# Timbre Transfer For Ship Radiated Noise

1st Nils Müller
*ATLAS Laboratory (A-LAB)*
*ATLAS ELEKTRONIK GmbH*
Bremen, Germany
nils.mueller@atlas-elektronik.com

2nd Jens Reermann
*Advanced Development Programs*
*ATLAS ELEKTRONIK GmbH*
Bremen, Germany
jens.reermann@atlas-elektronik.com

3rd Tobias Meisen
*TMDT Institute*
*Bergische Universtität Wuppertal*
Wuppertal, Germany
meisen@uni-wuppertal.de

***Abstract*—The objective of Underwater Acoustic Target Recognition is to classify vessels based on their unique acoustic signatures. Deep learning shows promise in UATR, but its effectiveness relies on large and diverse datasets. Existing public datasets like ShipsEar and DeepShip capture inter-ship but lack sufficient intra-ship variability, e.g. varying operational conditions. This limits the generalization and robustness of the model. To address this, we propose a timbre transfer approach using a hierarchical Vector Quantized Variational Autoencoder to separate static timbre features from dynamic noise. By incorporating ship-specific spectral characteristics with Adaptive Instance Normalization, our method generates realistic, variably conditioned acoustic signals, improving data augmentation, and enhancing recognition algorithms for maritime surveillance and environmental monitoring.**

***Index Terms*—sonar, ship radiated noise, UATR, timbre transfer, deep learning**

## I. Introduction

Ship-radiated noise is crucial in maritime applications such as environmental monitoring and naval operations, where it aids in identifying and classifying vessels using passive sonar. However, accurately distinguishing ships in real-world conditions is challenging due to the complexity of underwater acoustics, including factors such as varying operational speeds, environmental conditions, and background noise, which affect the spectral characteristics of the noise. Underwater Acoustic Target Recognition (UATR) has emerged as a key research area, using machine learning and deep learning to identify ships based on their acoustic signatures [1], [2]. Although deep learning models have shown strong performance, their generalization to varying operational conditions is hindered by the lack of sufficient intra-ship variability, i.e., recordings of the same ship under different conditions, in large-scale datasets such as ShipsEar [3] and DeepShip [4].

Generative models, particularly style transfer techniques, have shown success in both image and audio synthesis, excelling at generating diverse data while preserving domain-specific characteristics. Building on these advancements, this work introduces a novel audio style transfer framework for ship-radiated noise. By transferring the measured acoustic signatures of real ships to a controllable simulation, we can generate realistic, versatile ship noise data tailored to specific conditions. This study investigates whether a hierarchical Vector Quantized Variational Autoencoder (VQ-VAE) -based timbre transfer framework can enhance ship noise synthesis and improve the generalization of deep learning models in the UATR task.

We make three key contributions. First, we introduce a timbre transfer framework that learns the residual between simulated and real ship noise to improve realism. Second, we develop a controllable parametric narrowband model for adjusting key acoustic features like propulsion and propeller behaviour. Third, we present the first baseline for timbre transfer in underwater acoustics, paving the way for future work in data augmentation and scene simulation.

## II. Related Work

Early ship radiated noise models relied on statistical and physics-based approaches, such as Fourier synthesis and additive noise models, to reconstruct ship noise signatures [5], [6]. These methods, while useful, lacked adaptability to real-world variations and could not generalize across diverse operating conditions and sea regions. The lack of generalizability is also stated in the UATR review by Hummel et al. [7].

Various architectures have been explored to improve classification performance, such as the contrastive learning approach of Xie et al. [1] or the joint model of Tian et al. [2]. However, while the publicly available datasets capture inter-ship variability (differences between ships), they lack intra-ship variability (recordings of the same ship under different conditions). This limits the generalization of the model, making adaptation to varying speeds, loads, and environments challenging for real-world UATR applications.

Generative models have advanced in audio processing, with techniques such as style transfer enabling the transformation of sound sources while preserving acoustic features. Notable methods include RAVE by Caillon and Esling [8] and DDSP by Engel et al. [9]. However, because of the stochastic and nonharmonic nature of ship radiated noise, these approaches face challenges when applied to ship radiated noise.

