether the energy value of $e(m)$ is greater than half of the energy value at $e(m)'$ must be determined by a factor of a, and whether the energy value at $e(m)'$ is greater than the noise energy must be determined by a factor of b. If both conditions are simultaneously satisfied, the sampling point corresponding to the fundamental frequency will be changed to one-half of the frequency. This can be expressed as follows:

$$\begin{cases} e(m) > ae(m)' \\ e(m)' > bN_0 \end{cases} \quad (3)$$

In Formula (3), N_0 denotes the energy of the noise, and a and b refer to the coefficients of the decision.

2.1.2. Fitting

Peak-based extraction of time-frequency contours can result in non-smooth contours that deviate significantly from the actual time-frequency distribution. Curve fitting is employed to obtain smooth curves that accurately represent the data obtained from the method mentioned above while retaining sufficient signal details and smoothness. Thus, a two-layer fully connected neural network approach is used to accomplish this task. The fully connected neural network transforms the problem into a regression problem, and the network structure is depicted in Figure 1.

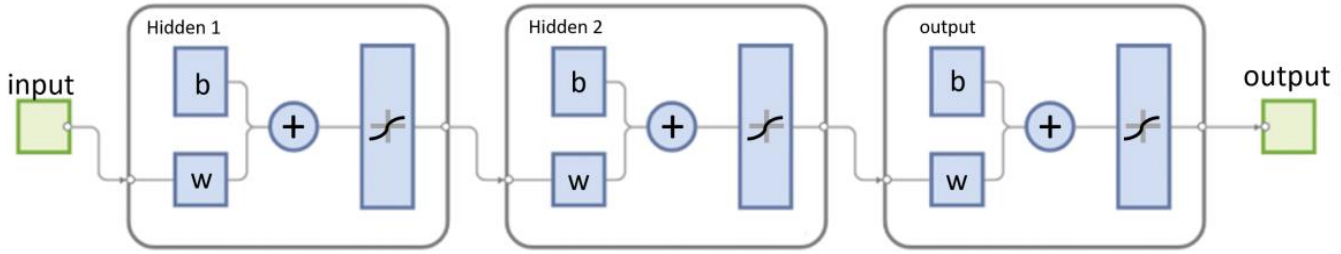

Figure 1. Fully connected neural network structure for curve fitting.

2.1.3. Synthesis

The dolphin signal is a frequency-tunable harmonic signal that can be modeled by its envelope, fundamental frequency, and harmonics. The frequency and amplitude parameters of the whistle signal after STFT are adopted and the sine wave model expressed by the following formula is used to synthesize the signal.

$$s[n] = \sum_{r=1}^{R} a_r[n] \sin(2\pi\phi_r[n]) \quad (4)$$

In Formula (4), R is the harmonic order; $a_r[n]$ indicates the amplitude of the nth point at the rth harmonic; and $\phi_r[n]$ is the phase of the nth point at the rth harmonic.

Therefore, the amplitude and phase of each harmonic in each sampling point of the signal must be clarified to synthesize the whistle signal. First, the amplitude $a_r[n]$ and phase $\phi_r[m]$ of the nth data block at the rth harmonic of the signal are calculated. Then, the amplitude $a_r[n]$ and phase $\phi_r[n]$ of each sampling point are obtained by interpolation.

1. Energy amplitude conversion: The formula for electrical signal power is $E = \frac{U_m{}^2}{R} \times t$, where E, U_m, R, t denote energy, maximum voltage, resistance, and time, respectively. Generally, it is considered that the resistivity of the signal circuit is very small, about 1 ohm. Thus, the formula can be written as $E = \frac{U^2}{2} \times t$, where U is the effective voltage. Therefore, $U = \sqrt{2E/t}$. To improve the signal-to-noise ratio, this paper uses $\sqrt{2}$ to amplify the signal. Assuming that the signal is stable within the D data range, it can be understood from Formula (4) that the energy of the mth data block of the rth harmonic is $e_r[m]$, so that each data sampling point becomes $a_r[mD] = 2\sqrt{e_r[m]/D}$. To obtain the value of each sampling point of the data block, the interpolation method is used to figure out the amplitude values of the remaining sample points $a_r[n]$.
2. Transient frequency phase conversion: The transient frequency represents the rotational speed of the complex plane vector argument, defined as the derivative of the phase. Hence, the estimate of the phase at each sample point becomes the integral of the transient frequency, that is, $\phi[n] = \sum_{i=1}^{n} f[i]$. As the D points are smooth, it is possible to use one of these points to represent the entire set of D points and achieve downsampling. After downsampling, the STFT is performed with a W point window, and the result is also smooth. Thus, within a certain window in the frequency domain, a single point can be used to represent the frequency of the whole window. Therefore, $f_r\ [m] = f\ [mDW]$, $\phi_r[m]$ is used as the sample point and the interpolation method is used to obtain the remaining phase values of $\phi_r[n]$ for the reconstruction of $\phi_r\ [n] = \phi\ [nD]$. As shown in Figure 2, the phases from the fundamental to the fifth harmonic are shown following the order from the bottom to the top.

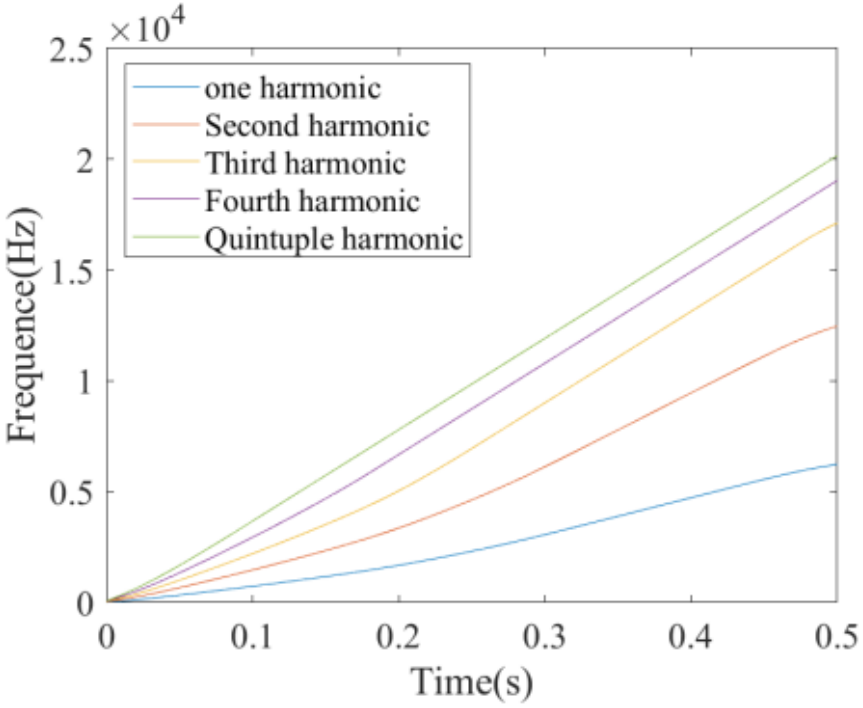

Figure 2. Phases of fundamental frequency and different harmonics.

