Received: 26 May 2025 | Accepted: 4 August 2025

DOI: 10.1111/2041-210X.70148

Methods in Ecology and Evolution

RESEARCH ARTICLE

Animal acoustic identification, denoising and source separation using generative adversarial networks

Mei Wang[1] | Kevin F. A. Darras[2] | Renjie Xue[3,4] | Fanglin Liu[1,3]

[1]Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China

[2]EFNO, ECODIV, INRAE, Domaine des Barres, Nogent-sur-Vernisson, Centre-Val de Loire, France

[3]Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China

[4]School of Computer Science and Artificial Intelligence, Hefei Normal University, Hefei, China

Correspondence
Fanglin Liu
Email: flliu@ipp.ac.cn

Handling Editor: Jeffrey Doser

Abstract

1. Soundscapes contain rich ecological information, offering insights into both biodiversity and ecosystem dynamics. However, the sheer volume of data produced by passive acoustic monitoring presents significant challenges for scalable analysis and ecological interpretation. While convolutional neural networks (CNNs) have advanced species classification in bioacoustics, they often struggle with identifying acoustic targets in acoustic space and quantifying soundscapes' characteristics.
2. In this study, we propose a novel spectrogram-to-spectrogram translation framework based on generative adversarial networks (GANs) to isolate and quantify acoustic sources within soundscape recordings. Our method is trained on paired spectrogram images: original full-spectrogram representations and target spectrogram representations containing only the vocalizations of specific sound labels. This design enables the model to learn source-specific mappings and perform both the species and community-level separation of acoustic components in soundscape recordings.
3. We developed and evaluated two GAN-based models: a species-level GAN targeting eight avian species, and a community-level GAN distinguishing among avian, insect and anthropogenic sound sources. The models were trained and tested using soundscape recordings collected from the Yaoluoping National Nature Reserve, eastern China. The species-level model achieved a mean F1 score of 0.76 for pixel-wise detection, while the community-level model reached 0.79 across categories. In addition to precise temporal-spectral localization, our approach captures sources' acoustic occupancy and frequency distribution patterns, offering deeper ecological insight. Compared to baseline CNN classifiers, our model achieved a mean F1 score of 0.97, demonstrating comparable classification performance to ResNet50 (0.95) and VGG16 (0.98) across multiple species. Our GAN approach for extracting sound sources also significantly outperformed conventional methods in denoising and source separation, as indicated by lower image-level mean squared error.

4. These results demonstrate the utility of GANs in advancing ecoacoustic analyses and biodiversity monitoring. By enabling robust source separation and fine-resolution signal mapping, the proposed approach contributes a scalable and transferable tool for soundscape quantification.

KEYWORDS
acoustic space, classification, denoise, generative adversarial network (GAN), source separation

1 | INTRODUCTION

Acoustic communication is common among animals, including insects, fish, amphibians, birds and mammals (Xie et al., 2020). These vocalizations, produced in the context of key behavioural activities, serve essential ecological functions, including territory defence, predator avoidance and mate attraction (Malavasi & Farina, 2012). Passive acoustic monitoring has enabled the collection of extensive soundscape datasets, capturing the interplay of biological, abiotic and anthropogenic sounds (Darras et al., 2025; Pijanowski et al., 2011). However, interpreting these diverse datasets remains challenging due to overlapping acoustic sources and species diversity (Gasc et al., 2013; Gibb et al., 2018).

Automated acoustic analysis methods, ranging from species-specific classifiers based on temporal, spectral or cepstral features, to holistic acoustic indices that characterize soundscape structure, have advanced the capacity to extract ecological insights from audio recordings (Keen et al., 2021; Molnár et al., 2008; Yip et al., 2021). Advances in deep learning, particularly convolutional neural networks (CNNs), have reframed bioacoustic tasks as image classification problems by treating spectrograms as input images (Stowell, 2022), enabling effective detection of species-specific vocalizations and anomalous acoustic events (Chronister et al., 2023; Napier et al., 2024). Nevertheless, these approaches often face significant challenges in accurate target sound event detection within acoustic space and cannot count sound signals, limiting their applicability in bioacoustics tasks such as abundance estimation, especially under high background noise (Napier et al., 2024). These limitations highlight the need for more robust and generalizable methods.

