POLITECNICO DI TORINO
Repository ISTITUZIONALE

A GAN-based framework for generating STFT spectrograms of rare acoustic events in structural health monitoring

Original
A GAN-based framework for generating STFT spectrograms of rare acoustic events in structural health monitoring / Farhadi, S., Iavarone, M., Corrado, M., Chatzi, E., Ventura, G.. - In: DATA-CENTRIC ENGINEERING. - ISSN 2632-6736. - 7:(2026), pp. 1-22. [10.1017/dce.2026.10071]

Availability:
This version is available at: 11583/3015991 since: 2026-09-27T17:08:28Z

Publisher:
Cambridge University Press

Published
DOI:10.1017/dce.2026.10071

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

23 October 2026

Data-Centric Engineering (2026), 7: e42
doi:10.1017/dce.2026.10071

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

A GAN-based framework for generating STFT spectrograms of rare acoustic events in structural health monitoring

Sasan Farhadi[1,2], Mariateresa Iavarone[2], Mauro Corrado[2], Eleni Chatzi[3] and Giulio Ventura[2]

[1]Institute of Engineering Geodesy and Measurement Systems, Graz University of Technology, Graz, Austria
[2]Politecnico di Torino, Italy
[3]ETH Zurich, Switzerland
Corresponding author: Sasan Farhadi; Email: sasan.farhadi@tugraz.at

Received: 17 December 2025 UTC; **Revised:** 02 June 2026 UTC; **Accepted:** 18 June 2026 UTC

Keywords: class imbalance; data augmentation; deep learning; generative adversarial networks; short-time Fourier transform; sound event detection

Abstract

Structural health monitoring plays a crucial role in ensuring the safety, reliability, and longevity of bridge infrastructures through early damage detection. Although recent advances in deep learning-based models have enabled automated event detection, their performance is often limited by data scarcity, environmental noise, and class imbalance. To address these challenges, this study introduces a customized generative adversarial network (GAN) model, STFTSynth, designed particularly for generating short-time Fourier transform spectrograms derived from acoustic-event signals. In contrast to augmentation techniques such as MixUp, GANs can synthesize spectrograms that visually and statistically resemble real event representations, providing a basis for representation-level dataset enrichment. The proposed model integrates dense residual blocks for spatial consistency with bidirectional gated recurrent units for temporal dependency modeling. Model performance is evaluated against three baseline generative models using qualitative inspection and quantitative metrics, including structural similarity index measure, peak signal-to-noise ratio, and Fréchet inception distance. Results show that STFTSynth outperforms baseline models, producing high-resolution, temporally consistent spectrograms that align closely with real-world data. These findings highlight the potential of GAN-based spectrogram synthesis for representation-level enrichment of rare acoustic-event datasets in bridge monitoring, particularly when real examples such as prestressing wire breakage are limited.

Impact Statement

Bridges are critical infrastructure systems, and rare damage-related acoustic events, such as prestressing wire breakage, are difficult to capture in sufficient quantity under real monitoring conditions. This scarcity limits the availability of representative data for developing and evaluating data-driven structural health monitoring (SHM) methods. This study investigates generative adversarial network (GAN)-based synthesis of short-time Fourier transform spectrograms as a representation-level strategy for enriching limited acoustic-event datasets. By generating synthetic spectrograms that visually and statistically resemble real event representations, the proposed framework provides a basis for studying rare-event data generation in bridge monitoring scenarios. The results demonstrate improved spectrogram fidelity and distributional similarity compared with selected GAN baselines. However, the generated samples are evaluated only at the representation level, and their effect on downstream SHM tasks, such as event classification, damage detection, or anomaly detection, requires further validation in future work.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

1. Introduction

Bridges are vital for supporting economic growth and enabling efficient transportation. However, their structural integrity is continually challenged by factors such as aging, traffic loads, and environmental conditions (Ko and Ni, 2005). Over time, these influences can lead to material deterioration, making early damage detection through structural health monitoring (SHM) critical for maintenance planning and ensuring structural safety. One severe form of internal deterioration in post-tensioned prestressed concrete bridges is the corrosion of the strands, which can lead to a progressive wire breakage and ultimately lead to catastrophic collapse of the structure, if not detected (Zingoni, 2019; Ferro et al., 2022). These incidents underscore the urgent need for proactive monitoring and timely interventions to maintain infrastructure safety and reliability.

Conventional SHM methods, notably visual inspection, remain mainly manual, expensive, and ineffective for detecting internal or subtle structural damage at early stages. In addition, advanced sensor-based methods often produce complex and noisy data, which complicates manual analysis and limits timely decision making (Webb et al., 2017; Saleem et al., 2021). In tackling these challenges, recent attention has turned toward automated processing techniques utilizing machine learning and deep learning (DL), which can efficiently interpret large datasets and directly identify structural anomalies from raw data (Zinno et al., 2022; Farhadi et al., 2024b). Convolutional neural networks (CNNs), for instance, have been widely applied to analyze vibration signals, indicating their efficiency in detecting structural anomalies (Parziale et al., 2024; Duran et al., 2025; Zhou et al., 2025). Similar approaches have been applied to detect damage in mechanical and structural components (Azad and Kim, 2024; Guo et al., 2024; Wang et al., 2025).

In recent studies (Farhadi et al., 2024a, 2024b, 2024c), the authors investigated the downstream task of sound event detection for classifying damage events, with particular focus on prestressing wire breakage —a critical yet rare and extremely short-lived occurrence in SHM. A comprehensive benchmarking of several deep learning models, including modified pre-trained models and a customized model (AcousticNet), confirmed the superior performance of short-time Fourier transform (STFT)-based representations compared to other time–frequency representations. Nevertheless, these studies also underscored persistent challenges: severe data scarcity and class imbalance, both of which hindered model generalization and robustness. Conventional augmentation strategies such as MixUp were explored as partial remedies, but their inherent limitations became clear. In particular, MixUp generates synthetic samples by linear interpolation, often producing unrealistic or ambiguous signals, especially for transient anomalies of very short duration. Consequently, such techniques proved insufficient to address the constraints posed by small, imbalanced, and noisy SHM datasets. This gap motivated the present work: the development of a generative framework for synthesizing realistic STFT representations of underrepresented event classes.

To address these challenges, the present study proposes a customized generative model designed to end-to-end synthesize realistic single-channel STFT spectrograms of bridge monitoring events. The model is based on the Wasserstein generative adversarial network with gradient penalty (WGAN-GP) architecture to ensure stable training and high-quality generation. Training is performed in an event-separated but unsupervised manner, meaning that although each GAN is trained on a specific event type (e.g., wire breakage or traffic), no class labels or downstream classifiers are involved. This setup enables the model to learn the spectral structure of each event class independently, enriching the dataset for rare classes without relying on manual annotation.

Unlike previous approaches that use three-channel (RGB) spectrograms, this study employs single-channel STFT inputs, simplifying the model and preserving the native structure of vibration data. The main contributions are as follows: (1) acquisition of a rare, high-quality dataset from controlled wire breakage experiments on a real post-tensioned concrete bridge in Italy slated for demolition, enabling safe, repeatable testing under realistic structural conditions; (2) development of STFTSynth, a WGAN-GP-based generative model tailored to synthesize realistic single-channel STFT spectrograms of various events; and (3) comprehensive benchmarking of STFTSynth against baseline models (deep convolutional

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

GAN [DCGAN], standard WGAN-GP, and least-squares GAN [LSGAN]) using both visual inspection and quantitative metrics. This work demonstrates the potential of GAN-based data generation for representation-level synthesis of rare SHM acoustic-event spectrograms. The contribution is focused on synthetic STFT spectrogram generation and representation-level benchmarking, while the effect of the generated samples on downstream SHM tasks remains a direction for future work.

2. Background

Despite the significant potential of DL methods in SHM, several challenges limit their generalization and deployment in real-world scenarios. A major issue is data scarcity, as rare but critical events such as prestressing wire breakage are extremely difficult and costly to capture in sufficient quantity for robust training (Chen et al., 2024; Farhadi et al., 2024a). Furthermore, class imbalance, due to the dominance of normal structural states, can bias model performance and reduce sensitivity to damage (Gao et al., 2021). Real-world datasets are often affected by environmental noise, such as traffic or construction vibrations, which complicate anomaly detection and diminish signal clarity (Bull et al., 2021; Tran et al., 2021; Xie et al., 2022). Although modern SHM methods increasingly support real-time and cloud-based monitoring, their underlying models are typically trained offline and require large, labeled datasets. These conditions make supervised DL difficult to scale for critical but infrequent events. To overcome these limitations, generative modeling has appeared as a promising strategy for synthesizing additional representations of underrepresented events, such as wire breakage. However, the practical value of such synthetic samples for SHM deployment requires downstream validation.

GANs (Goodfellow et al., 2014) have been widely adopted for synthetic data generation in audio and vibration domains. Early waveform-based approaches (e.g., WaveGAN [Donahue et al., 2018]) demonstrated the feasibility of adversarial learning for temporal signals but often suffered from training instability and limited spectral coherence, particularly for complex structural monitoring data. Consequently, later studies shifted toward spectrogram-based GANs, where time–frequency representations such as STFT enable more stable training and improved representation of spectral content (e.g., SpecGAN [Donahue et al., 2018], GANSynth [Engel et al., 2019], MelGAN [Kumar et al., 2019], EVA-GAN [Liao et al., 2024]). These works collectively support the use of STFT-domain representations as an effective and practical basis for adversarial audio generation.

Recent work has begun exploring the potential of GANs in SHM applications, particularly for generating structural monitoring signals. Luleci et al. (2023) proposed a one-dimensional Wasserstein deep convolutional GAN with gradient penalty (1D WDCGAN-GP) to generate synthetic acceleration time series for damage detection. Their study demonstrated that integrating synthetic samples with real datasets improved the performance of downstream classifiers, particularly in addressing data imbalance. These efforts confirmed the viability of GAN-based augmentation for SHM but remained limited to waveform-level generation. Similarly, Tran et al. (2021) explored the use of GANs to generate vibration signals under varying operational conditions, highlighting GANs' potential to model complex data distributions in noisy SHM setting. Beyond waveform data, GANs have also been employed for structural defect simulation. Li et al. (2024) applied CycleGAN to generate crack images on bridge surfaces, aiding in image-based defect classification. In related work, Le et al. (2020) used WGANs to generate surface defect patterns on metallic structures. Although these image-based approaches are promising, they do not address the spectral and temporal complexity inherent in time-series-based (vibration-based) SHM. Besides, in parallel to data-driven approaches, physics-informed GAN have indicated potential for sound and demonstrating improved interpretability and fidelity (Ren et al., 2013; Su et al., 2023). These studies confirm the application of GANs for structural data augmentation; however, very few works have investigated GAN-based generation of time–frequency representations, particularly single-channel STFT spectrograms, in the context of SHM. This gap is significant, as spectrograms provide a compact, interpretable, and structure-preserving way to represent acoustic signals, especially for rare and short-lived events like prestressing wire breakage. The current study

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

addresses this gap by introducing a GAN-based framework specifically designed to synthesize realistic STFT spectrograms of structural events.