Based on the above model, the measured data are reconstructed and the results are shown in Figure 3.

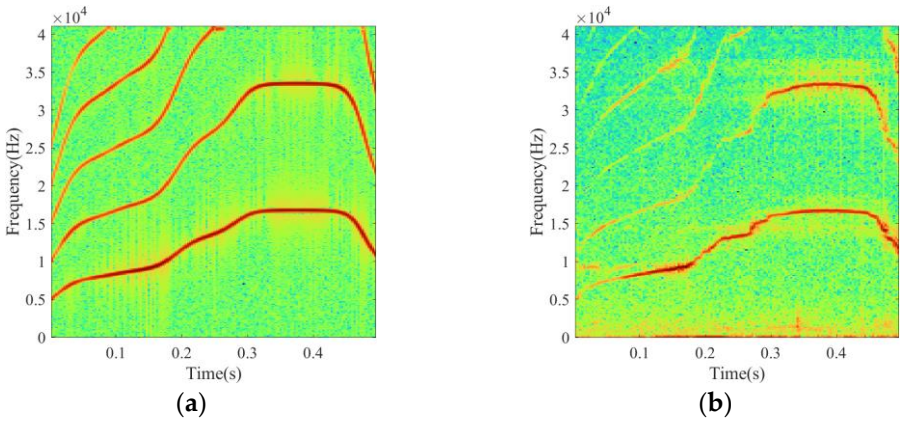

Figure 3. STFT spectrogram for dolphin signal: (**a**) synthesized signal; (**b**) original signal.

2.2. Data Augmentation Scheme

The frequencies of whistle signals of different individual dolphins are not the same, and there may also be large differences in the communication frequencies between the different populations. However, the time-frequency contours of the whistle signals of geographically adjacent dolphin populations are similar, and the sampled small sample data usually have only a small number of dolphins' signals, so the whistle signals of different dolphin signals can be simulated using frequency shifts. The different dolphin signals can be divided into multiple signals according to the time-frequency contours, and two types of signals are mainly explored here to achieve the generation of dolphin signals, so different weights are used for the weighting of the same type of signals. The average can effectively pull in the distance of different dolphin signals.

1. Frequency shift: Since more attention is hereby paid to the waveform of the signal in signal generation, the frequency of the original signal is shifted up and down as a whole to expand the data of the signal. This method can improve the robustness of subsequent tasks to frequency changes in different signals, and can be expressed as follows:

$$f_{a'}(t) = f_a(t) \pm m \tag{5}$$

In Formula (5), $f_{a'}(t)$ represents the new signal, $f_a(t)$ denotes the original signal, and m is the base of the offset.

2. Signal scaling: In the case of the inclusion of dolphin calls, there may be a Doppler shift, so signal spreading can be used for the simulation of this situation. In addition, this method can improve the feature extraction capability of the subsequent tasks [23], and the signal stretching can be expressed as follows:

$$f_{a'}(t) = \beta f_a(t) \tag{6}$$

In Formula (6), $f_{a'}(t)$ represents the new signal, $\beta f_a(t)$ refers to the original signal, and β is the scaling factor.

In the specific implementation, after normalizing the above two signals to a fixed length, their waveforms are the same. Thus, a small part of the signal at the beginning and end of the signal is cut out, and the signal is then normalized to the same length.

3. Weighted average: In a data-limited scenario, the data distribution may not be complete, and the data elements of the same type of data may be far apart, which will cause trouble for subsequent tasks. Therefore, the weighted average of the signals between the same type of signals can be achieved [24]. Obtaining a new signal biased towards one of the signals also results in an average signal of the two signals. The

method of weighted average makes the distribution of similar signals more uniform, and this method can be expressed as follows:

$$f(t) = \mu f_a(t) + (1 - \mu) f_b(t) \tag{7}$$

In Formula (7), $f(t)$ represents the new signal, $f_a(t)$ $f_b(t)$ denotes the same category in the original signal, and μ refers to the weight.

Figure 4 shows the results of frequency shifting, stretching, mixing, and the original signal on the same signal when the above three methods are used for data augmentation. For the same type of waveform, the basic waveform has not changed.

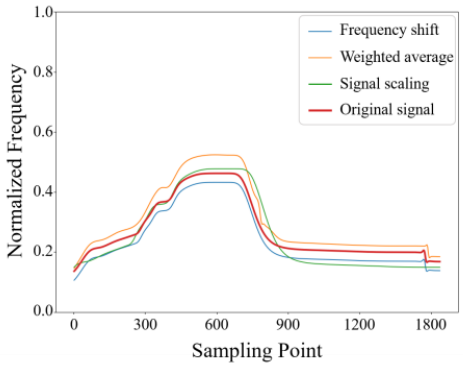

Figure 4. Changes to the waveform by the data augmentation scheme.

2.3. Proposed Method

MAE-GAN, as shown in Figure 5, mainly includes an autoencoder and a GAN. The encoder structure mainly improves the network generation effect and enhances the compatibility of UUVs and other equipment. GAN transforms a random noise into a hidden representation learned by an autoencoder. In the training phase, the autoencoder reduces the loss of reconstructed data and original data, whereas the generator reduces the loss of real hidden representation and generated data, and the discriminator reduces the loss of generated data and real data.

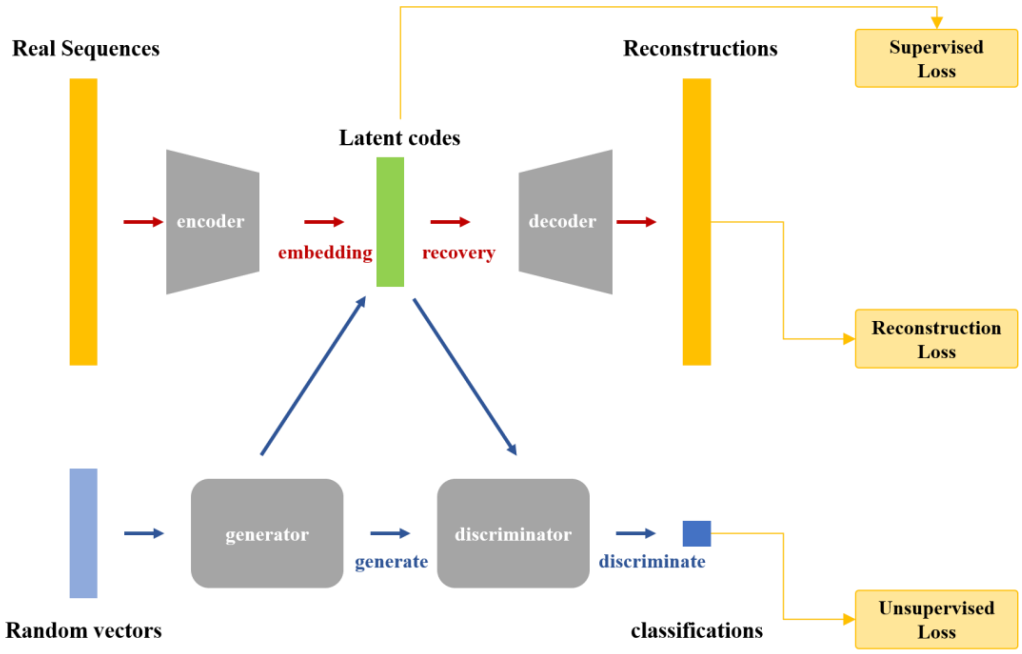

Figure 5. The structure of MAE-GAN.