A range of bioacoustic denoising methods, such as spectral subtraction, wavelet-based filtering and deep learning techniques, has been developed to reduce noise in natural recordings (Xie et al., 2021). Approaches, such as independent component analysis (ICA), principal component analysis (PCA), non-negative matrix factorization (NMF) and supervised deep learning models, have been employed to track target sources in sound mixtures (Lin & Tsao, 2020; Sun, Yen, & Lin, 2022). However, due to the inherent complexity of acoustic environments in nature, reliably separating or identifying biological signals remains an open and unresolved challenge.

An emerging alternative is to reframe sound analysis as a generative problem. Unlike discriminative models that aim to distinguish between predefined classes, generative approaches learn the underlying distribution of the data, enabling the synthesis of new samples that capture latent signal structures (Metri & Mamatha, 2021; Wu et al., 2009). This perspective offers advantages for modelling complex, noisy soundscapes by capturing intrinsic acoustic characteristics rather than relying solely on explicit class boundaries. Despite its promise, the application of generative models for source separation and ecological interpretation in ecoacoustics remains underexplored.

Generative adversarial networks (GANs) have emerged as a powerful generative framework within deep learning (de Souza et al., 2023). A standard GAN architecture consists of a generator that produces synthetic data and a discriminator that distinguishes between real and generated samples (Metri & Mamatha, 2021). Conditional variants, such as pix2pix (Abdelmotaal et al., 2021), extend this framework to supervised image-to-image translation by conditioning the generation process on paired inputs. In the context of ecoacoustics, GAN-based models have been explored primarily for data augmentation purposes. For example, WaveGAN has been employed to synthesize raw environmental audio waveforms, providing additional training data for a sound classification model (Zhang et al., 2022). DCGAN has been applied to generate mel-frequency cepstral coefficient spectrograms of environmental sounds, expanding the variability of input data for improved model robustness (Bahmei et al., 2022). ACGAN has been used to generate class-conditional spectrograms of bird calls, enabling more balanced datasets and improved species classification performance (Fu et al., 2023). These studies demonstrate GANs' potential to enrich training datasets and boost classification models.

However, to our knowledge, GAN-based image-to-image translation has not been applied to soundscape recordings due to the lack of paired training data, where each full-spectrogram input is matched with a target spectrogram containing only target signals. Developing such a framework could disentangle mixed acoustic sources, facilitating the identification and quantification of species- or group-specific vocal activity within complex soundscapes.

In this study, we present a novel spectrogram translation method using the pix2pix GAN architecture to isolate and quantify target acoustic sources from real-world ecoacoustic data.

Leveraging spectrograms of field recordings from the Yaoluoping National Nature Reserve (YNNR), China, we trained two models: one at the species level, targeting eight bird species and another at the community level, distinguishing among birds, insects and anthropogenic sounds. By learning from paired spectrograms, originals and manually annotated target masks, we enable the generator to reconstruct source-specific spectrograms from mixed-source recordings.

To evaluate model performance, we employed perceptual similarity metrics including structural similarity index measure (SSIM) (Wang et al., 2004) and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018), assessing the fidelity of the translated outputs. The trained models also quantified vocal activity across frequency, time and amplitude dimensions, providing insights into ecological acoustic niches and interspecific overlap. Comparative experiments against baseline classifiers trained with the same data investigate species recognition performance. We further evaluated model performance in denoising and source separation by comparing it with conventional approaches.

2 | MATERIALS AND METHODS

2.1 | Audio data collection

YNNR is situated in the western region of Anhui Province, China (116°02′ E–116°11′ E, 30°57′ N–31°06′ N). Nestled within the Tapieh Mountains, YNNR is a critical water conservation forest and a refuge for nationally protected and endangered flora and fauna. According to prior studies (Li et al., 2017; Wang et al., 2023), the dominant sound sources in this region include birds and insects, with intermittent contributions from anthropogenic activities.

To capture acoustic data representative of the area's habitat diversity, five recording sites were established across a gradient of altitudes. Automated sound recorders (SM4+; Wildlife Acoustics, Maynard, MA, USA) were mounted on tree trunks at approximately 1.5 m above ground to reduce understory obstruction and optimize detection of airborne vocalizations.