Moreover, several studies have explored generative modeling in SHM context, primarily with the aim of mitigating data scarcity and class imbalance. Existing SHM-focused works have mainly employed GAN-based approaches for waveform-level synthesis of vibration or acceleration signals, including labeled data augmentation and signal reconstruction tasks (Hou et al., 2022; Luleci et al., 2022; Mousavi et al., 2025). Related efforts have also explored GAN-based augmentation of acoustic emission signals for damage-related event analysis (Fu et al., 2024). Although these studies demonstrate the potential of generative models in SHM, they mainly operate in the time domain and do not address the generation of time–frequency representations. In contrast, the present study focuses on the generation of single-channel STFT spectrograms for rare acoustic events, addressing a gap in the current SHM literature where time–frequency-based generative augmentation remains largely unexplored.

3. Methods

3.1. Data preparation

Starting from the continuous recordings, single acoustic events with a duration of 0.0104 s were extracted. Considering the acquisition rate of 96 kHz, each signal consists of 1000 samples. The length of the signals was chosen based on experimental observations and numerical modeling, which indicate that the duration of a breakage event is approximately 0.01 s. As discussed earlier, high-quality labeled data are critical for enhancing model robustness. In this study, the dataset comprises four signal classes: wire breakage (202 signals), hammering (264 signals), electric trimmer (459 signals), and traffic noise (415 signals). The distribution highlights a class imbalance, with critical events such as wire breakage being underrepresented. This imbalance can adversely affect downstream classification model performance by biasing predictions toward the more frequent classes. Figure 1 provides an example of the representations in the time domain of each signal class, highlighting the distinctive patterns captured during the acquisition process. To ensure uniformity and meaningful comparison, all signals were normalized before further processing. This preprocessing step reduces the bias from varying amplitude ranges and can simplify model training. The next step is to extract features that can capture necessary information for model training, as discussed in the next section.

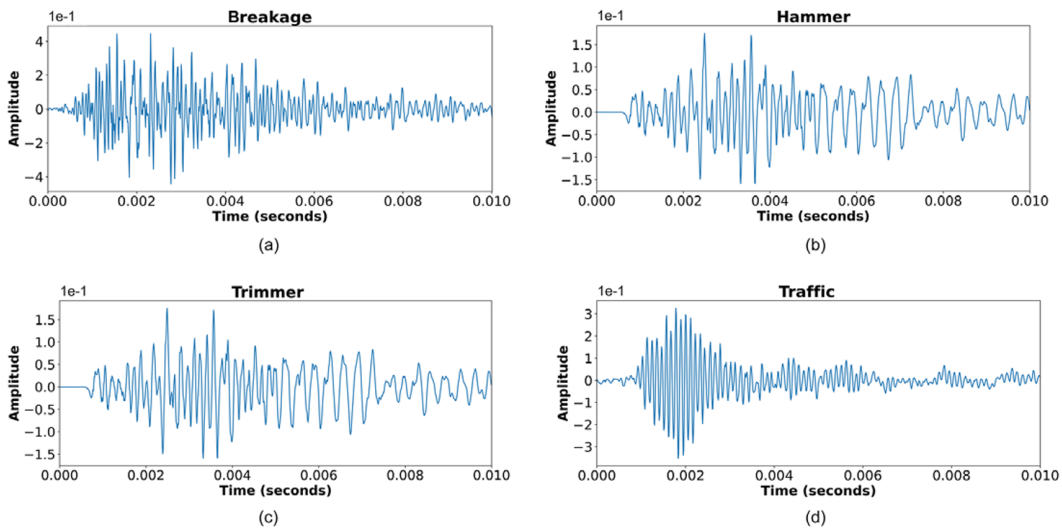

Figure 1. *Time-domain representation of the recorded acoustic events.*

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

3.2. Feature extraction

STFT was employed as the primary feature extraction method for in-situ acoustic signals, transforming them from the time domain into the time–frequency domain to capture both spectral and temporal characteristics. This choice aligns with prior work by the authors and supports dynamic signal representation for GAN-based data generation. All STFT computations employed a Hann window function with 50% overlap. This choice was supported by preliminary tests, which indicated that Hann windows produced stable and informative spectrograms for wire breakage signals, while a 50% overlap maintained sufficient temporal resolution without introducing excessive redundancy.

Window size is a key STFT parameter, directly affecting the time–frequency resolution trade-off (**Gabor limit**). Figure 2 compares spectrograms for four window sizes: small windows (e.g., 64 samples) yield high temporal resolution but poor frequency resolution, while large windows (e.g., 512 samples) achieve the opposite. To balance these effects, window sizes of 128 and 256 were selected, offering both spectral clarity and temporal precision. The resulting spectrogram dimensions, expressed as (frequency bins, time windows), were $(65, 15)$ and $(129, 7)$, respectively. All STFT outputs were treated as single-channel images during training to reduce computational complexity.

3.3. Sequence modeling: Gated recurrent unit

Sequential modeling was initially developed to address the limitations inherent in CNNs, particularly their inability to capture long-term dependencies in sequential data. Early models include Jordan networks (Jordan, 1986) and Elman networks (Elman, 1990), both foundational forms of recurrent neural networks (RNNs) designed to process sequential information. Among the more advanced architectures, gated recurrent units (GRUs) (Cho et al., 2014) were introduced as a computationally efficient alternative to long short-term memory (LSTM) networks. GRUs incorporate gating mechanisms that effectively preserve long-range dependencies and mitigate the vanishing gradient problem. By combining the roles of LSTM's input and forget gates into a single update gate, GRUs reduce model complexity while maintaining comparable performance for various sequence modeling tasks (Figure 3).

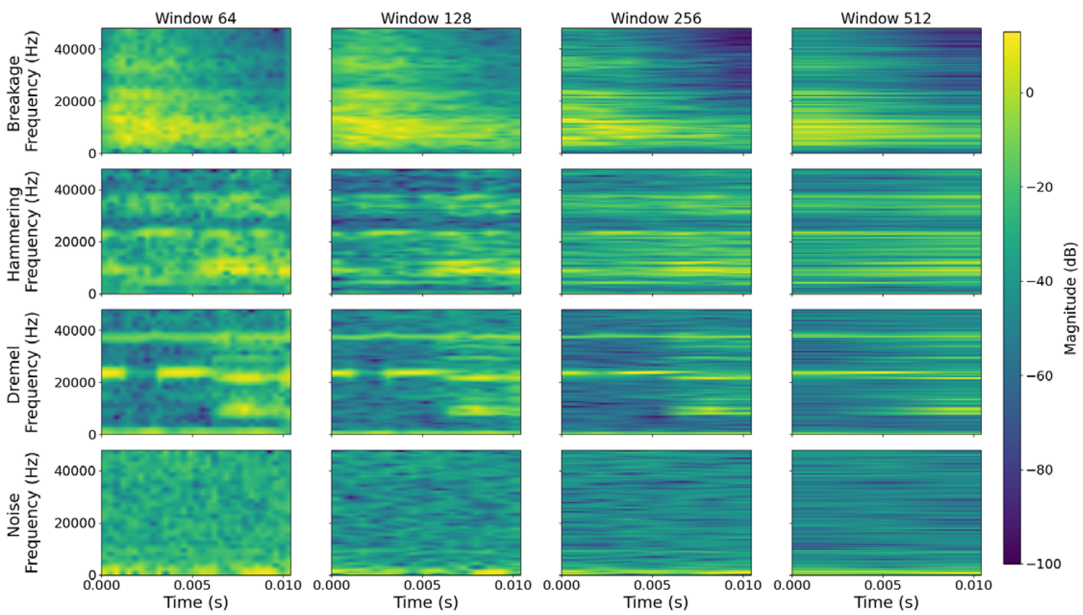

Figure 2. *STFT spectrograms of events across four window sizes, illustrating the impact of resolution on signal representation and feature extraction.*

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

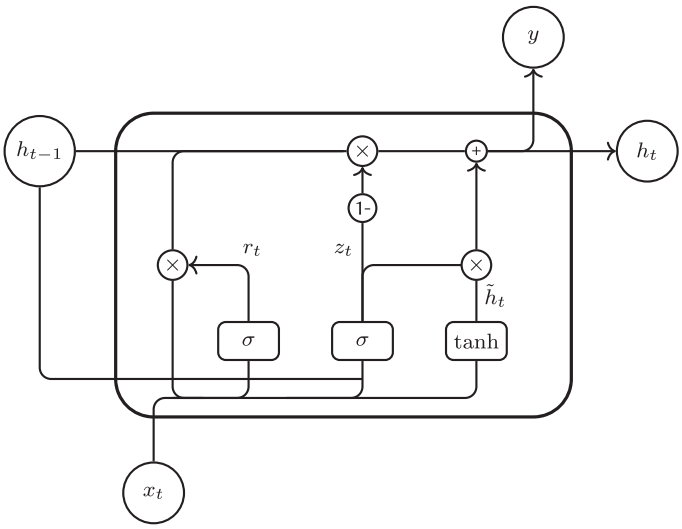

Figure 3. *Architecture of a gated recurrent unit (GRU).*

At each time step t, the hidden state $h(t)$ is updated using a linear interpolation between the previous hidden state $h(t-1)$ and the candidate activation $\tilde{h}(t)$, weighted by the update gate $z(t)$. The updated equation is given as (Cho et al., 2014; Chung et al., 2014)

$$h(t) = (1 - z(t)) \otimes h(t-1) + z(t) \otimes \tilde{h}(t) \tag{1}$$

where $\otimes$ represents element-wise multiplication, and $z(t)$, known as the update gate, determines the extent to which the hidden state should be updated with new information. The update gate is computed as

$$z(t) = \sigma(w_z \cdot x(t) + w_h \cdot h(t-1) + b(z)) \tag{2}$$

where w_z and w_h are the weight matrices for the input and hidden state, respectively, and b_z is the bias vector. The update gate in GRU simultaneously performs the role of the input and output gates in LSTM. The sigmoid function, σ, constrains $z(t)$ to values between 0 and 1, determining the proportion of $h(t-1)$ that should be preserved versus the amount of $\tilde{h}(t)$ (candidate activation) to incorporate. This update gate effectively regulates the flow of information from the previous hidden state to the current one. The candidate activation $\tilde{h}(t)$ is computed similarly to the hidden state in a conventional RNN, with an additional modulation by the reset gate $r(t)$ as follows:

$$\tilde{h}(t) = \tanh\left(w_x \cdot x(t) + w_{\tilde{h}} \cdot (r(t) \otimes h(t-1)) + b_{\tilde{h}}\right) \tag{3}$$

where w_x and $w_{\tilde{h}}$are weight matrices for the input and the reset-modulated hidden state, respectively. $r(t)$ is known as the *reset gate*, determines how much of the previous hidden state $h(t-1)$ can be retained when calculating the candidate activation. The reset gate is computed as

$$r(t) = \sigma(w_x \cdot x(t) + w_h \cdot h(t-1) + b_r) \tag{4}$$

where w_x and w_h are the weights for the input and hidden state in the reset gate, and b_r is the corresponding bias vector.