For the GANs, the quality of generated data is more emphasized. Thus, for the combined network of the Autoencoder and GAN [25], improving the ability of the decoder or generator can effectively improve the quality of the generated data. It is easier to improve the ability of the decoder to handle limited data. Considering the finite computing resources and real-time requirements in UUVs and other equipment, the proposed method is different from most of the time-series data generation methods. Using Recurrent Neural Network (RNN) as an autoencoder, the proposed method, MAE-GAN, is based on convolution [26].

MAE-GAN uses the autoencoder and CWGAN-GP for joint training. The autoencoder uses the encoder to obtain the latent codes of the original data and the decoder to reconstruct the original signal. Specifically, the classical autoencoder structure of the input to latent codes can be expressed as follows:

$$l = f(x) \tag{8}$$

In Formula (8), x and l denote input and latent codes, and f represents the function mapping relationship from the input to the latent codes.

In denoising autoencoder, this process can be described as $l = f(n(x))$, with n denoting noise addition, and the hidden layer to the output layer can be written as follows:

$$\hat{x} = g(l) \tag{9}$$

In Formula (9), $\hat{x}$ denotes the output and g refers to the function mapping relationship from the latent codes to the output.

The goal of the optimization algorithm is to minimize the distance between the input and the output, which can be expressed as follows:

$$MinimizeLoss = dist(x, \hat{x}) \tag{10}$$

According to the principle of the autoencoder, it can be used for dimensionality reduction. On this basis, the denoising autoencoder can enhance the robustness of the data to noise and improve the stability of the implicit representation. It is attractive to use the latent representation for downstream tasks in computationally resource-sensitive devices. However, in the case of sparse samples, although both autoencoders and denoising autoencoders can perform the implicit representation of the data, overfitting or underfitting can easily occur. In addition, the generated implicit representation may differ from the original number of data categories, which adds new variables to the downstream task.

Herein, MAE is adopted to eliminate the influence of the above problems. Masks are used to randomly cover a part of the input data so that the encoder can only receive part of the original signal, and the decoder uses the complete original signal (MAE is an asymmetric self-encoder from this perspective) so that a more challenging task can be constructed to express the complete signal from the incomplete signal. In this case, compared to the above AE expression, the process of inputting MAE to latent codes can be expressed as follows:

$$l_m = mask(f(n(x))) \tag{11}$$

In Formula (11), *mask* denotes the process of adding masks. In this way, the process from latent codes to the output can be expressed as follows:

$$\hat{x} = g(l_m) \tag{12}$$

The rest is the same as other autoencoders, and this overall structure is shown in Figure 6.

MAE has been recently introduced by Kaiming He [27] on the framework of Transformer [28] and Vision Transformer (VIT) [29], and it also presented good results in other small sample size fields [30]. However, due to the performance considerations of UUV and other equipment, the calculation complexity and model size of the transformer model is costly. Therefore, MAE is a simulation of this architecture using a convolutional neural network. At the same time, He's MAE deals with image problems, in which transform-

ing two-dimensional data to one-dimensional data requires the introduction of position coding. Nevertheless, a convolution-based MAE structure for one-dimensional data is hereby proposed, which naturally has position information, so there is no need to add positional encoding.

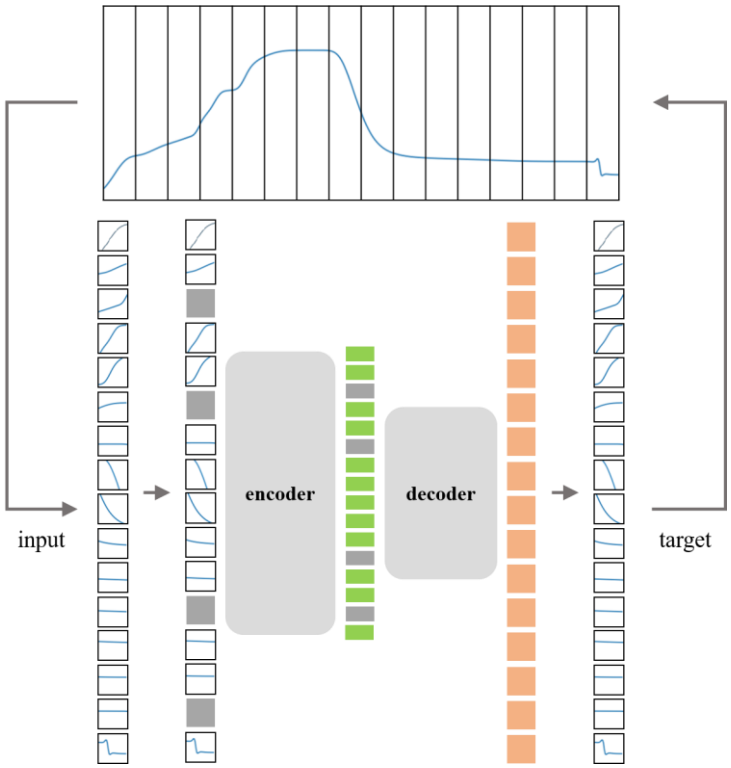

Figure 6. The structure of MAE in MAE-GAN.

In the data augmentation scheme, there is also the practice of discarding part of the data [31]. The present experiments have found that this method can also improve the effect of final data generation with the classic autoencoder structure, but it is still difficult to explain whether the improvement comes from the improvement of the feature extraction ability of the autoencoder or the improvement of the reconstruction ability. Based on the results of this paper, an autoencoder with a more powerful reconstruction capability is more effective.