Each device was configured to record at a sampling rate of 24 kHz, capturing signals within the 0–12 kHz frequency range. This recording strategy covers the frequency range of most bird songs (Mikula et al., 2020). Recordings were saved in uncompressed WAV format on SD cards. Our temporal sampling followed a duty cycle of 5-min recordings every 30 min throughout the 24-h cycle. This systematic temporal monitoring was conducted from April 5, 2019, to April 30, 2020. During the peak-breeding season from April to July in 2019, dawn periods were additionally monitored using continuous recordings to maximize the capture of heightened vocal activity.

Field recordings were conducted in the YNNR under a long-term institutional agreement between the University of Science and Technology of China and the Reserve. The acoustic monitoring was non-invasive, and no specific licence or permit was required for this fieldwork.

2.2 | Data pre-processing

2.2.1 | Target acoustic species and communities

To facilitate model training and analysis, the recordings were processed using AnalysisPrograms.exe (Towsey et al., 2018). Each audio file was segmented into 5-s units, converted from stereo to mono by mixing, to ensure standardization across the dataset.

Vocalizing taxa were identified through a combination of auditory and visual inspection using Raven Pro (K. Lisa Yang Center for Conservation Bioacoustic, 2014), which enabled simultaneous audio playback and spectrogram analysis. For species-level modelling, we selected eight representative and acoustically distinctive bird species commonly found within the reserve: Jungle Nightjar (*Caprimulgus indicus*, 0.3–2.2 kHz), Eurasian Jay (*Garrulus glandarius*, 0.4–11.7 kHz), Koklass Pheasant (*Pucrasia macrolopha*, 0.2–8.6 kHz), Oriental Scops Owl (*Otus sunia*, 0.7–1.5 kHz), Lesser Cuckoo (*Cuculus poliocephalus*, 1.0–3.2 kHz), Brownish-flanked Bush Warbler (*Horornis fortipes*, 1.1–8.5 kHz), Alström's Warbler (*Phylloscopus soror*, 1.8–8.7 kHz) and Hartert's Leaf Warbler (*Phylloscopus goodsoni*, 1.9–9.8 kHz). These frequency ranges were determined by examining 10 A-rated recordings for each species from Xeno-Canto, focusing on harmonics and dominant energy bands. We note that under certain behavioural or environmental conditions (e.g. courtship displays, alarm calls) or favourable recording conditions (e.g. when the source is close), some harmonics or call elements may extend slightly beyond these primary ranges.

We additionally developed a community-level model designed to generalize across broader acoustic categories. This model included three primary vocal communities: avian vocalizations, insect-produced sounds and anthropogenic noises. Human-related acoustic events were restricted to public forest protection announcements, advertisement broadcasts and vehicle horns. Non-target signals, including dog barks, anuran calls, unidentified vocalizations and abiotic ambient sound (e.g. wind, rainfall), were regarded as background.

2.2.2 | Original and target images

Spectrograms were used to represent the acoustic space by mapping time (*x*-axis) and frequency (*y*-axis), with sound intensity encoded using a colour gradient. To generate spectrograms, raw audio files were processed using Python libraries (wave and matplotlib) and converted via a short-time Fourier transform employing a Hanning window of 512 samples and 50% overlap to capture temporal–spectral continuity. Each spectrogram image, corresponding to a 5-s audio segment, was then plotted at a resolution of 256 × 256 pixels to match the square input size used in the model.

For species-level data, spectrograms containing vocalizations of the eight selected bird species were manually annotated. Each species was assigned a unique colour for pixel-wise labelling, while non-target sounds were labelled in white. Similarly, for the

community-level dataset, which included birds, insects and anthropogenic sources, three colours represented the target communities, with non-target pixels again marked white.

Although sound sources may overlap, each pixel in the spectrogram was assigned a single colour class. Therefore, based on auditory cues and visual differences in the spectrograms, each pixel was labelled with only one corresponding sound category. Non-target backgrounds were manually removed using Adobe Photoshop's eraser tool. The resulting images were further refined in Python with the scikit-maad library (Ulloa et al., 2021) using intensity thresholding to ensure spectral fidelity and ecological relevance. On average, annotating 100 images took approximately 2 h. This is considerably more time-consuming than audio segment-level annotation, but offers finer temporal-spectral resolution.

Each spectrogram served as an input image, paired with a corresponding colour-coded target mask, forming the training set for the generative model. The species-level dataset contained 4150 image pairs: 500 samples per species and 150 background-only samples. It was randomly split into 80% training (3320 samples) and 20% testing (830 samples). The community-level dataset included 1200 image pairs with a roughly balanced representation of bird, insect and human sounds, though these sources sometimes co-occurred within the same image.