In underwater acoustics, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been explored for data augmentation and improved recognition performance. GANs have been widely utilized to generate time-frequency representations for ship radiated noise [10], [11], with Ashraf et al. introducing AN-GAN to enhance the signal-to-noise ratio [12]. However, these methods rely on spectrogram-based representations, which require transformations back to the time domain, which is non-trivial due to the lack of phase information. Accurate phase reconstruction is crucial in this application, as phase carries fine-grained temporal details and that are essential for accurately preserving multipath propagation and Doppler effects, for instance. Changing towards raw waveform generation, Atanackovic et al. [13] used GANs to model ship radiated noise, and Li et al. [14] proposed a GAN-based approach to generate realistic ship noise. Qiu et al. [15] explored VAEs and autoregressive models to generate radiated ship noise, introducing cross-domain pretraining and controllable noise synthesis.

These advancements highlight the potential of generative models for ship radiated noise synthesis. However, these methods struggle to integrate controllability over operating conditions with the objective of enriching variability. We therefore propose a solution by providing controllable variability with a deterministic simulation, while preserving authenticity using a style transfer approach.

## III. Method

This section presents the proposed framework for ship-radiated noise synthesis, built on a Vector Quantized Variational Autoencoder (VQ-VAE) with Adaptive Instance Normalization (AdaIN) to match the simulation output features with the style reference. Our approach uses a two-stage training process focusing on content preservation and style adaptation respectively: First, the model reconstructs simulated content signals, and second, it transfers real ship noise characteristics onto these signals. The following subsections detail the core components of the method, including the architecture and training procedure.

### A. Adaptive Instance Normalization

Our approach builds on the adaptive instance normalization proposed by Huang et al. [16], which matches the feature statistics of arbitrary reference inputs to a given content input. This work is based on the observation made in [17], [18], where the authors demonstrate that style characteristics can be described using feature statistics. Although AdaIN has been successful in image tasks, its application to audio is limited. In [19], the authors use AdaIN in the audio domain for spectrogram-based disentanglement of speech content and emotion. The authors in [20] successfully extended AdaIn to raw waveforms for speech conversion tasks. The RAVE model [8] also applied AdaIN to raw waveform inputs. In our work, we extend AdaIN to ship noise synthesis, using the output of the deterministic simulation as the content, and samples from the DeepShip dataset as the style references. The stylized content feature $t$ using AdaIN is determined by normalizing the features of the content signal $z(c)$ to the features of the style reference $z(s)$, where $z$ represents the latent vector and $s$ and $c$ are the style and content signals, respectively by

$$t\left(c,s\right)=\sigma\left(z\left(s\right)\right)\left(\frac{z\left(c\right)-\mu\left(z\left(c\right)\right)}{\sigma\left(z\left(c\right)\right)}\right)+\mu\left(z\left(s\right)\right) \qquad (1)$$

The latent content representation is normalized and scaled by the spatial-wise mean $\mu$ and standard deviation $\sigma$ of the style representation, where $t$ corresponds to the stylized content feature.

### B. Architecture

The core architecture is inspired by the VQ-VAE architecture from Oord et al. [21] and the hierarchical structure presented in [22], where the intuition is for the lower levels to capture style and texture, while higher levels focus on semantic features. This structure naturally supports style-content separation, as suggested in [16], and enhances signal fidelity. The encoder and decoder architectures are similar to the model presented by Dhariwal et al. [22], which is known for processing raw waveforms in a VQ-VAE setup. As ship radiated noise is characterized by transient and long-term acoustic patterns, we inspired our model setup by the Open AI hierarchical Jukebox model which is able to captures both local and global features. The encoder and decoder use stacked residual layers with smooth down- and upsampling to minimize aliasing artifacts. Combining the strengths of [16], [22], this architecture effectively handles raw waveform-based style transfer. The proposed architecture is shown in Figure 1.

The content signal and style references are input to the encoder as one-dimensional raw waveforms. Following [22], we use decoupled encoders to map inputs to different compressions to emphasize capturing short and long term context at the different hierarchy layers. However, applying the nested encoder structure from [23] led to insufficient information flow to deeper layers, as early encoders overfit. Both content and style signals are encoded using residual 1D convolutional layers. After encoding, content features $z_i\left(c\right)$ are normalized and scaled to style features $z_i\left(s\right)$ via the AdaIN layer. Vector quantization begins with high-level features ($E_1$ in Figure 1). The stylized feature vector $t_i\left(c,s\right)$ is mapped to the nearest codewords in the codebook $C_i$. These quantized features are decoded to the next lower layer ($E_0$), added to the stylized features, and the process repeats for each hierarchy level. The lowest-level decoder combines and maps features back to the original input dimension.