From an implementation perspective, the structure of the autoencoder in the MAE is shown in Figure 7. First, the original signal is unified to the length (1824) by using the interpolation method. Then, the unified signal is divided into several (16) patches, and a small CNN is used by each patch of the original signal for feature extraction. For the unexposed data, the features of this part are directly replaced by 0, and then, those of these parts are spliced to obtain the implicit representation of the overall data. To be specific, let the input be s and divide the input into n equal parts, which can be expressed as Formula (13)

$$s = [s_1, s_2, \ldots, s_n] \tag{13}$$

The corresponding input to the function of the latent codes can be expressed as Formula (14):

$$f = [f_1, f_2, \ldots, f_n] \tag{14}$$

The generated mask (m) consists of a vector of length n with a definite proportion of 0 and 1, and then the latent codes of MAE can be expressed as follows:

$$l_m = [m_1 f_1(s_1), m_2 f_2(s_2), \ldots, m_n f_n(s_n)] \tag{15}$$

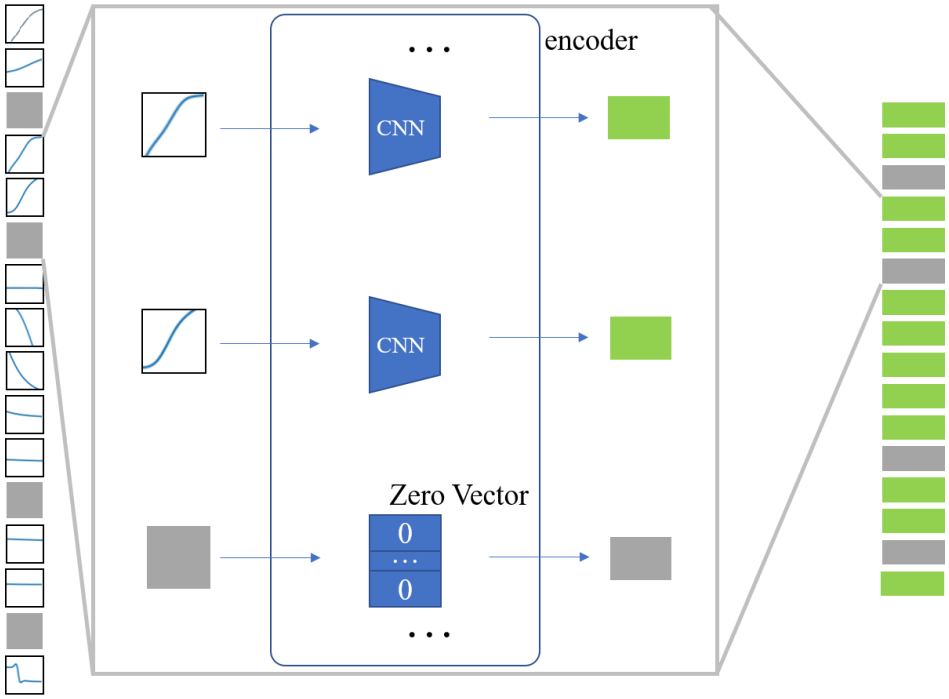

Figure 7. The structure of the encoder in MAE.

Afterward, the masked hidden variables are input into the decoder to restore the original complete input signal. Through this process, on the one hand, the decoder will learn global features from the missing latent coeds, thus avoiding the possible overfitting phenomenon of AE for small sample data and improving the recovery ability of the decoder, while on the other, the fusion of different features can be realized by randomly replacing the data of a specific signal with features from different signals. New signal data can be generated after passing through the decoder.

The hidden vector of the trained MAE is used as the learning goal of GAN. Considering the potential fusion of the previously mentioned features, CWGAN-GP is used in this paper. Hidden vectors are generated by using random noise, and a discriminator is used to discriminate against the hidden vectors generated by the generator. In this process, the inference of the mapping from the normal distribution noise to the distribution of latent vectors is implemented.

3. Data and Implementation

3.1. Dataset

The vocal sound signals of bottlenose dolphins can be divided into three main categories, i.e., click echolocation signals, whistle communications signals, and emergency burst signals. Dolphins are capable of precise individual communication in extremely complex environments with various noise disturbances. The signal is represented as a Frequency Modulated (FM) signal presenting certain time-frequency characteristics. According to the time-frequency contours, it can be subdivided into six categories, including the fixed frequency signal, the up-sweep frequency signal, the down-sweep frequency signal, the concave signal, the convex signal, and the sinusoidal signal.

The total number of signals in the dataset of this paper is 279, of which 13 are from the Woods Hole Oceanographic Institution database [32]. These signals include two categories, i.e., convex signal and up-sweep frequency signal. The sampling frequency of the signal is 81.9 kHz, and the effective frequency range of the signal is 3–40 kHz. In this paper, the fifth harmonic is taken into account, which can be regarded as a multiple of the fundamental frequency, so each data set is constructed by six rows, with one indicating the frequency

and the other five representing the amplitude. In this case, the size of the original data set is (279, 6, 1824).

3.2. Implementation Detail

In data augmentation, the frequency range of the time-frequency contour of the data is normalized from 0 to 40 k to 0–1. In the data expansion stage, the frequency offset operation is performed for the same segment of signals with parameters −0.02, −0.010, 0.010, and 0.02., respectively, whereas signal aliasing is performed for signals with parameters 0.015, 0.03125, and 0.0625, respectively. Signal aliasing with weights of 0.25, 0.5, and 0.75 is used for two signals of the same class.

The initial parameters of MAE-GAN included a batch size of 1824 and an epoch of 3000. Adam is selected as the optimizer for MAE. The learning rate is ε, the exponential decay rate for moment estimates is β_1 and β_2, and the values of ε, β_1, and β_2, are set as 1×10^{-4}, 0.5, and 0.9, respectively. For CWGAN-GP, using RMSPROP as the optimizer, the learning rate is 1×10^{-4}, and the hyper-parameter λ equals 10. The generator is updated once after the discriminator has been updated 5 times.

PyTorch 1.10.0 and Python 3.9 are applied for the training and evaluation of the proposed MAE-GAN. All experiments are based on a Windows 10 PC equipped with an Intel Core i5-10400 CPU, an Nvidia GeForce GTX 1070ti GPU, and 16 G RAM. The Nvidia CUDA version and cuDNN version are 11.7 and 8.2.1, respectively.

The construction of the dataset is implemented on the MATLAB R2022a platform, and the data extension work as well as the construction and operation of the generated model are based on the Python environment.

The implementation details of MAE-GAN are shown in Table 1.

Table 1. MAE-GAN internal details.