2.3 | GAN architecture

The pix2pix framework is a conditional GAN developed for image-to-image translation tasks, learning a mapping from input to target images (Isola et al., 2017). As illustrated in Figure 1, the architecture comprises two primary components: a generator and a discriminator.

The generator adopts an encoder-decoder structure implemented via the U-Net architecture (Ronneberger et al., 2015). It receives an original spectrogram as input and generates a synthetic image mimicking the annotated target. The encoder progressively compresses the input through a series of downsampling operations to a bottleneck representation, while the decoder symmetrically upsamples the latent vector back into an output image. Skip connections between encoder and decoder layers preserve fine-grained features and spatial resolution by enabling direct feature propagation.

The discriminator follows the PatchGAN design (Isola et al., 2017), which focuses on the discrimination of local regions (patches) in the

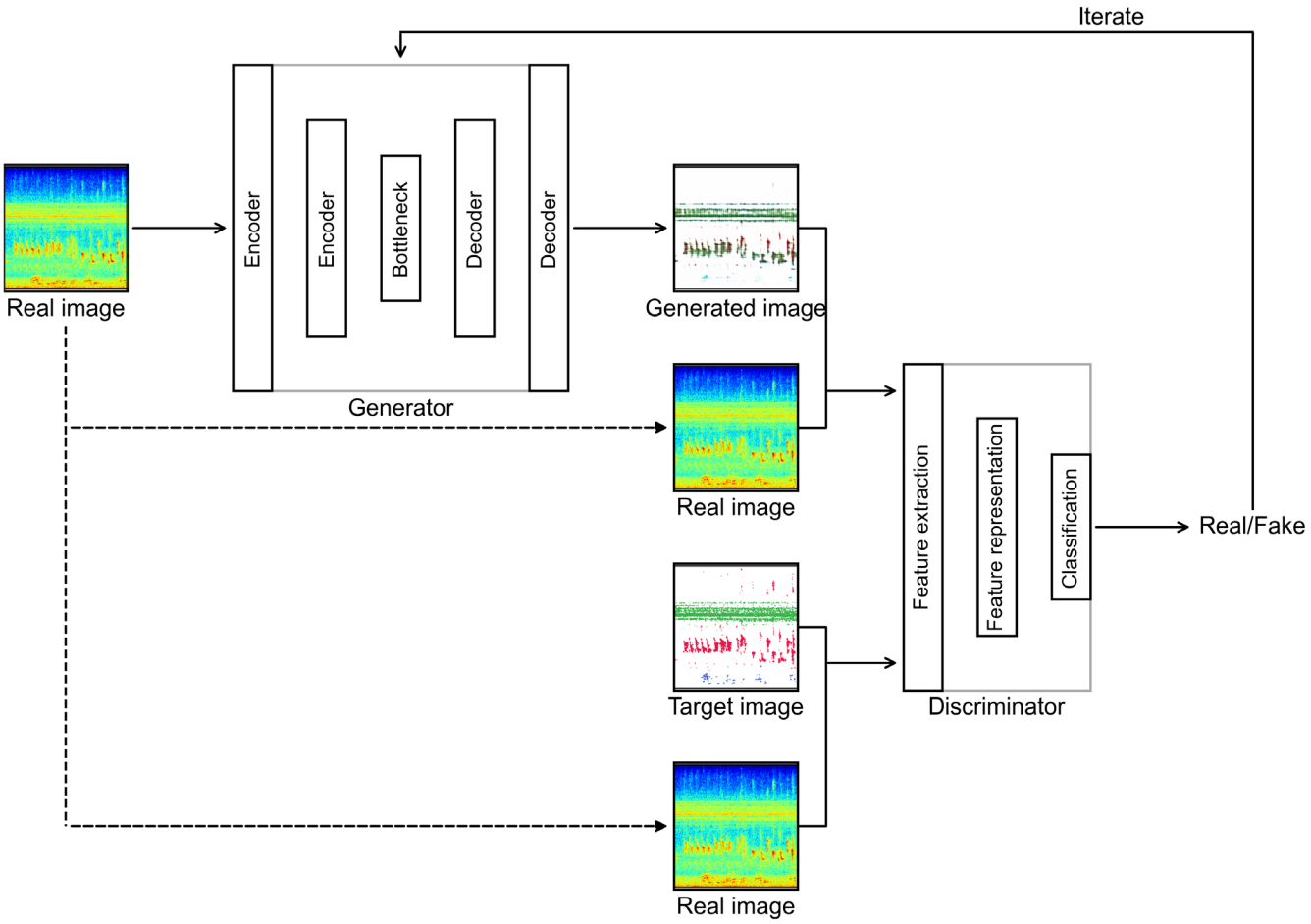

FIGURE 1 The generative adversarial network (GAN) model consists of a generator and a discriminator. The generator produces a new image based on the given original images, while the discriminator evaluates whether the generated image closely resembles the target. The process is iterated until the generator can generate sufficiently realistic images to deceive the discriminator.

image, thereby focusing on high-frequency structural fidelity at the local level. The discriminator outputs a probability map indicating whether each patch is real (from targeted images) or fake (from the generator), guiding the generator to produce perceptually realistic textures and boundaries.

Model training is guided by a composite loss function (de Souza et al., 2023). The generator is optimized using a combination of adversarial loss and L1 loss (mean absolute error). The adversarial loss is formulated as binary cross-entropy between predicted and true patch labels, while the L1 loss measures the pixel-wise distance between the generated and reference images. The discriminator, in contrast, is trained solely using adversarial loss to maximize its ability to distinguish real from synthesized samples.

The model training process was conducted on a GeForce GTX 4060 GPU. The model was trained using Python's torch library with the YNNR training set. Input and output resolutions were 256 × 256 pixels. The learning rate was set to 0.0002 with a batch size of 1. The ratio of adversarial loss to L1 loss for the generator was 1:100. While hyperparameters followed established literature (Isola et al., 2017), image augmentations such as cropping, rotation and flipping were omitted due to the frequency-time positional significance of audio signals in the spectrogram. Typically, GAN models do not converge in the conventional sense but reach a balance between the generator and discriminator (Abdelmotaal et al., 2021), making it difficult to determine when training should