### C. Training

Our training follows the two-stage approach from Huang et al. [16]. In the first stage, we pretrain our hierarchical VQ-VAE to reconstruct simulated ship noise, creating a latent space that encodes key acoustic properties. In the second stage, a training scheme similar to the original AdaIN approach [16] is used to align the characteristics of generated noise with real-world ship recordings.

In the first phase, encoders map the raw waveform into a scaled quantized latent space which is decoded back to the original signal dimensions and evaluated using a reconstruction loss. The training objective includes a time-domain L1 loss for fidelity and phase preservation, alongside a spectral convergence loss to ensure accurate frequency bin reconstruction. This phase establishes a structured latent space for the subsequent style transfer.

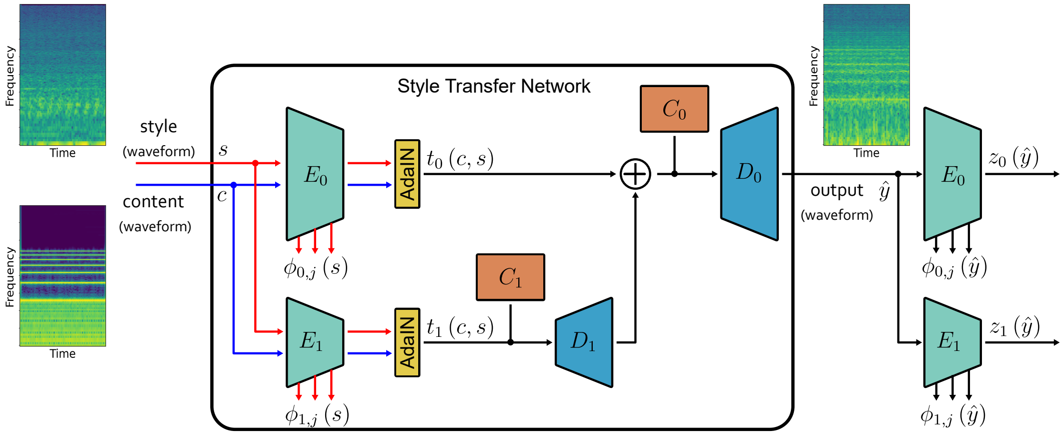

Fig. 1: Proposed hierarchical architecture for time domain timbre transfer using adaptive instance normalization and vector quantization. In our work we utilize an analogous three-level hierarchy.

In the second stage, the pre-trained encoders and codebooks are frozen. The combined decoder (shown in blue in Figure 1) is trained using a weighted loss function that combines content- and style-based losses. The weighted loss is defined by $\mathcal{L} = \mathcal{L}_c + \lambda\mathcal{L}_s$, where $\mathcal{L}$ resembles the total loss, $\mathcal{L}_c$ the content based, $\mathcal{L}_s$ the style based loss which is weighted by the factor given by $\lambda$. The content loss $\mathcal{L}_c$ is the Euclidean distance between the encoded content features $z_j(c)$and the encoded features of the output signal $z_j(\hat{y})$. The distances are simply added for every $i$-th hierarchical layer, where $L$ represents the total number of levels. The style loss matches the statistics of the style features to the content features. Therefore, the L1 loss of the mean $\mu$ and the standard deviation $\sigma$ in different intermediate encoding layers $\phi$ is calculated by

$$\begin{aligned}\mathcal{L}_c &= \sum_{i=1}^{L} ||\, z_i(\hat{y}) - z_i(c)\,||_2 \\ \mathcal{L}_s &= \sum_{i=1}^{L}\sum_{j=1}^{K} ||\, \mu(\phi_{i,j}(\hat{y})) - \mu(\phi_{i,j}(s))\,||_2 \\ &+ \sum_{i=1}^{L}\sum_{j=1}^{K} ||\, \sigma(\phi_{i,j}(\hat{y})) - \sigma(\phi_{i,j}(s))\,||_2\end{aligned} \tag{2}$$

where the index $j$ represents the layer index of the $K$ intermediate encoder outputs.

## IV. Experiments

We employ a three-level hierarchy compressing the input waveform by $1/4$, $1/8$, and $1/16$, ensuring sufficient reconstruction despite some high-frequency artefacts. Higher compression ratios led to reconstruction collapse during training.