Model	1D CNN Layer §							Normalization	Activation
	Layer	Convolution Method	Input Channels	Output Channels	Kernel Size	Stride	Padding		
Encoder Block *	1	Conv1d	6	32	8	1	0	LayerNorm	LeakyReLu †
	2	Conv1d	32	64	8	1	0	LayerNorm	
	3	Conv1d	64	128	8	2	0	LayerNorm	
	4	Conv1d	128	64	5	1	0	LayerNorm	
Decoder	1	ConvTranspose1d	64	512	8	2	0	LayerNorm	LeakyReLu
	2	ConvTranspose1d	512	1024	8	2	0	LayerNorm	LeakyReLu
	3	ConvTranspose1d	1024	512	8	2	0	LayerNorm	LeakyReLu
	4	ConvTranspose1d	512	6	8	2	0	LayerNorm	LeakyReLu
	5	Linear	602	512	-	-	-	-	LeakyReLu
	6	Linear	512	256	-	-	-	-	LeakyReLu
	7	Linear	256	512	-	-	-	-	Sigmoid
CWGAN-GP-generator	1	ConvTranspose1d	100	512	4	1	0	LayerNorm	ReLu
	2	ConvTranspose1d	512	256	4	2	1	LayerNorm	ReLu
	3	ConvTranspose1d	256	128	4	2	1	LayerNorm	ReLu
	4	ConvTranspose1d	128	64	4	2	1	LayerNorm	ReLu
	5	Conv1d	64	64	4	2	1	LayerNorm	Tanh
CWGAN-GP-Discriminator	1	Conv1d	64	64	4	1	1	LayerNorm	LeakyReLu
	2	Conv1d	64	128	4	2	1	LayerNorm	LeakyReLu
	3	Conv1d	128	256	4	2	1	LayerNorm	LeakyReLu
	4	Conv1d	256	512	4	2	1	LayerNorm	LeakyReLu
	5	Conv1d	512	1	3	1	0	-	-

* In MAE-GAN, the encoder consists of several encoder blocks; † the parameter α of LeakyReLu is 0.2; § for Linear layer, the Input Channels and Output Channels represent the in features and the out features, respectively.

4. Experiments and Analysis

4.1. Evaluation Metrics

Herein, raw data, t-SNE graph [33], discriminative score [34], and space and time complexity are used as evaluation metrics.

Among these metrics, the raw data plot, instead of hearing, is used to observe the difference between the generated whistle time-frequency profile and the original whistle signal profile. The effect of signal generation can be observed from the difference in images. The t-SNE plots show the low-dimensional distribution of the generated and the original data; the closer they are to each other in terms of distribution, the better the effect will be. The discriminative score is an evaluation metric for the generated data obtained through supervised learning by using a fully connected network. The network is trained to label the two classes of real signals in the used dataset, followed by the creation of a classification model. The generated data are then subjected to a classification task, and the classification error rate of the discriminant is used to evaluate the quality of the generated data. Mathematically, the discriminant score can be expressed as follows:

$$DS = N_{wrong} / N_{all} \tag{16}$$

In Formula (16), N_{wrong} denotes the total number of misclassifications and N_{all} represents the number of all signals.

The space and time complexity mainly considers the performance requirements of the algorithm operations. For the discriminative score and the space and time complexity, the lower the value, the better the generated effect.

4.2. Evaluation Results

Herein, the Time-GAN [34] network [35] is used as the baseline model for comparison. The evaluation results of the comparative experiments are as follows. To visualize the features of the time-frequency profile, only the fundamental frequency profile is shown.

4.2.1. Raw Data

The whistle signal of dolphins has a high variability, the fundamental wave of its spectrogram can roughly reflect the characteristics of the signal, and the signals of different whistle signal spectrograms are not the same. However, in general, the fundamental frequency waveform of the whistle signal can be divided into several fixed categories. Figure 8 shows the waveforms of the two types of real signals. The waveform of the actual signal is indicated by the red line, and the waveform of the generated data is indicated by the blue and green lines. Comparing the two generated data, the generated data waveform of the proposed method is closer to that of the real signal. Although the baseline network is capable of extracting the outline of the waveform effectively, it tends to exhibit fluctuations in waveform details and experience feature fusion. For example, the waveform on the left illustrates fluctuations at its peak, which is common in the signal type on the right. Therefore, the proposed method was found to perform better in generating the whistle signals of the dolphin.

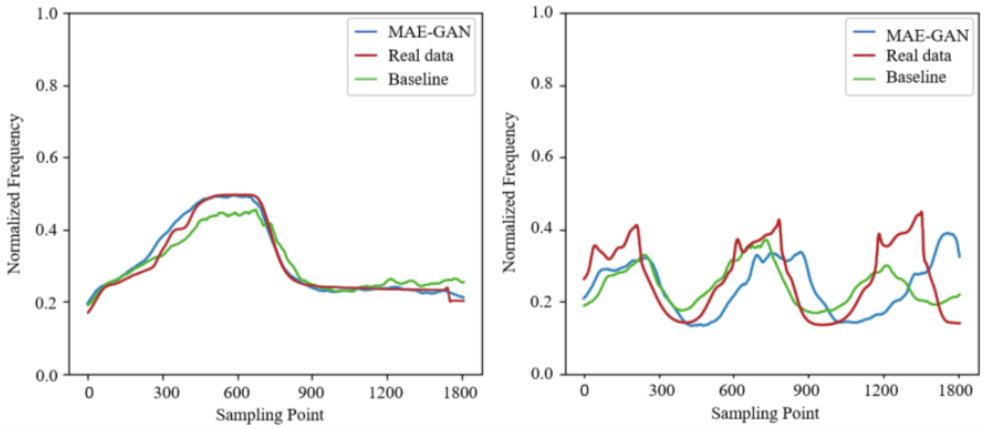

Figure 8. Fundamental frequency waveforms of real and generated dolphin signals.

4.2.2. t-SNE

To evaluate the similarity between generated and real data, t-SNE, an embedding model that can map data in a high-dimensional space to a low-dimensional space and preserve the local characteristics of the dataset, is hereby used to plot visualization example graphs of data point distributions mapped to two dimensions. t-SNE can be regarded as one of the most effective data dimensionality reduction and visualization methods. After the transformation, the data are not separable in the case of no separability in low-dimensional space, which could be attributed to either the dataset being inherently inseparable or the data within the dataset being unsuitable for projection into a low-dimensional space, and vice versa. Therefore, to verify the authenticity of the generated data, the application of t-SNE will contribute significantly. Based on its properties, the generated outcome is considered better when the distribution of the generated data closely or coincidentally matches that of the original data.

In Figure 9, the blue data represent the real data, whereas the orange represents the generated data. Subfigure (a) shows the data generated using the baseline model for the used dataset, and Subfigure (b) presents the data generated using MAE-GAN. It can be seen that the data generated using the proposed method perform better than those of the baseline model. In Subfigure (a), although the generated data partially fit the real data, there is a partial distance between the generated model and the true data distribution. In Subfigure (b), although MAE-GAN still has some offset, the overall generation mode has been dramatically improved compared with the baseline model.