3.4. Generative adversarial networks

GANs, first introduced by Goodfellow et al. (2014), are generative models that learn to produce data resembling a target distribution through an adversarial process. A GAN consists of two neural networks: the generator, which maps random latent vectors ($z \sim p_z(z)$) to synthetic samples ($\tilde{x}$) and the discriminator

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

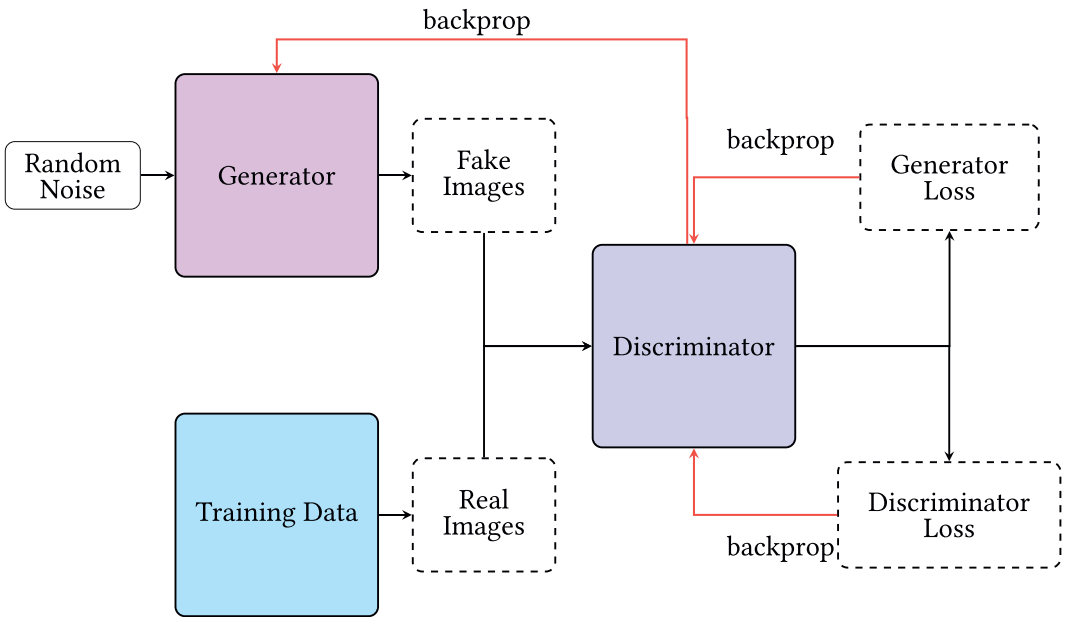

Figure 4. *Overview of GAN workflow.*

(or critic) which distinguish real samples ($x \sim p_{data}$) from the generated ones (Figure 4). The two networks are trained in a minimax game (Xia et al., 2022; Gui et al., 2023):

$$\min_{G} \max_{D} \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{5}$$

where $D(x) \in (0,1)$ is the discriminator's estimated probability that x is real, p_{data} is the distribution of the real dataset, and $p_z(z)$ is the prior distribution over latent vectors (often Gaussian or uniform). The generator G aims to minimize this objective by producing realistic samples that maximize $D(G(z))$, while the discriminator D aims to maximize it by correctly distinguishing real from generated data.

In practice, vanilla GANs often suffer from unstable training and mode collapse. Variants such as DCGAN (Radford et al., 2016) introduced convolutional architectures, LSGAN (Mao et al., 2017) replaced the cross-entropy loss with a least-squares loss to improve gradient quality, and WGAN (Arjovsky et al., 2017) employed the Wasserstein distance for more stable optimization. The WGAN-GP (Gulrajani et al., 2017) further improved stability by enforcing a soft 1-Lipschitz constraint via a gradient penalty term:

$$\mathcal{L}_D = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \lambda \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2\right] \tag{6}$$

where $\mathbb{P}_r$ and $\mathbb{P}_g$ are the real and generated distributions, $\hat{x}$ is an interpolation between real and fake samples, and λ is the penalty coefficient. WGAN-GP produces smoother gradients and reduces mode collapse (Bishop and Bishop, 2024), making it suitable for high-dimensional data generation.

3.5. Proposed GAN framework

Convolutional GANs such as DCGAN and LSGAN are effective at modeling local spatial structure in spectrograms. However, without explicit sequential modeling, they may not fully capture long-range temporal dependencies across frames, particularly when events span multiple STFT windows. Sequence-aware components can complement convolutional layers to improve temporal consistency. To address this limitation, this research proposed STFTSynth architecture which integrates convolutional residual

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

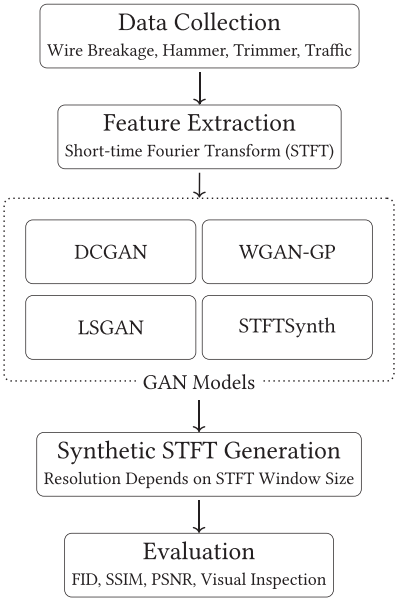

Figure 5. *Overview of the proposed GAN-based data augmentation workflow using STFT spectrograms.*

blocks for spatial feature extraction with bidirectional gated recurrent units (BiGRUs) to explicitly model temporal dependencies in both forward and backward directions. The residual blocks enhance gradient flow and enable multi-scale feature learning. On the other hand, BiGRUs capture the evolution of spectral patterns over time, ensuring coherent temporal transitions between adjacent STFT frames.

STFTSynth is implemented within the WGAN-GP framework to utilize its stable training dynamics and improved gradient behavior. The generator combines dense residual connections with BiGRUs to produce spectrograms that contain realistic local spectral texture and consistent temporal structure. The discriminator employs spectral normalization and multi-scale feature extraction to improve robustness against overfitting and mode collapse. This hybrid design enables STFTSynth to generate high-fidelity, temporally consistent STFT spectrograms for rare SHM events, such as prestressing wire breakage, that are difficult to capture in sufficient quantity for supervised training.

The overall workflow of the proposed GAN-based STFT data generation is illustrated in Figure 5. It outlines the progression from raw event collection to feature extraction, GAN-based spectrogram synthesis, and evaluation using quantitative and qualitative metrics. Implementation details, training procedures, and hyperparameter settings of the proposed architecture are provided in Section 3.6

3.6. Model developments and training procedure

This section presents the development, implementation, and training procedures for four GAN models: DCGAN, WGAN-GP, LSGAN, and the proposed STFTSynth. The baseline models were selected due to their widespread adoption and established performance in image generation tasks, providing a reliable foundation for comparison. However, architectures such as DCGAN are known to suffer from issues including mode collapse and training instability, particularly when applied to small and imbalanced datasets. To address these limitations, WGAN-GP and LSGAN introduce alternative loss formulations that improve training stability and enhance sample diversity. Despite these improvements, the baseline models still exhibit limitations in capturing long-range temporal dependencies present in STFT spectrograms. This limitation motivates the development of STFTSynth, a task-specific architecture designed to improve both temporal consistency and spectral realism. Highly complex architectures such as StyleGAN were not considered due to their significant computational requirements and dependence on large-scale,

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

diverse datasets, which are not available in this study. Additionally, hyperparameter tuning was performed to further mitigate issues related to mode collapse and overfitting. All models were trained using a dataset collected from in-situ bridge monitoring.

The implementation environment and computational setup are described as follows. Training experiments were conducted on a workstation equipped with an NVIDIA RTX A6000 GPU (48 GB memory). GPU acceleration enabled the customized models to complete each training epoch in approximately 15 seconds on average, depending on model complexity and batch size. In contrast, the selected baseline models required less than 2 seconds per epoch. All models were implemented using Keras with a TensorFlow 2.10 backend, within a Python 3.9 environment and its associated scientific computing libraries. The workstation operated on Windows Server 2019, with GPU performance optimized using CUDA Toolkit 11.2 and cuDNN 8.0.

3.6.1. DCGAN

The DCGAN implementation followed Radford et al. (2016), adapted for single-channel STFT spectrograms (size = H × W). The generator starts from a 100-dimensional latent vector, followed by four transposed convolutional layers (kernel = 4, stride = 2, padding = 1) with batch normalization and Leaky ReLU activations, producing the target resolution. The discriminator mirrors this structure in reverse using strided 2D convolutions with spectral normalization and leaky ReLU activations, ending with a sigmoid output for real/fake classification. All optimizer settings, initialization schemes, and other training hyperparameters are summarized in Table 1.