For the content data we model 1000 ships with random configurations (in terms of propeller blades, cylinders, stroke, gear ratios, harmonics, amplitudes, and phase offsets) according to the frequency relations outlined in [24]. Each fundamental frequency is accompanied by an additive harmonic set. For each ship 20 three-second recordings at 16 kHz with constant crank shaft ratios are generated. This method simplifies the process while effectively capturing the essential characteristics of ship-radiated noise in a controlled manner. We choose a signal length of three seconds as it is long enough to capture the minimal expected frequency of 1 Hz multiple times. The style reference comes from DeepShip [4], sampled at 16 kHz and split into non-overlapping 3 s frames. Training, validation, and testing use ship-wise separated hold-out sets in a $0.7/0.2/0.1$ ratio. Random Gain and Polarity Inversion augmentation help prevent overfitting. Mixing typical low-frequency signals with high-frequency simulated signals improves high-frequency encoding, ensuring harmonic frequencies reach up to 6 kHz. For all training stages, the Adam optimizer (lr $= 0.001$, linear decay $0.1$ over 20 epochs) was used.

## V. Evaluation

Generative models are evaluated using either qualitative metrics like the Mean Opinion Score (MOS) or quantitative measures. In style transfer tasks, the focus is on balancing content preservation and style adaptation. To evaluate our timbre transfer framework, we conduct experiments across two key aspects. **Spectral Fidelity**: How closely does the synthesized ship noise match real-world spectra? We assess this using qualitative waveform analysis and logarithmic spatial distance (LSD). **Content Preservation vs. Style Adaptation**: Does the model maintain narrowband components from the simulated input while adapting the broadband structure of the style reference? We introduce a masked frequency ratio (MFR) to compare the preservation of the harmonic frequencies of the input along additional correlation analysis to compare the global structure.

### A. *Qualitative Evaluation*

We analyse content preservation and style adaptation by expecting a combination of broadband style characteristics with narrowband content components. Figure 2 shows the spectrograms of the generated mixtures given different content and style references. It can be seen that the proposed approach is able to preserve most of the harmonic structure in the generated signals. Generally, the preservation of higher frequencies is more stable than for low frequency (see B-I). A reason for this might be the large receptive fields required for low frequencies when handling raw waveforms. Additionally, when providing a single sine wave (II) the model automatically outputs corresponding harmonic frequencies, indicating that the model learns and assumes an intrinsic harmonic structure for given inputs. The two examples from the validation set A and B show that the model adapts to the narrowband characteristic of the content input and to the broadband characteristic of the style reference. The examples for C and D demonstrate the suppression of content-like frequency lines from style references, as artificially added frequency lines and chirp signal are fully suppressed in the generated output. The generated samples for the style reference E shows that the model fails to transfer amplitude variations, such as fading and short time bursts to the output. However, we expect that thus issue can be addressed with directed augmentations. The examples for the style reference F show that the model is able to generalize to unseen style references to some extent, nonetheless the results still differ in the general broadband structure. However, it is important to note that the sample from the ShipsEar has a higher low frequency energy density compared to the DeepShip samples, indicating a difference in the data distribution.

### B. Quantitative Evaluation

We quantitatively assess content preservation, style adaptation, and spectral fidelity by calculating reconstruction loss with a pre-trained autoencoder on style data from the DeepShip dataset. The autoencoder learns to reconstruct the stylized signals. A low reconstruction loss for the generated samples indicates alignment with realistic ship noise. We compare time-domain L1 and spectral distance measures between content and style signals. Table I reports reconstruction differences via median, Inter Quartile Range (IQR), skewness, and kurtosis. The temporal L1 distance shows the generated signals align more closely with the style reference than the content signal, demonstrating strong style adaptation. The L1 distribution further supports this alignment in time-domain characteristics. In the spectral domain, Log Spectral Distance shows similar medians but differing skew, indicating broadband and narrowband alignment. Despite some frequency variations, overall spectral characteristics remain comparable. The largest difference appears in Spectral Convergence, distinguishing content from style, though the generated signals align well with style, showing minimal deviations.

TABLE I: Deviation of reconstruction loss distribution compared to style reference for temporal and spectral L1 loss, as well as the Log Spectral Distance (LSD) and Spectral Convergence (SC). Lower Values are better.

| | | median | IQR | skewness | kurtosis |
|---|---|---|---|---|---|
| L1 (temp.) | content | -0.002 | -0.056 | 0.404 | 1.888 |
| | generated | **0.000** | **0.007** | **0.091** | **0.161** |
| LSD | content | 0.397 | -0.288 | 2.746 | 15.463 |
| | generated | **-0.135** | **0.065** | **0.399** | **-0.749** |
| SC | content | 1.407 | 0.416 | 0.788 | 11.258 |
| | generated | **0.025** | **0.015** | **-0.539** | **-2.135** |