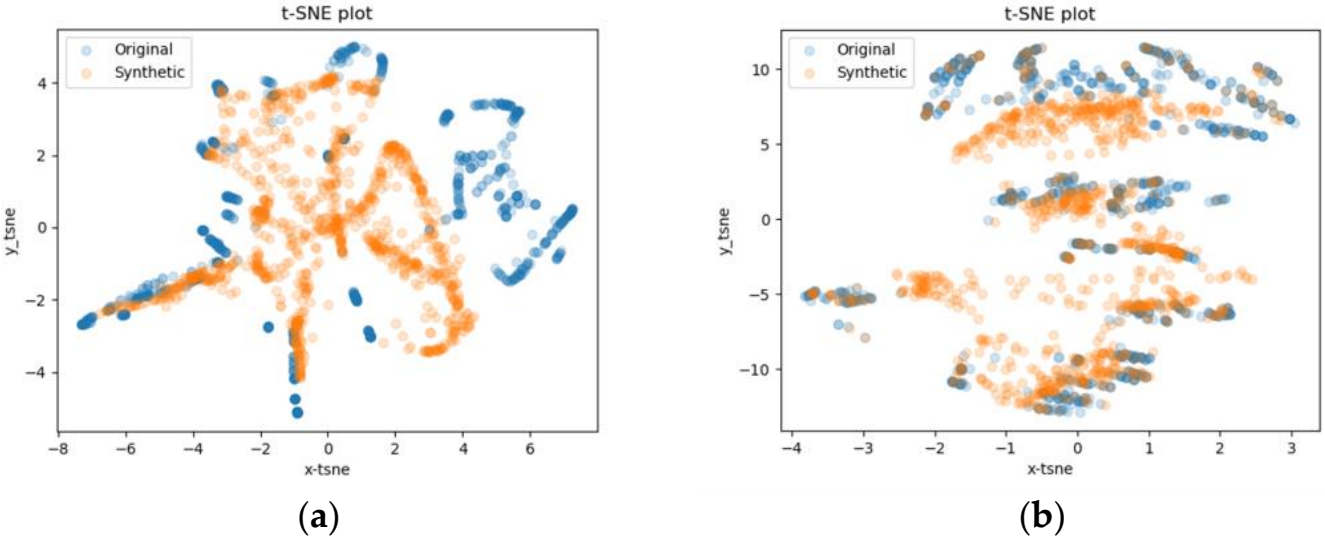

Figure 9. t-SNE results (blue for raw data and orange for generated data): (**a**) baseline model results; (**b**) MAE-GAN results.

4.2.3. Discriminative Score

In the discrimination process, it is not convincing enough to use only naked-eye observation. To this end, a discriminative neural network based on supervised learning is adopted to analyze the authenticity of the generated data. First, the dataset is constructed for the discriminant network using an equal amount of generated and real data, and the label "true" is assigned to the real data and "false" to the generated data. Then, a discriminant network is built. Herein, a three-layer fully connected neural network is used to classify the data. After 1000 epoch iterations, the target discriminant neural network converges. Finally, the error rate of the generated data test set is used as the evaluation index; therefore, the lower the discriminative score, the better the quality of the generated model. Table 2 shows the scores of the baseline model and MAE-GAN with the best results in bold, where the total generated data used is 15,000.

Table 2. Discriminative scores of Time-GAN and MAE-GAN generated data.

Method	Discriminative Score
Baseline model	0.362
MAE-GAN	**0.074**

It can be observed from Table 2 that MAE-GAN exhibits the lowest error rate. The relatively high error rate of the benchmark algorithm is attributed to the substantial discrepancy between the waveform profile of the fundamental wave generated by the baseline algorithm and the actual data.

4.2.4. Space and Time Complexity

To verify whether the proposed method can be adapted to real-time operation in low-energy and low-computing devices, this paper compares the parameters and calculations of MAE-GAN, Time-GAN, and deep WAVEGAN to generate the same data. Among these models, it can be concluded from Table 3, where the best results are shown in bold, that the Time-GAN network has the least number of parameters, but it requires a large amount of calculation. The deep WAVEGAN model demonstrates highly satisfactory results, but it has the most parameters and calculation resources, and its high calculation cost is unaffordable for UUVs. Additionally, the proposed MAE-GAN has fewer parameters and requires the least computing resources, i.e., 24% more parameters are required but only 30.2% of computation recourse is required.

Table 3. Comparison of time complexity and space complexity of different models.

Method	Param (M)	MFLOPs
MAE-GAN	1.65	**68.5**
Baseline model	**1.24**	226.7
Deep WAVEGAN	3.73	86,951.3

In practical implementation, the energy consumption of computing resources significantly surpasses that of memory, making the proposed method particularly well-suited for real-time operation in UUVs.

4.3. Ablation Experiment

The experiments in this paper use a new data enhancement scheme and a masked autoencoder. The following ablation experiment is designed to demonstrate the effectiveness of the simultaneous operation of these two modules. Based on the above experiments, three experiments have been added, including experiments using MAE but not data enhancement scheme; experiments using data enhancement scheme alone without MAE; and experiments that do not use either MAE or data enhancement scheme. When MAE is not used, a denoising autoencoder is used instead for comparison.

Figure 10 shows the raw data and t-SNE graphs of using only MAE without the data enhancement scheme; without using MAE but only the data enhancement scheme; and without both MAE and the data enhancement scheme. Figure 11 shows corresponding t-SNE plots, and Table 4 presents the corresponding discriminative scores.

The experimental results show that both the data augmentation scheme and MAE can improve the generated effect, but the practical improvement of a single scheme is relatively limited.

The discrimination scores corresponding to the above ablation experiments are shown in Table 3. The fusion of the two methods together brings a significant improvement, with an increase of 13.6%, compared to the best mono-method. Meanwhile, the effect of using MAE only is 8% higher than that without data augmentation and MAE, which can be attributed to the fact that although MAE possesses good feature extraction capabilities,

it still relies on a certain amount of data for support, which can be supplemented by the data enrichment approach. The reason for the severe degradation of generation quality when data augmentation methods are not used is that MAE is limited by the operating environment and can only use smaller convolutional kernels and depths, making it difficult to achieve better feature extraction capabilities. Meanwhile, the excessive data differences within the few-shot dataset, such as contour feature differences, worsen the above problem.