3.6.2. WGAN-GP

The WGAN-GP model retained the same convolutional generator–discriminator architecture as DCGAN for a fair comparison but replaced the BCE loss with the Wasserstein loss and added a gradient penalty ($\lambda = 12$) to enforce the 1-Lipschitz constraint. The discriminator (critic) outputs real-valued scores rather than probabilities, and spectral normalization was used to improve stability. Both generator and critic employed Leaky ReLU activations in all hidden layers, with Tanh at the generator output. Training followed the optimizer and regularization settings listed in Table 1.

3.6.3. LSGAN

The LSGAN architecture also matched DCGAN in both generator and discriminator structure but replaced the BCE loss with a least-squares loss to improve gradient quality and reduce saturation. The discriminator minimized the mean squared error between real samples and label 1, and fake samples and label 0. Leaky ReLU activations were used in all hidden layers, with Tanh at the generator output. All other training settings followed Table 1.

3.6.4. STFTSynth

STFTSynth extends the WGAN-GP architecture with dense residual blocks (DRBs) and BiGRU layers to address the strong temporal dependencies in STFT spectrograms of SHM events. DRBs are inserted after upsampling stages in the generator to enhance gradient flow, encourage feature reuse, and integrate multi-scale spectral details. BiGRUs are positioned after intermediate feature maps to model temporal evolution in both forward and backward directions, improving coherence across consecutive STFT frames. The discriminator employs spectral normalization and multi-scale convolutional layers to improve robustness against overfitting and mode collapse. Training follows the WGAN-GP scheme with $\lambda = 12$ and matched optimizer settings. A full architectural specification is given in Table 2, and a schematic is shown in Figure 6.

The dataset used for training consists of four event classes: wire breakage (202 samples), hammering (264 samples), electric trimmer (459 samples), and traffic noise (415 samples). For each class, a small portion of samples (approximately 15%) was held out as a validation set and used exclusively for evaluation. Each GAN model was trained under multiple epoch settings (1000–5000) to evaluate training

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

Table 1. *Summary of common training hyperparameters and configurations for GAN variant*

Parameter	Optimal value	Search space
General training parameters		
Learning rate (generator)	2e−5	5e−7 to 1e−3
Learning rate (discriminator)	2e−6	5e−7 to 1e−4
Learning rate adjustments	Factor: 0.15, Patience: 5	Factor: 0.1–0.5, Patience: 3–10
Batch size	16	16–64
Epochs	Max 2000	1000–5000
Optimizer	Nadam (Adam with Nesterov momentum)	Adam, Nadam, RMSprop
Weight initialization	Xavier	Xavier and He
Latent space dimension Z	100	50–500
Regularization	Batch normalization, dropout rate: 0.2	Dropout: 0.1–0.4
GAN-specific parameters		
Loss function (DCGAN)	Binary cross-entropy (BCE)	BCE
Loss function (LSGAN)	Least-squares loss	Least squares, Huber loss
Loss function (WGAN-GP)	Wasserstein loss with gradient penalty	Wasserstein loss
Gradient penalty λ_{GP} (WGAN-GP)	12	1–20
Gradient clipping (if applicable)	Not used	0.01–0.1
Activation function (generator)	Leaky ReLU (all layers except last), Tanh	ReLU, Leaky ReLU, ELU, Tanh
Activation function (discriminator)	Leaky ReLU (all layers except last), sigmoid	ReLU, Leaky ReLU, ELU, Sigmoid
Generator layers	4 Transposed convolutions	3–8
Discriminator layers	5 Convolutional layers	4–8
Kernel size (conv layers)	5 × 5	3 × 3 to 5 × 5
Stride (conv layers)	2	1–2

Table 2. *Comparison of architectural components across GAN variants*

Component	DCGAN	WGAN-GP	LSGAN	STFTSynth
Generator layers	4 ConvTrans-pose2D	4 ConvTrans-pose2D	4 ConvTrans-pose2D	4 ConvTrans-pose2D, 4 DRB and 2 BiGRU
Discriminator layers	5 Conv2D	5 Conv2D	5 Conv2D	5 Conv2D
Normalization	BatchNorm	BatchNorm	BatchNorm	BatchNorm
Activation (generator)	Leaky ReLU + Tanh	Leaky ReLU + Tanh	Leaky ReLU + Tanh	Leaky ReLU + Tanh
Activation (discriminator)	Leaky ReLU	Leaky ReLU	Leaky ReLU	Leaky ReLU
Loss function	BCE	Wasserstein + GP	Least squares	Wasserstein + GP

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

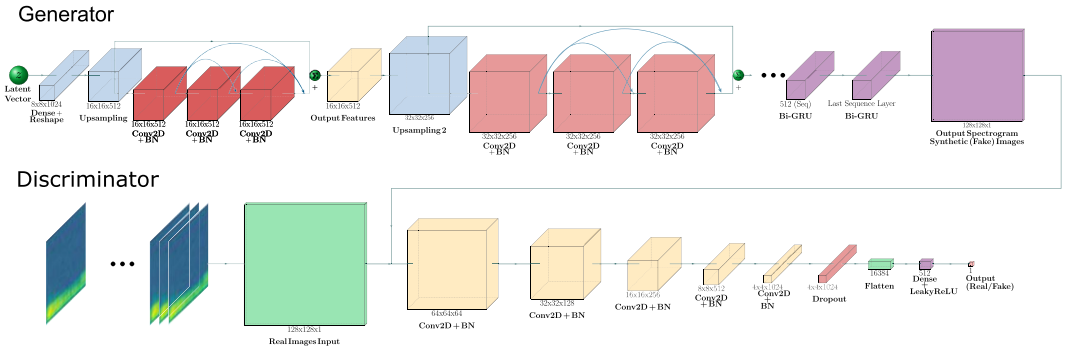

Figure 6. *Architecture of the proposed GAN model, STFTSynth.*

stability and sample quality. Although mode collapse occasionally occurred in early or mid-stages, extended training generally allowed the generators to recover and produce diverse, high-quality spectrograms. Based on comparative evaluations (Table 1), 2000 epochs were selected as the optimal configuration, offering a balance between stability, fidelity, and computational cost.

3.7. Evaluation metrics

3.7.1. Structural similarity index measure

The structural similarity index measure (SSIM) is a perception-based image quality metric that evaluates the similarity between two images by considering structural information (Li et al., 2018). In contrast to conventional metrics such as mean square error (MSE), SSIM accounts for the inter-dependencies of neighboring pixels, aligning more closely with human visual perception. SSIM consists of three components, including luminance $I(x,y)$, contrast $C(x,y)$, and structure $S(x,y)$:

$$\mathrm{SSIM}(x,y) = I(x,y) \cdot C(x,y) \cdot S(x,y) \tag{7}$$

In Equation (7) (Zhou Wang et al., 2004):

$$I(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}$$

$$C(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}$$

$$S(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

Here, x and y are image patches extracted from real and generated images, respectively. μ_x and μ_y are the mean intensity values. σ_x and σ_y are the standard deviations of the pixel values that represent the contrast. σ_{xy} is the covariance between the corresponding pixel values that capture structural similarity. c_1, c_2, and c_3 are small constants to avoid division by zero (Zhou Wang et al., 2004). SSIM score ranges between 0 and 1, where a value closer to 1 means more similarity to the real images.

3.7.2. Peak signal-to-noise ratio

The peak signal-to-noise ratio (PSNR) measures the similarity between two monochrome images to assess the quality of a generated image compared to a real one (Smith, 1997). A higher value of PSNR indicates better image quality. PSNR is computed as follows:

$$\mathrm{PSNR}(\mathrm{R},\mathrm{G}) = 10\log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \tag{8}$$

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

In Equation (8), R and G refer to the real and generated images, respectively, and MAX_I is the maximum pixel value of the image and MSE can be computed as follows:

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{i=0}^{n-1}(R(m,n) - G(m,n))^2$$

3.7.3. Fréchet inception distance

The Fréchet inception distance (FID) was introduced by Heusel et al. (2017) to evaluate the performance of GAN models. It measures the difference between real and generated data distributions in a feature space extracted from a pre-trained model, such as Inception-v3. The FID score is computed as follows:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right) \tag{9}$$

In Equation (9), μ_r, Σ_r and μ_g, Σ_g are the mean and covariance matrices of real and generated image features, respectively. A lower FID score (typically less than 10 for high-quality images) indicates better image quality and diversity.

4. Data acquisition and problem setting

Developing an effective automated SHM system requires a robust data acquisition strategy to accurately capture vibration signals under real-world conditions. High-quality, well-labeled datasets are particularly critical for tasks such as synthetic data generation and structural event classification. This section describes the detailed procedures used for field data acquisition on post-tensioned concrete bridges, with the goal of providing a transparent and replicable experimental framework.

4.1. Acquisition setup

The experimental framework was conducted on the post-tensioned concrete bridge Cerqueta, located along the A24 highway in L'Aquila, Italy. As the structure was scheduled for demolition, it offered a rare opportunity to perform controlled breakage tests on its internal prestressing tendons (Figure 7). Cerqueta is a two-span bridge, each span consisting of four longitudinal prestressed reinforced concrete I-girders with a cast-inplace deck slab; the longitudinal girders are connected by cross beams. I-girders are prestressed by post-tensioned cables, placed in grouted corrugated metallic ducts. The prestressing system of the I-beams consisted of seven tendons with parabolic path having 12 strands with seven twisted wires each. A detailed summary of the structural and experimental parameters for the bridge is provided in Table 3.

These tests were made possible through a collaboration between Politecnico di Torino and Strada dei Parchi S.p.A., which allowed data collection during the partial deconstruction phase. Due to the considerable height and limited accessibility of the bridge, a mobile elevated work platform was utilized to access the side beams where instrumentation and testing were performed. The experimental procedure began by carefully removing sections of concrete in the web of the I-beams to open an operational window to access the internal steel ducts housing the prestressing tendons. In total, four windows were opened, two for each of east-facing side beams of the two spans (see Figure 8). Controlled wire breakage was performed by cutting the cross section of wires with an electric trimmer tool until generating a spontaneous tensile breaking. Considering the actual stress in the wires, they break when about half the cross section is cut.