Next, we quantify content preservation using a masked frequency ratio. The narrowband spectral lines of the content signals are well-defined due to deterministic simulation. As the content lacks a broadband component, a precise frequency mask which filters only the narrowband lines is determined. This metric evaluates how well these frequency lines are preserved in the generated signal by comparing their relative magnitudes. The Masked Frequency Ratio (MFR) can therefore determined by

$$MFR = \frac{\sum_k f_{mask,k} \mid \mathcal{F}(\hat{y}) \mid_k}{\sum_k f_{mask,k} \mid \mathcal{F}(c) \mid_k} \qquad (3)$$

where $f_{mask,k}$ describes the masked spectrum for the $k$-th narrowband component. This metric outputs $1.0$, when all the masked frequency magnitudes in the generated signal $\hat{y}$ match the masked frequency magnitude of the content signal $c$. Additionally, the similarity between the content and generated time-domain, as well as spectrum are evaluating using the Pearson's correlation factor $\sigma_{temp}$ and $\sigma_{freq}$. Table II displays the content preservation metrics for the four test signals, white noise, pink noise, style and the generated outputs.

TABLE II: Evaluation of the content signal preservation between the style reference and the generated outputs in terms of the mean MFR over the test set alongside the correlation of the time domain signal and corresponding spectrum. Higher Values are better.

| | MFR | correlation (temp.) | correlation (freq.) |
|---|---|---|---|
| | | $\sigma_{temp}$ | $\sigma_{freq}$ |
| style | 0.122 | 0.000 | 0.244 |
| generated | **0.358** | **0.380** | **0.533** |

Table II shows the highest MFR and correlation factors for the generated signals. In the time domain, none of the signals correlate with the content signal, but the best frequency-domain correlation is with the style reference. The highest MFR for the generated signal is $0.358$, indicating some content loss during style transfer. Improving content preservation requires balancing with style adaptation, which can be achieved through stronger content loss penalization. The results demonstrate that our method effectively transfers real-world spectral characteristics to simulated ship noise while preserving its structure.

## VI. Conclusion

In this work, we present a novel approach for generating realistic and authentic ship-radiated noise by refining a deterministic and controllable simulation with real-world data. Using a hierarchical VQ-VAE model combined with AdaIn,

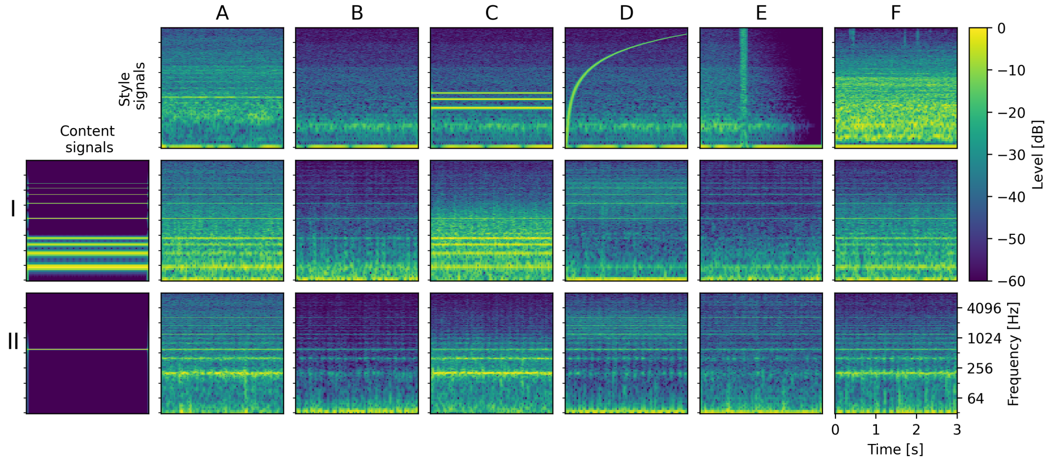

Fig. 2: Roman numerals I and II show a simulated content signal and a 600 Hz sine wave. Style references A–F include various examples: A and B are DeepShip samples, C–E are modified versions of B with added sine, chirp, and noise effects, and F is a sample from the ShipsEar dataset. Intersections of content I, II with styles A–F display the generated outputs.

our method effectively transfers the spectral characteristics of real ship noise to the simulated content. To our knowledge, this work that investigates style transfer for ship radiated noise synthesis, offering an innovative framework that can serve as a foundation and benchmark for future research. Through both qualitative and quantitative evaluations, we demonstrate the model's ability to preserve key features of the original content signal while adapting the broadband structure to real-world ship noise. Despite some limitations in low-frequency preservation, the model successfully replicates the overall spectral characteristics and texture of ship-radiated noise. The results emphasize the tradeoff between content preservation and style adaptation, suggesting that refining content loss penalization could further enhance the model. Overall, this approach provides a strong starting point for generating realistic ship noise.