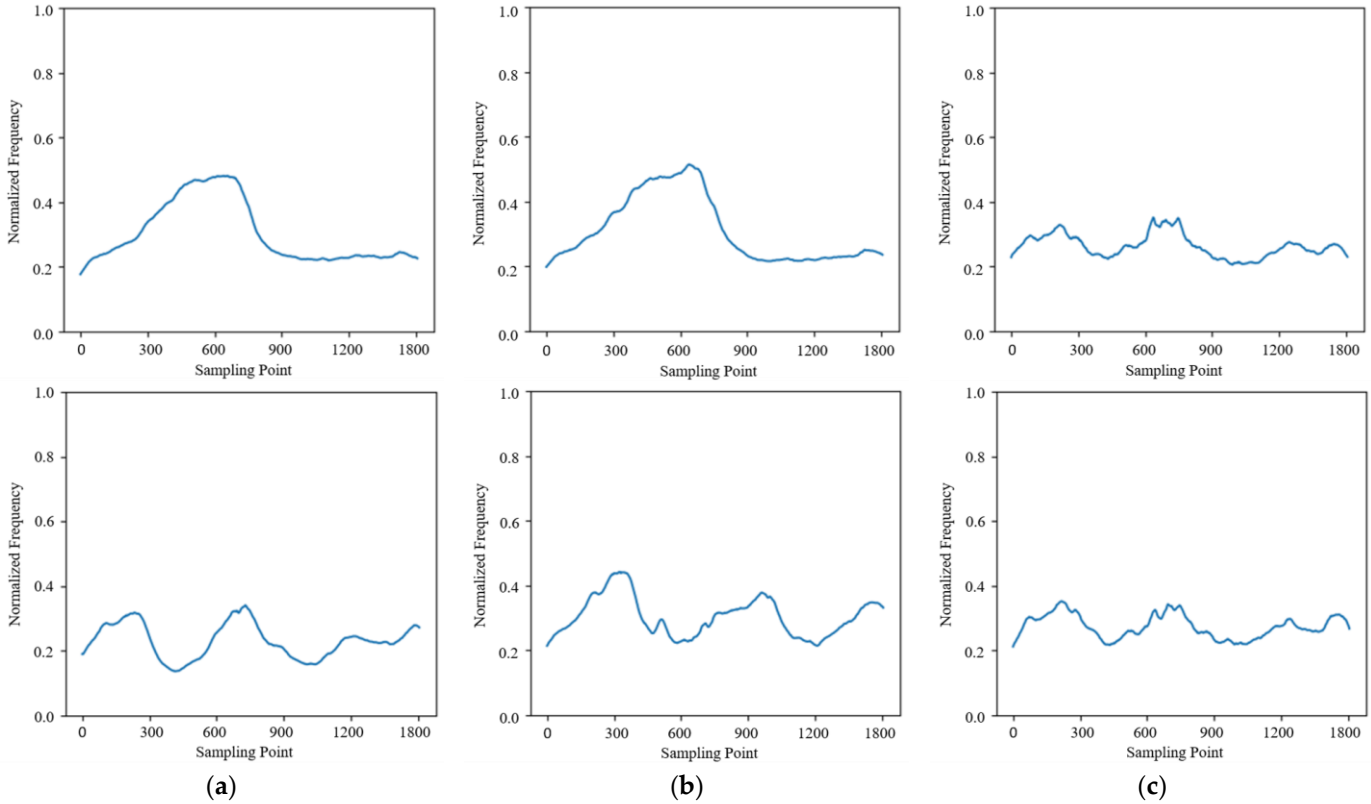

Figure 10. Fundamental frequency waveforms of generated dolphin signal: (**a**) data augmentation without MAE; (**b**) MAE without data augmentation; and (**c**) without data augmentation and MAE.

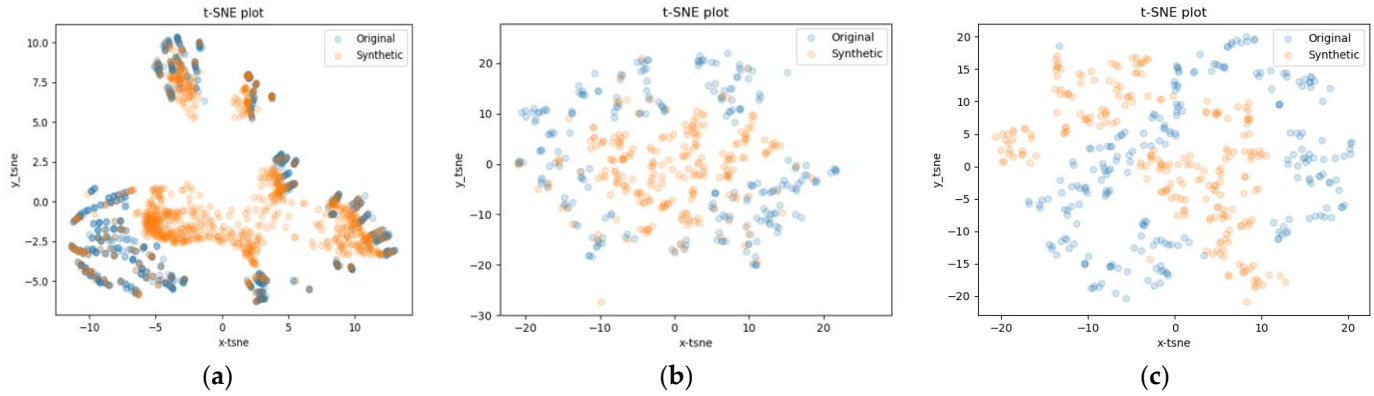

Figure 11. t-SNE diagram in the ablation experiment: (**a**) data augmentation without MAE; (**b**) MAE without data augmentation; and (**c**) without data augmentation and MAE.

Table 4. Discriminative scores of Ablation experiment results.

Method	Discriminative Score
Data augmentation without MAE	0.21
MAE without data augmentation	0.84
Without data augmentation and MAE	0.92

5. Conclusions

Herein, a novel approach for the real-time generation of dolphin whistle signals is proposed by using a data augmentation scheme and a convolution-based MAE-GAN network. The experimental results indicate that the proposed method can enhance the quality of dolphin signal generation under small sample conditions, and their combined application is more effective than that of either method alone. The proposed approach generates realistic dolphin whistle signals, as evidenced by the original data waveforms and t-SNE plots, and the discriminative scores proposed are upgraded by 28.8%, which confirms the good quality of the generated signals. To reduce computational resources, a CNN structure instead of the RNN structure used in the baseline model is hereby employed. The proposed method requires only 30.2% of the computing resources of the comparison network and enhances the expression capability of the decoder of autoencoder by using a masking mechanism. These two modifications significantly reduce the time complexity of the algorithm, making it suitable for devices with limited computational resources.

In conclusion, the proposed MAE-GAN network demonstrates outstanding performance in terms of both generation accuracy and efficient utilization of computational resources. However, the purity of the generated signal based on the sinusoidal model may require further research on the reconstruction method of the whistle signal. Furthermore, the application of a discard strategy to the mask may result in insufficient recovery of the self-encoder, causing occasional unavailability during the generation process.