4.2. Data acquisition

The acquisition system included two piezo-ceramic mono-axial accelerometers (Model 805 M1, DSPM Industria srl) having frequency response from 0.4 Hz to 12 kHz, and natural frequency of 34 kHz. Both accelerometers were mounted on the web of the side beam where the wires were cut, 4.5 m far from the

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

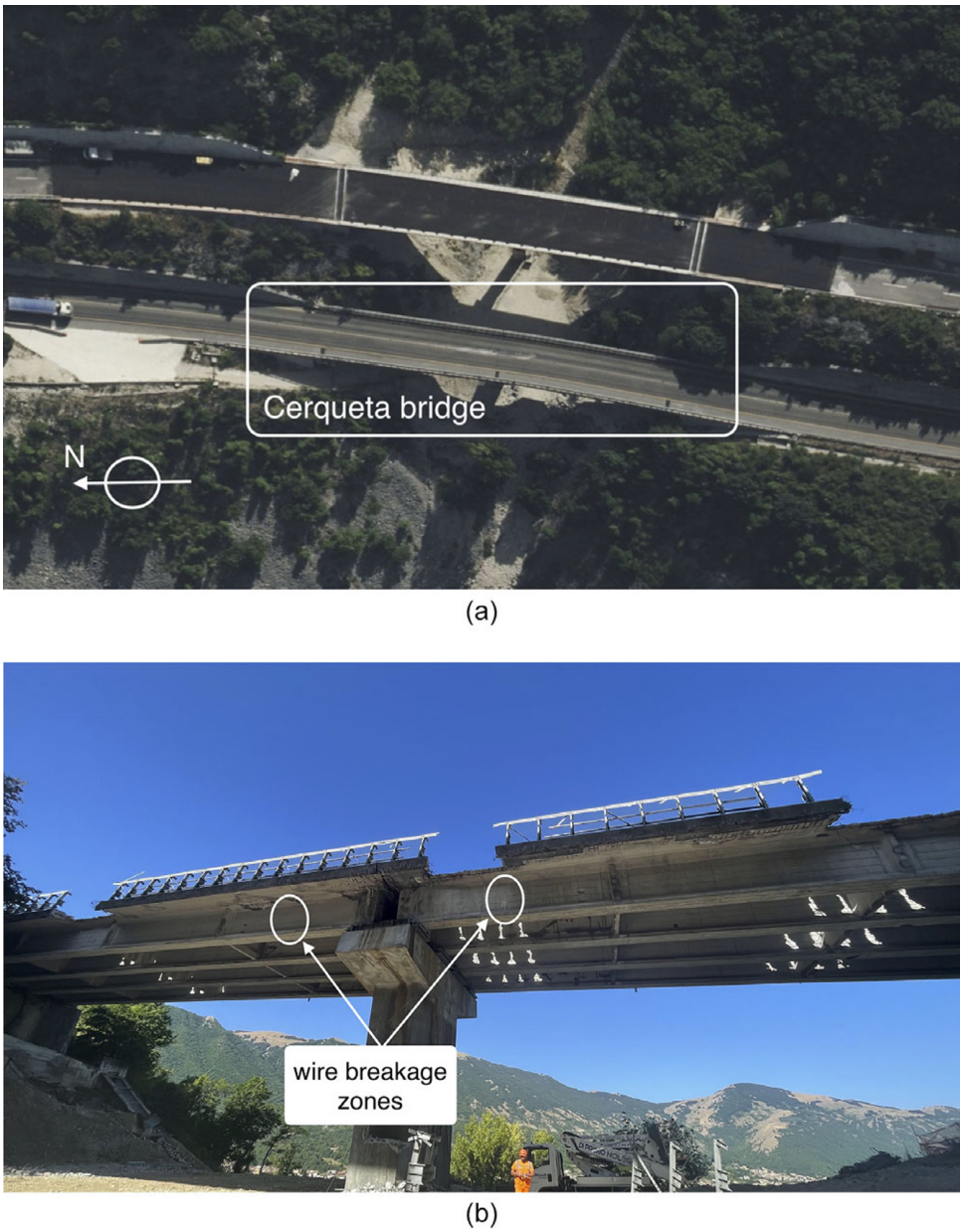

Figure 7. *Cerqueta bridge along the A24 highway selected for controlled prestressing tendon breakage tests: (a) plan view; (b) lateral view from East.*

cutting area, toward the midspan. The two accelerometers were installed to catch vibrations in the longitudinal direction and in the vertical direction, with respect to the bridge alignment (Figure 9). Although conventional structural monitoring typically targets low frequencies, up to about 50 Hz, to study the natural frequencies of the bridge, in our sound event detection-based approach, the accelerometers are used as microphones to capture the acoustic events propagating through the structure. That is the reason for the selected frequency range of the accelerometers. Consistently, an acquisition rate typical of sound recorders has been set, equal to 96 kHz. Besides a technical consistency of the devices with the developed

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

Table 3. *Summary of structural and experimental parameters for Cerqueta bridge*

Parameter	Cerqueta bridge
Location	A24 Highway, L'Aquila
Span length	38 m
Beam height	2.12 m
Orientation	North–South longitudinal axis
Test side	East-facing side beam
Height above ground	15 m
Number of tendons monitored	4 (12 strands each, with twisted wires)
Accelerometers	2 (4.5 m from cut area)
Sampling rate	96 kHz
Data acquisition mode	Dual-channel, synchronized

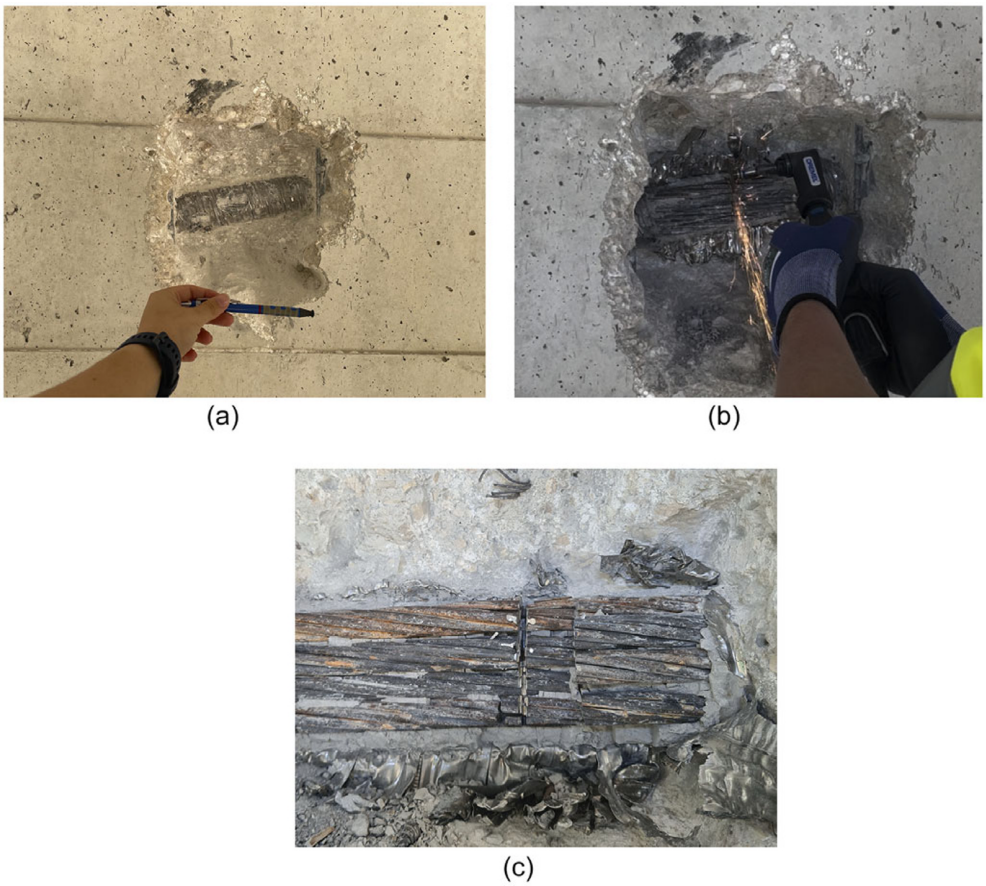

Figure 8. *Controlled wire breakage test on the Cerqueta bridge: (a) operational window opened in the beam's web; (b) wire cutting using an electric trimmer; (c) strands with twisted wires after breakage.*

approach, a key feature of the acquisition system is its simplicity and low cost. Data were recorded in continuous via a dual-channel acquisition system labeled as left and right channels, each capturing different orientations of acoustic signals. This setup provided a more detailed view of the events, enabling a richer analysis of time–frequency characteristics and signal propagation.