## References

[1] Y. .Xie, J. Ren, and J. Xu, "Guiding the underwater acoustic target recognition with interpretable contrastive learning," *OCEANS 2023 - Limerick, OCEANS Limerick 2023*, 2023.

[2] S.-Z. Tian, D.-B. Chen, Y. Fu, and J.-L. Zhou, "Joint learning model for underwater acoustic target recognition," *Knowledge-Based Systems*, vol. 260, 2023.

[3] David Santos-Domínguez, Soledad Torres-Guijarro, Antonio Cardenal-López, and Antonio Pena-Gimenez, "Shipsear: An underwater vessel noise database," *Applied Acoustics*, vol. 113, pp. 64–69, 2016.

[4] M. Irfan, Z. Jiangbin, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, "Deepship: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification," *Expert Systems with Applications*, vol. 183, p. 115270, 2021.

[5] J. Luo and Y. Yang, "Simulation model of ship-radiated broadband noise," in *2011 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pp. 1–5, IEEE, 9/14/2011 - 9/16/2011.

[6] Y. Zheng, C. Hu, and H. Zhao, "On the active control of spectral lines of ship radiated noise," in *2014 12th International Conference on Signal Processing (ICSP)*, pp. 2422–2425, 2014.

[7] Hilde I. Hummel, Rob van der Mei, and Sandjai Bhulai, "A survey on machine learning in ship radiated noise," *Ocean Engineering*, vol. 298, p. 117252, 2024.

[8] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *CoRR*, vol. abs/2111.05011, 2021.

[9] J. H. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: differentiable digital signal processing," *CoRR*, vol. abs/2001.04643, 2020.

[10] Fan Liu, Qingzeng Song, and Guanghao Jin, "Expansion of restricted sample for underwater acoustic signal based on generative adversarial networks," pp. 1222–1229, SPIE, 2019.

[11] H. Yang, X. Huang, and Y. Liu, "Infogan-enhanced underwater acoustic target recognition method based on deep learning," in *Proceedings of 2022 International Conference on Autonomous Unmanned Systems (ICAUS 2022)* (W. Fu, M. Gu, and Y. Niu, eds.), (Singapore), pp. 2705–2714, Springer Nature Singapore, 2023.

[12] Hina Ashraf, Babar Shah, Afaque Manzoor Soomro, Qurat-ul-Ain Safdar, Zahid Halim, and Said Khalid Shah, "Ambient-noise free generation of clean underwater ship engine audios from hydrophones using generative adversarial networks," *Computers and Electrical Engineering*, vol. 100, p. 107970, 2022.

[13] L. Atanackovic, V. Vakilian, D. Wiebe, L. Lampe, and R. Diamant, "Stochastic ship-radiated noise modelling via generative adversarial networks," pp. 1–8, 2020.

[14] Y. Li, F.-X. Ge, Y. Bai, and M. Li, "Pseudo ship-radiated noise generation based on adversarial learning," in *2021 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pp. 222–227, 2021.

[15] J. Qiu, W. Liu, Z. Chen, H. Liu, and J. Liu, "Research on autoregressive ship-radiated noise synthesis technology based on prompt signal constraints," pp. 1732–1737, 2024.

[16] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *CoRR*, vol. abs/1703.06868, 2017.

[17] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images."

[18] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer."

[19] C.-J. Chang, L. Zhao, S. Zhang, and M. Kapadia, "Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis," *Computer Animation and Virtual Worlds*, vol. 33, no. 3-4, 2022.

[20] K. Ko, D. kim, K. Oh, and H. Ko, *WaveVC: Speech and Fundamental Frequency Consistent Raw Audio Voice Conversion*. 2023.

[21] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *CoRR*, vol. abs/1711.00937, 2017.

[22] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," 2020.

[23] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," *CoRR*, vol. abs/1906.00446, 2019.

[24] C. Zhu, T. Gaggero, N. C. Makris, and P. Ratilal, "Underwater sound characteristics of a ship with controllable pitch propeller," *Journal of Marine Science and Engineering*, vol. 10, no. 3, p. 328, 2022.