Author Contributions: Conceptualization, Q.H., H.W. and Z.W.; methodology, X.W., H.W. and Y.H.; software, X.W., H.W., X.H. and C.H.; validation, H.W.; formal analysis, Q.H., X.W. and C.H.; investigation, H.W., X.H. and Y.H.; resources, Q.H.; data curation, H.W.; writing—original draft preparation, H.W.; writing—review and editing, Q.H., X.W. and H.W.; visualization, H.W.; supervision, Q.H. and X.W.; project administration, Q.H.; funding acquisition, Q.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Major Program of the National Natural Science Foundation of China under Grant No. 61890961, the General Program of the National Natural Science Foundation of China under Grant No. 61971412, the Basic Research Project of China under Grant No. JCKY2020110C074, and the Rapid Support Fund Project of China under Grant No. 61404150405.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, C.; Jiang, J.; Wang, X.; Sun, Z.; Li, Z.; Fu, X.; Duan, F. Bionic Covert Underwater Communication Focusing on the Overlapping of Whistles and Clicks Generated by Different Cetacean Individuals. *Appl. Acoust.* **2021**, *183*, 108279. [CrossRef]
2. Jiang, J.; Sun, Z.; Duan, F.; Fu, X.; Wang, X.; Li, C.; Liu, W.; Gan, L. Synthesis and Modification of Cetacean Tonal Sounds for Underwater Bionic Covert Detection and Communication. *IEEE Access* **2020**, *8*, 119980–119994. [CrossRef]
3. Gregorietti, M.; Papale, E.; Ceraulo, M.; de Vita, C.; Pace, D.S.; Tranchida, G.; Mazzola, S.; Buscaino, G. Acoustic Presence of Dolphins through Whistles Detection in Mediterranean Shallow Waters. *J. Mar. Sci. Eng.* **2021**, *9*, 78. [CrossRef]
4. Kipnis, D.; Diamant, R. Graph-Based Clustering of Dolphin Whistles. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2216–2227. [CrossRef]

5. Li, P.; Liu, X.; Palmer, K.J.; Fleishman, E.; Gillespie, D.; Nosal, E.-M.; Shiu, Y.; Klinck, H.; Cholewiak, D.; Helble, T.; et al. Learning Deep Models from Synthetic Data for Extracting Dolphin Whistle Contours. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–10.
6. Lin, C.-F.; Chung, Y.-C.; Zhu, J.-D.; Chang, S.-H.; Wen, C.-C.; Parinov, I.A.; Shevtsov, S.N. The Energy Based Characteristics of Sperm Whale Clicks Using the Hilbert Huang Transform Analysis Method. *J. Acoust. Soc. Am.* **2017**, *142*, 504. [CrossRef]
7. Yan, S.; Zhou, X.; Hu, J.; Hanly, S.V. Low Probability of Detection Communication: Opportunities and Challenges. *IEEE Wirel. Commun.* **2019**, *26*, 19–25. [CrossRef]
8. van der Merwe, J.R.; du Plessis, W.P.; Maasdorp, F.D.V.; Cilliers, J.E. Introduction of Low Probability of Recognition to Radar System Classification. In Proceedings of the 2016 IEEE Radar Conference (RadarConf), Philadelphia, PA, USA, 2–6 May 2016; pp. 1–5.
9. Stove, A.G.; Hume, A.L.; Baker, C.J. Low Probability of Intercept Radar Strategies. *IEE Proc. Radar Sonar Navig.* **2004**, *151*, 249–260. [CrossRef]
10. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. AutoAugment: Learning Augmentation Policies from Data. *arXiv* **2019**, arXiv:1805.09501.
11. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical Automated Data Augmentation with a Reduced Search Space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.
12. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
13. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. *arXiv* **2018**, arXiv:1710.09412.
14. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
15. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
16. Li, Y.; Zhang, R.; Lu, J.; Shechtman, E. Few-Shot Image Generation with Elastic Weight Consolidation. *arXiv* **2020**, arXiv:2012.02780.
17. Li, K.; Zhang, Y.; Li, K.; Fu, Y. Adversarial Feature Hallucination Networks for Few-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13470–13479.
18. Subedi, B.; Sathishkumar, V.E.; Maheshwari, V.; Kumar, M.S.; Jayagopal, P.; Allayear, S.M. Feature Learning-Based Generative Adversarial Network Data Augmentation for Class-Based Few-Shot Learning. *Math. Probl. Eng.* **2022**, *2022*, e9710667. [CrossRef]
19. Xiao, J.; Li, L.; Wang, C.; Zha, Z.-J.; Huang, Q. Few Shot Generative Model Adaption via Relaxed Spatial Structural Alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2022; pp. 11204–11213.
20. Sinha, A.; Song, J.; Meng, C.; Ermon, S. D2C: Diffusion-Decoding Models for Few-Shot Conditional Generation. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Curran Associates, Inc.: Cambridge, MA, USA, 2021; Volume 34, pp. 12533–12548.
21. Hazra, D.; Byun, Y.-C. SynSigGAN: Generative Adversarial Networks for Synthetic Biomedical Signal Generation. *Biology* **2020**, *9*, 441. [CrossRef]
22. Zhang, L.; Huang, H.-N.; Yin, L.; Li, B.-Q.; Wu, D.; Liu, H.-R.; Li, X.-F.; Xie, Y.-L. Dolphin Vocal Sound Generation via Deep WaveGAN. *J. Electron. Sci. Technol.* **2022**, *20*, 100171. [CrossRef]
23. Varghese, T.; Ophir, J. Enhancement of Echo-Signal Correlation in Elastography Using Temporal Stretching. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **1997**, *44*, 173–180. [CrossRef]
24. Forestier, G.; Petitjean, F.; Dau, H.A.; Webb, G.I.; Keogh, E. Generating Synthetic Time Series to Augment Sparse Datasets. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; pp. 865–870.
25. Lu, J.; Yi, S. Autoencoding Conditional GAN for Portfolio Allocation Diversification. *AEF* **2022**, *9*, 55. [CrossRef]
26. Zhang, Q.; Lin, J.; Song, H.; Sheng, G. Fault Identification Based on PD Ultrasonic Signal Using RNN, DNN and CNN. In Proceedings of the 2018 Condition Monitoring and Diagnosis (CMD), Perth, WA, Australia, 23–26 September 2018; pp. 1–6.
27. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Cambridge, MA, USA, 2017; Volume 30.
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
30. Mao, J.; Zhou, H.; Yin, X.; Xu, Y.C.B.N.R. Masked Autoencoders Are Effective Solution to Transformer Data-Hungry. *arXiv* **2023**, arXiv:2212.05677.

31. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13001–13008. [CrossRef]
32. Sayigh, L.; Daher, M.A.; Allen, J.; Gordon, H.; Joyce, K.; Stuhlmann, C.; Tyack, P. The Watkins Marine Mammal Sound Database: An Online, Freely Accessible Resource. *Proc. Mtgs. Acoust.* **2016**, *27*, 040013. [CrossRef]
33. Arora, S.; Hu, W.; Kothari, P.K. An Analysis of the T-SNE Algorithm for Data Visualization. In Proceedings of the Conference On Learning Theory, PMLR, Stockholm, Sweden, 6–9 July 2018; pp. 1455–1462.
34. Pei, H.; Ren, K.; Yang, Y.; Liu, C.; Qin, T.; Li, D. Towards Generating Real-World Time Series Data. In Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 7–10 December 2021.
35. Yoon, J.; Jarrett, D.; van der Schaar, M. Time-Series Generative Adversarial Networks. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates, Inc.: Cambridge, MA, USA, 2019; Volume 32.