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

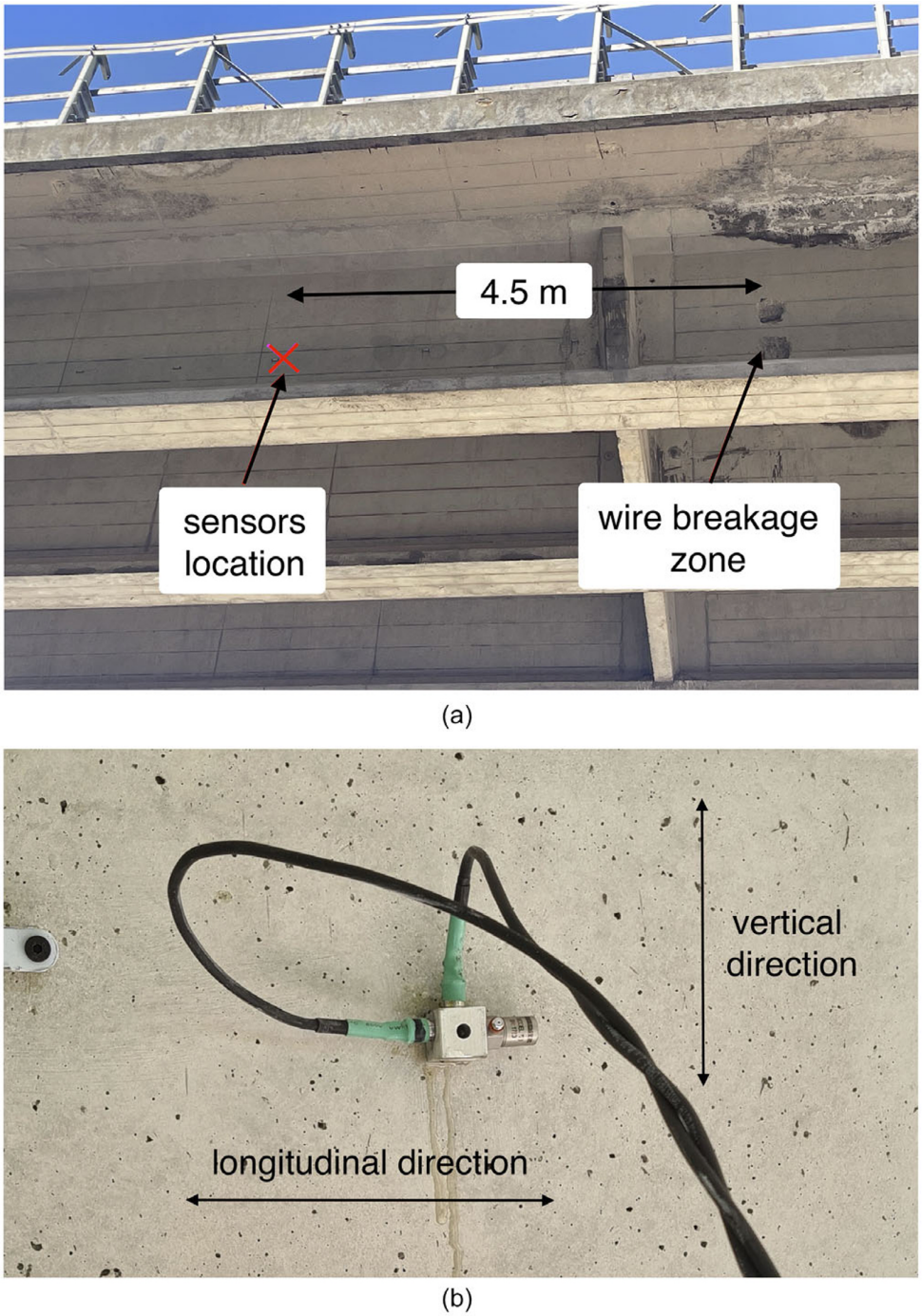

Figure 9. *Sensor positioning: (a) distance from the cutting areas; (b) orientation of the two accelerometers (green devices) with respect to the bridge alignment.*

In addition to wire breakage acoustic signals, two other classes of acoustic events were recorded on the Cerqueta bridge, namely electric trimmer operation and light hammering. Furthermore, to further enrich the dataset, traffic noise was recorded, using the same acquisition system, on a bridge of the same type as

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

Cerqueta. All in all, this experimental activity ensures that the dataset reflects a realistic scenario during structural monitoring of post-tensioned systems, enabling the development of more generalized and high-performing SHM systems. It is also important to highlight that acoustic data from each sensor were modeled independently, with each uniaxial signal treated as a separate input for the generative model. This decision was made to increase the number of input signals and its diversity for the models to avoid bias.

5. Results and discussion

5.1. Quantitative results

This section presents a quantitative evaluation of the proposed GAN models for generating STFT spectrograms. The objective is to assess how well the generated samples preserve the structural and statistical characteristics of real spectrograms. Three widely used metrics were adopted: structural similarity index (SSIM), PSNR, and FID. Table 4 summarizes the performance of all models across different event classes and STFT window sizes (128 and 256).

SSIM evaluates the perceptual similarity between generated and real STFT spectrograms, where higher values indicate better performance. The results in Table 4 show that STFTSynth outperforms all other models across all events and window sizes, achieving the highest SSIM scores. For instance, STFTSynth achieves an SSIM of 0.488 for the breakage event (window size 128) and 0.532 for the trimmer event (window size 256), both of which are the highest recorded values. This indicates that the proposed STFTSynth model generates STFT spectrograms with better similarity to real spectrograms, effectively preserving key spectral features. Contrarily, WGAN-GP exhibits the lowest SSIM scores across almost all event classes, particularly for the breakage event (window size 256), with a score of 0.180, and the traffic event (window size 128), with a score of 0.170. This highlights that WGAN-GP struggles to preserve the fine-grained structural details present in real STFT spectrograms.

PSNR measures reconstruction quality by quantifying the distortion between generated and real spectrograms, where higher values indicate lower distortion and improved clarity. Similar to SSIM, STFTSynth demonstrates the best overall performance, achieving the highest PSNR scores across all event classes and window sizes. The trimmer event obtained a score of 14.302, and the breakage event achieved a score of 14.009, both using a window size of 256. These results show that STFTSynth generates high-quality spectrograms with minimal distortion. In contrast, WGAN-GP consistently

Table 4. *Performance metrics of GAN variants for generating STFT spectrograms. Bold values indicate the best performance for each class and window size, with higher values preferred for SSIM and PSNR and lower values preferred for FID.*

		DCGAN		WGAN-GP		LSGAN		STFTSynth	
Metrics	Class	128	256	128	256	128	256	128	256
SSIM	Breakage	0.358	0.325	0.367	0.180	0.360	0.185	**0.488**	**0.459**
	Trimmer	0.336	0.393	0.342	0.403	0.347	0.404	**0.456**	**0.532**
	Hammer	0.222	0.225	0.208	0.232	0.212	0.224	**0.326**	**0.287**
	Traffic	0.189	0.264	0.170	0.186	0.176	0.184	**0.265**	**0.295**
PSNR	Breakage	13.348	13.451	13.409	11.972	13.523	12.249	**13.731**	**14.009**
	Trimmer	12.772	13.258	**13.424**	13.621	13.284	13.719	**13.422**	**14.302**
	Hammer	10.956	11.427	10.520	11.289	10.689	11.393	**11.338**	**11.934**
	Traffic	11.521	11.530	11.374	12.020	11.928	11.957	**12.422**	**12.232**
FID	Breakage	0.253	**0.229**	0.219	0.331	0.274	0.231	**0.210**	0.249
	Trimmer	0.251	0.242	**0.173**	0.270	0.186	**0.190**	0.219	0.235
	Hammer	0.209	0.177	0.221	0.194	0.237	0.241	**0.102**	**0.117**
	Traffic	0.209	0.279	**0.179**	0.237	0.198	**0.167**	0.249	0.246

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

produces lower PSNR values, particularly for the Hammer event using a window size of 128, obtaining a score of 10.520, which indicates higher noise levels and lower clarity in the generated STFT spectrograms. However, this model achieves comparable results to STFTSynth for the trimmer event using a window size of 128. The other two models, DCGAN and LSGAN, exhibit moderate performance compared to STFTSynth.

FID measures the similarity between real and generated data distributions by comparing their feature representations. Lower FID values indicate that generated spectrograms are more realistic and closely resemble real STFT data. As shown in Table 4, STFTSynth generally achieves the lowest FID scores, notably for the breakage event (0.210, window size 128) and hammer events (0.102 and 0.117, window sizes 128 and 256, respectively). These results confirm STFTSynth's capability to generate spectrograms closely aligned with real data distributions. However, STFTSynth's performance is not uniformly superior in all cases. For the breakage event at window size 256, DCGAN achieves a slightly lower FID (0.229) compared to STFTSynth's 0.249, demonstrating DCGAN's competitiveness in specific instances. Additionally, for the trimmer event at window size 256, LSGAN (0.190) outperforms STFTSynth (0.235), suggesting LSGAN's ability to better capture realistic STFT spectrogram features under certain conditions.

Moreover, WGAN-GP also shows competitive performance in some cases. Specifically, for the trimmer (0.173) and traffic (0.179) events at a window size of 128, WGAN-GP achieves the lowest FID, surpassing both LSGAN and STFTSynth. Despite this strength, WGAN-GP demonstrates inconsistent performance, with significantly higher FID scores for other events such as Breakage and Hammer. These variations suggest that WGAN-GP struggles to consistently capture high-quality spectrogram features across diverse event types. Lastly, DCGAN typically produces higher FID scores compared to the other evaluated models, indicating limitations in accurately modeling complex structural details and capturing realistic data distributions.

5.2. Qualitative analysis

In addition to the quantitative evaluation, a qualitative visual assessment of the generated STFT spectrograms was performed. The generated STFT spectrograms were visually compared against real ones to examine how accurately each model captures the key time–frequency characteristics of real events. Figures 10 and 11 illustrate the real and generated STFT spectrograms for wire breakage across the different GAN models, using STFT window sizes of 128 and 256, respectively. A visual inspection confirms that STFTSynth consistently produces the sharpest and most structurally detailed spectrograms, closely matching the real-world patterns. Among the baseline models, LSGAN generates spectrograms closest in quality to STFTSynth. In contrast, DCGAN and WGAN-GP show notable limitations, particularly in silent or less-structured regions, resulting in visible distortions or missing spectral details.

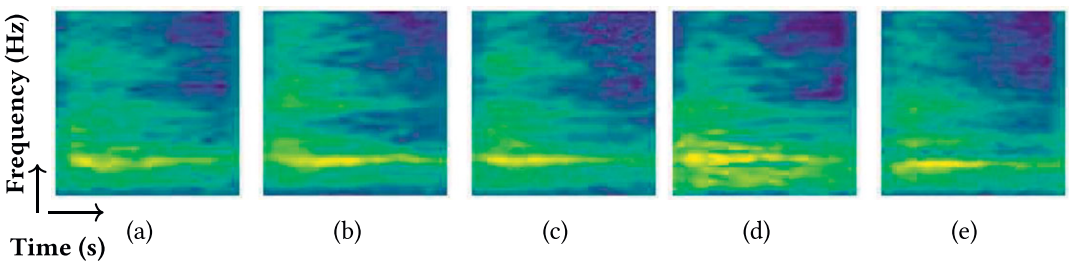

Figure 10. *STFT representations of real and generated wire breakage spectrograms (window size = 128): (a) Real STFT, (b) DCGAN, (c) WGAN, (d) LSGAN, and (e) STFTSynth (full model). The horizontal and vertical arrows indicate the time and frequency axes, respectively.*

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

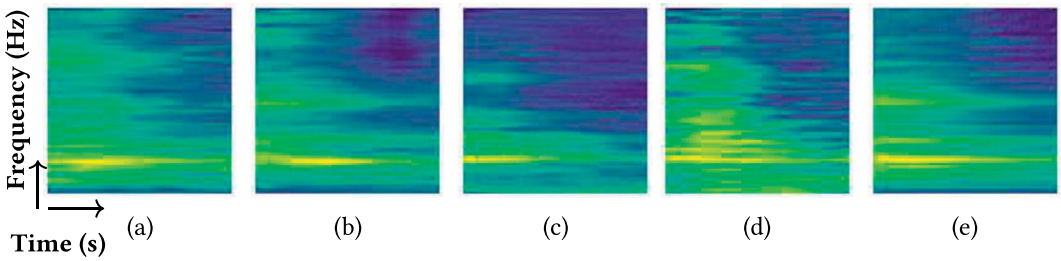

Figure 11. *STFT representations of real and generated wire breakage spectrograms (window size = 256): (a) Real STFT, (b) DCGAN, (c) WGAN, (d) LSGAN, and (e) STFTSynth (full model). The horizontal and vertical arrows indicate the time and frequency axes, respectively.*

It is important to highlight that the evaluation metrics used in this study assess the quality of the generated STFT spectrograms at the representation level. Improvements in SSIM and PSNR indicate better preservation of spectral patterns, while FID quantifies the similarity between the distributions of real and generated samples. These metrics provide evidence that the synthetic spectrograms capture the key statistical and structural characteristics of the original data. However, they do not directly measure performance in downstream SHM tasks such as event classification or detection. Therefore, they should be interpreted as necessary but not sufficient indicators of the potential usefulness of synthetic data in subsequent monitoring applications.

5.3. Ablation study

An ablation study was conducted to investigate the individual contributions of the DRBs and BiGRU layers within the STFTSynth architecture. Different variants of STFTSynth were developed by selectively removing these components to assess their impact on model performance. The evaluation specifically focused on wire breakage STFT spectrograms.

The results presented in Table 5 clearly illustrate the essential roles of DRB and BiGRU layers in enhancing STFTSynth's performance for generating high-quality STFT spectrograms. The full STFTSynth model, incorporating both DRB and BiGRU, achieves the highest performance across all metrics, with the best SSIM (0.488), highest PSNR (13.731), and lowest FID (0.210). These values confirm its strong capability to generate realistic and structurally detailed spectrograms.

Removing the DRB layers from STFTSynth leads to a performance decline across all metrics, especially reflected by an increased (worsened) FID score, rising from 0.210 to 0.260. This demonstrates the importance of the DRB layers in capturing detailed spectral structures. However, removing the DRB has a minimal effect on computational cost, as indicated by the training time per epoch (TTE), which remains nearly unchanged. The removal of BiGRU layers has a greater degradation in performance, with SSIM dropping to 0.346, PSNR reducing to 12.831, and the FID score increasing notably to 0.330. This substantial drop highlights BiGRU's essential role in modeling temporal dependencies, critical for

Table 5. *Performance metrics of different STFTSynth variants on wire breakage STFT spectrograms*

STFTSynth variants	SSIM	PSNR	FID	TTE [s]*
Full model	0.488	13.7310	0.210	15.50
No DRB	0.431	13.1517	0.260	15.00
No BiGRU	0.346	12.8312	0.330	2.50
No DRB & BiGRU	0.367	13.4090	0.219	1.10

*TTE, Training Time per Epoch (in seconds).

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

accurately capturing realistic STFT spectrogram patterns. However, eliminating the BiGRU layers significantly improves computational efficiency, reducing the TTE to just 2.50 seconds per epoch. Thus, the absence of BiGRU introduces a clear trade-off between computational efficiency and the quality of generated spectrograms. The effectiveness of the BiGRU layers can be attributed to the temporal characteristics of the STFT spectrogram, which encodes signal evolution across both time and frequency. Although convolutional layers are effective at capturing local spatial features, they are intrinsically limited in modeling long-range temporal dependencies. In contrast, BiGRUs process sequences in both forward and backward directions, enabling the model to learn temporal dependencies from both past and future contexts. This bidirectional processing is particularly beneficial for vibration signals, where structural anomalies such as wire breakage may exhibit subtle patterns that evolve over time and are not strictly one-directional. This way, BiGRUs can enhance the generator's ability to produce more realistic and temporally coherent spectrograms.

Removing both DRB and BiGRU layers returns STFTSynth to its original WGAN-GP baseline. This variant performs slightly better than scenarios where only one of these components is removed, achieving an FID score of 0.219, lower than both the "No DRB" and "No BiGRU" variants. However, the SSIM (0.367) and PSNR (13.409) values remain substantially lower compared to the full STFTSynth, clearly emphasizing that the integration of both DRB and BiGRU significantly enhances the model's ability to produce realistic STFT spectrograms. This baseline variant, however, offers the fastest computational efficiency, achieving the lowest TTE of only 1.10 seconds per epoch.

Overall, the ablation study conclusively demonstrates that integrating both DRB and BiGRU layers into the STFTSynth architecture provides substantial improvements in the quality of generated STFT spectrograms. Despite increased computational costs, the integration of both layers is justified by the enhanced performance.

6. Limitations and future work

Even though the proposed framework considers all generated samples uniformly during training, their practical value in SHM application is not uniform. For instance, rare, high-energy transient events such as wire breakage can benefit more from synthetic augmentation than operational or environmental noise signals as they are more difficult to reproduce experimentally. From a data-centric perspective, synthetic data are particularly useful in scenarios where datasets are highly imbalanced or when specific events are extremely rare, which is commonly the case for damage events such as prestressing wire breakage in SHM. In such situations, collecting additional real-world data cannot be possible or it might be costly. It is also important to note that simply increasing the amount of data does not always lead to better model performance. When synthetic samples do not introduce additional variability beyond what is already present in the real data, the benefits of augmentation may be saturated. Understanding how different proportions of synthetic data influence model performance is an important topic for future research, particularly in downstream SHM tasks. Besides, some learning models can perform reasonably well under rare and imbalanced data conditions; their results may still lack reliability and generalizability due to the limited availability of representative samples. These considerations highlight the need for data-centric strategies that carefully balance synthetic augmentation and real data acquisition, which we identify as a promising direction for future work.

Although the proposed GAN-based framework shows promising results, it remains constrained by the diversity and quality of the available real-world dataset. Synthetic generation can only reflect the variability present in the original samples. Expanding data collection across different structures and operational conditions is therefore essential, and future work should also investigate semi-supervised and self-supervised approaches to improve performance under limited labeling. In addition, the present study focuses on generating synthetic spectrograms without integrating them into a downstream classification task. Due to the rarity of real wire breakage events, building a statistically meaningful and balanced classification benchmark was not feasible in the current setting. Future studies should therefore evaluate the effect of GAN-augmented datasets on SHM classifiers in order to assess their practical utility more directly.

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

STFT was chosen for its fixed-size outputs and compatibility with GAN training; however, alternative representations such as wavelet scalograms may also be beneficial to capture highly non-stationary or overlapping events. Diffusion-based generative models also present a promising alternative, offering improved stability, sample diversity, and the potential to reconstruct time-domain signals for more direct physical validation. Finally, real-world monitoring often involves overlapping acoustic events, which were absent in this controlled dataset. Addressing this challenge will require integrating source separation techniques to separate concurrent signals, improving both data generation and downstream event detection.

7. Conclusion

This study proposed a GAN-based framework for representation-level synthesis of STFT spectrograms associated with rare acoustic events in SHM. A novel customized model, STFTSynth, was introduced to generate realistic STFT spectrograms of various structural events. By integrating dilated residual blocks and bidirectional GRU layers, the model enhances both spectral and temporal coherence of the generated data. STFTSynth was evaluated against established baselines—DCGAN, WGAN-GP, and LSGAN—using quantitative metrics (SSIM, PSNR, and FID) and qualitative visual assessments.

The results demonstrate that STFTSynth consistently achieves the highest SSIM and PSNR values, indicating its superior ability to generate high-quality spectrograms with minimal distortion. FID scores further confirm that STFTSynth produces the most realistic samples in the majority of cases, although LSGAN shows competitive performance for specific classes, such as the trimmer event. In contrast, WGAN-GP consistently exhibited the weakest performance, with lower SSIM and PSNR values, highlighting its limitations in preserving fine-grained structural and visual details. Qualitative assessments align closely with the quantitative results, with STFTSynth-generated spectrograms visually resembling real-world patterns more accurately than those produced by other models. The findings validate STFTSynth as an effective representation-level generator of single-channel STFT spectrograms for rare acoustic events in SHM. The results demonstrate improved spectrogram fidelity and distributional similarity compared with the selected GAN baselines. The effect of the generated samples on downstream event classification, damage detection, or anomaly detection remains a direction for future work.

Data availability statement. The data used in this study are not publicly available due to confidentiality and infrastructure security restrictions associated with the monitored site.
A limited subset of the data, or derived features, may be made available from the corresponding author upon reasonable request, subject to approval by the data owner.

Acknowledgements. The authors would like to express their sincere gratitude to Strada dei Parchi SpA for providing the opportunity and support necessary to conduct the experimental tests.

Author contribution. S.F.: Conceptualization, Data Curation, Methodology, Software, Validation, Formal analysis, Investigation, Supervision, Visualization, Writing—original draft & editing. Co-author M.I.: Methodology, Software, Formal analysis. Co-author M.C.: Supervision, Resources, Writing—original draft, Writing—review & editing. Co-author E. C.: Supervision, Writing—review & editing. Co-author G. V.: Supervision, Writing—review & editing. All authors approved the final manuscript.

Competing interests. The authors declare no competing interests.

AI statement. Generative AI tools were used to improve the clarity and readability of the manuscript (e.g., grammar and style suggestions). All technical content, experimental design, results, and interpretations were produced and verified by the authors, who take full responsibility for the manuscript.

References

Arjovsky M, **Chintala S and Bottou L** (2017) Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214–223). Pmlr.

Azad MM and Kim HS (2024) Hybrid deep convolutional networks for the autonomous damage diagnosis of laminated composite structures. *Composite Structures 329*, 117792. https://doi.org/10.1016/j.compstruct.2023.117792.

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

Bishop CM and Bishop H (2024) *Deep Learning: Foundations and Concepts*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-45468-4.

Bull LA, **Gardner P**, **Rogers TJ**, **Cross EJ**, **Dervilis N and Worden K** (2021) Probabilistic inference for structural health monitoring: New modes of learning from data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 7*(1), 03120003. https://doi.org/10.1061/AJRUA6.0001106.

Chen J, **Chan I and Brilakis I** (2024) Shifting research from defect detection to defect modeling in computer vision-based structural health monitoring. *Automation in Construction 164*, 105481. https://doi.org/10.1016/j.autcon.2024.105481.

Cho K, **van Merrienboer B**, **Bahdanau D and Bengio Y** (2014) On the properties of neural machine translation: Encoder-decoder approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111. Association for Computational Linguistics.

Cho K, **van Merrienboer B**, **Gulcehre C**, **Bahdanau D**, **Bougares F**, **Schwenk H and Bengio Y** (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics.

Chung J, **Gulcehre C**, **Cho K and Bengio Y** (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS 2014 Deep Learning and Representation Learning Workshop.*

Donahue C, **McAuley J and Puckette M** (2018) Adversarial audio synthesis. Version number: 3.

Duran B, **Emory D**, **Azam YE and Linzell DG** (2025) A novel CNN architecture for robust structural damage identification via strain measurements and its validation via full-scale experiments. *Measurement 239*, 115393. https://doi.org/10.1016/j.measurement.2024.115393.

Elman JL (1990) Finding structure in time. *Cognitive Science 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1.

Engel J, **Agrawal KK**, **Chen S**, **Gulrajani I**, **Donahue C and Roberts A** (2019) GANSynth: Adversarial neural audio synthesis. Version number: 2.

Farhadi S, **Corrado M**, **Borla O and Ventura G** (2024a) Prestressing wire breakage monitoring using sound event detection. *Computer-Aided Civil and Infrastructure Engineering 39*(2), 186–202. https://doi.org/10.1111/mice.13079.

Farhadi S, **Corrado M and Ventura G** (2024b) Acoustic event-based prestressing concrete wire breakage detection. *Procedia Structural Integrity 64*, 549–556. https://doi.org/10.1016/j.prostr.2024.09.305.

Farhadi S, **Corrado M and Ventura G** (2024c) Automated acoustic event-based monitoring of prestressing tendons breakage in concrete bridges. *Computer-Aided Civil and Infrastructure Engineering 39*(24), 3700–3720. https://doi.org/10.1111/mice.13321.

Ferro GA, **Restuccia L**, **Falliano D**, **Devitofranceschi A, and Gemelli A** (2022) Collapse of existing bridges: From the lesson of La Reale viaduct to the definition of a partial safety coefficient of variable traffic loads. *Journal of Structural Engineering 148*(11), 04022181. https://doi.org/10.1061/(ASCE)ST.1943-541X.0003458.

Fu W, **Zhou R and Guo Z** (2024) Concrete acoustic emission signal augmentation method based on generative adversarial networks. *Measurement 231*, 114574. https://doi.org/10.1016/j.measurement.2024.114574.

Gao Y, **Zhai P and Mosalam KM** (2021) Balanced semisupervised generative adversarial network for damage assessment from low-data imbalanced-class regime. *Computer-Aided Civil and Infrastructure Engineering 36*(9), 1094–1113. https://doi.org/10.1111/mice.12741.

Goodfellow IJ, **Pouget-Abadie J**, **Mirza M**, **Xu B**, **Warde-Farley D**, **Ozair S**, **Courville A and Bengio Y** (2014) Generative adversarial networks. Version number: 1.

Gui J, **Sun Z**, **Wen Y**, **Tao D and Ye J** (2023) A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering 35*(4), 3313–3332. https://doi.org/10.1109/TKDE.2021.3130191.

Gulrajani I, **Ahmed F**, **Arjovsky M**, **Dumoulin V and Courville A** (2017) Improved training of Wasserstein GANs. Version number: 3.

Guo Z, **Zhou R**, **Gao Y**, **Fu W and Yu Q** (2024) Towards ultrasonic guided wave fine-grained damage detection on hierarchical multi-label classification network. *Mechanical Systems and Signal Processing 218*, 111582. https://doi.org/10.1016/j.ymssp.2024.111582.

Heusel M, **Ramsauer H**, **Unterthiner T**, **Nessler B and Hochreiter S** (2017) GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, **Vol**. *30*, pp. 6626–6637.

Hou J, **Jiang H**, **Wan C**, **Yi L**, **Gao S**, **Ding Y and Xue S** (2022) Deep learning and data augmentation based data imputation for structural health monitoring system in multi-sensor damaged state. *Measurement 196*, 111206. https://doi.org/10.1016/j.measurement.2022.111206.

Jordan MI (1986) *Serial Order: A Parallel Distributed Processing Approach*. Tech. rep., Institute for Cognitive Science, University of California, San Diego.

Ko J and Ni Y (2005) Technology developments in structural health monitoring of large-scale bridges. *Engineering Structures 27*(12), 1715–1725. https://doi.org/10.1016/j.engstruct.2005.02.021.

Kumar K, **Kumar R**, **de Boissiere T**, **Gestin L**, **Teoh WZ**, **Sotelo J**, **de Brebisson A**, **Bengio Y and Courville A** (2019) MelGAN: Generative adversarial networks for conditional waveform synthesis. Version number: 3.

Le X, **Mei J**, **Zhang H**, **Zhou B and Xi J** (2020) A learning-based approach for surface defect detection using small image datasets. *Neurocomputing 408*, 112–120. https://doi.org/10.1016/j.neucom.2019.09.107.

Li Y, **Gan Z**, **Shen Y**, **Liu J**, **Cheng Y**, **Wu Y**, **Carin L**, **Carlson D and Gao J** (2018) StoryGAN: A sequential conditional GAN for story visualization. Version number: 2.

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.

Li B, **Guo H and Wang Z** (2024) Data augmentation using CycleGAN-based methods for automatic bridge crack detection. *Structure 62*, 106321. https://doi.org/10.1016/j.istruc.2024.106321.

Liao S, **Lan S and Zachariah AG** (2024) EVA-GAN: Enhanced various audio generation via scalable generative adversarial networks. Version number: 1.

Luleci F, **Catbas FN and Avci O** (2022) Generative adversarial networks for data generation in structural health monitoring. *Frontiers in Built Environment 8*, 816644. https://doi.org/10.3389/fbuil.2022.816644.

Luleci F, **Catbas FN and Avci O** (2023) Generative adversarial networks for labeled acceleration data augmentation for structural damage detection. *Journal of Civil Structural Health Monitoring 13*(1), 181–198. https://doi.org/10.1007/s13349-022-00627-8.

Mao X, **Li Q**, **Xie H**, **Lau RYK**, **Wang Z and Smolley SP** (2017) Least squares generative adversarial networks. *arXiv: 1611.04076*. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2794–2802.

Mousavi V, **Rashidi M**, **Ghazimoghadam S**, **Mohammadi M**, **Samali B and Devitt J** (2025) Transformer-based time-series GAN for data augmentation in bridge digital twins. *Automation in Construction 175*, 106208. https://doi.org/10.1016/j.autcon.2025.106208.

Parziale M, **Silva PH**, **Giglio M and Cadini F** (2024) Explainability of convolutional neural networks for damage diagnosis using transmissibility functions. *Structure 69*, 107583. https://doi.org/10.1016/j.istruc.2024.107583.

Radford A, **Metz L and Chintala S** (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations (ICLR 2016)*.

Ren Z, **Yeh H and Lin MC** (2013) Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics 32*(1), 1–16. https://doi.org/10.1145/2421636.2421637.

Saleem MR, **Park J-W**, **Lee J-H**, **Jung H-J and Sarwar MZ** (2021) Instant bridge visual inspection using an unmanned aerial vehicle by image capturing and geo-tagging system and deep convolutional neural network. *Structural Health Monitoring 20*(4), 1760–1777. https://doi.org/10.1177/1475921720932384.

Smith SW (1997) *The Scientist and Engineer's Guide to Digital Signal Processing*, 1st Edn. San Diego, CA: California Technical Publ.

Su K, **Qian K**, **Shlizerman E**, **Torralba A and Gan C** (2023) Physics-driven diffusion models for impact sound synthesis from videos.

Tran N-T, **Tran V-H**, **Nguyen N-B**, **Nguyen T-K and Cheung N-M** (2021) On data augmentation for GAN training. *IEEE Transactions on Image Processing 30*, 1882–1897. https://doi.org/10.1109/TIP.2021.3049346.

Wang D, **Zhang Y**, **Zhang H**, **Zhuang Y**, **Gao S and Li Y** (2025) Bearing fault diagnosis method based on multisensor hybrid feature fusion. *IEEE Transactions on Instrumentation and Measurement 74*, 1–11.

Webb GT, **Vardanega PJ**, **Hoult NA**, **Fidler PRA**, **Bennett PJ and Middleton CR** (2017) Analysis of fiber-optic strain-monitoring data from a prestressed concrete bridge. *Journal of Bridge Engineering 22*(5), 05017002. https://doi.org/10.1061/(ASCE)BE.1943-5592.0000996.

Xia X, **Pan X**, **Li N**, **He X**, **Ma L**, **Zhang X and Ding N** (2022) GAN-based anomaly detection: A review. *Neurocomputing 493*, 497–535. https://doi.org/10.1016/j.neucom.2021.12.093.

Xie X, **Li Y**, **Chen W**, **Ouyang K**, **Jiang L and Zheng H** (2022) Global mixup: Eliminating ambiguity with clustering. Version number: 1.

Zhou Y, **Liu Y**, **Lian Y**, **Pan T**, **Zheng Y and Zhou Y** (2025) Ambient vibration measurement-aided multi-1D CNNs ensemble for damage localization framework: Demonstration on a large-scale RC pedestrian bridge. *Mechanical Systems and Signal Processing 224*, 111937. https://doi.org/10.1016/j.ymssp.2024.111937.

Zhou Wang AB, **Sheikh H and Simoncelli E** (2004) Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861.

Zingoni A (ed.) (2019) Advances in engineering materials, structures and systems: innovations, mechanics and applications. In *Proceedings of the Seventh International Conference on Structural Engineering, Mechanics, and Computation, 2–4 September 2019, Cape Town, South Africa*. Boca Raton: CRC Press, Taylor & Francis Group.

Zinno R, **Haghshenas SS**, **Guido G and VItale A** (2022) Artificial intelligence and structural health monitoring of bridges: A review of the state-of-the-art. *IEEE Access 10*, 88058–88078. https://doi.org/10.1109/ACCESS.2022.3199443.

Cite this article: Farhadi S, Iavarone M, Corrado M, Chatzi E and Ventura G (2026). A GAN-based framework for generating STFT spectrograms of rare acoustic events in structural health monitoring. *Data-Centric Engineering*, 7, e42. doi:10.1017/dce.2026.10071

Downloaded from https://www.cambridge.org/core. 24 Sep 2026 at 09:01:26, subject to the Cambridge Core terms of use.