stop. As training progresses, image quality improves, but more epochs do not guarantee better results. Therefore, the final generator model is selected based on the quality of the generated images rather than the total training time. This can be achieved by loading each model and performing a translation on the original images from the training set for evaluation. Both species and community models were trained for 200 epochs, with checkpoints saved for monitoring and evaluation.

2.4 | Model evaluation

2.4.1 | Generated images quality assessment

To evaluate model performance, test-set spectrograms were input into the trained conditional GAN, generating synthetic spectrogram-like images with colours representing individual bird species or broader acoustic communities.

To quantitatively assess the similarity between the generated outputs and their corresponding annotated targets, we computed two complementary metrics: the SSIM (Wang et al., 2004) and the LPIPS metric (Zhang et al., 2018). SSIM evaluates the structural fidelity between two images by comparing luminance, contrast and spatial structure; its values range from 0 (no similarity) to 1 (perfect similarity). LPIPS measures perceptual dissimilarity in the deep feature space of pretrained networks, where lower scores indicate higher similarity.

In addition to these quantitative assessments, visual inspection qualitatively validated the fidelity of model outputs. Spectrograms from the model with optimal SSIM and LPIPS scores were visually compared with target images to verify that the GAN consistently reconstructed source-specific acoustic structures. This combined evaluation framework ensured both structural and perceptual integrity of the translated spectrograms and confirmed the applicability of the generative model to real-world soundscape scenarios.

2.4.2 | Pixel translation performance evaluation

To assign class labels to pixels in the generated spectrogram images, a colour-based thresholding approach was employed. Each species or acoustic community was represented by a distinct RGB colour in the annotated images, with 256 intensity levels per channel (Shailesh et al., 2016). For each pixel in the generated image, if its RGB values deviated by no more than ±15 intensity levels from the assigned class colour, it was attributed to that category; otherwise, it was classified as background. The threshold setting was based on visually perceived colour separability to ensure consistent and interpretable outputs from the GAN model. This procedure allowed the generated images to function as semantic colour masks, enabling per-pixel classification of acoustic sources.

In the species-level model, pixels were classified into nine categories: eight birds and background sounds. For the community-level model, pixel-wise classification was restricted to four categories: bird vocalizations, insect sounds, anthropogenic noise and background.

Using annotated test images as true labels and colour-mapped outputs as predictions, we evaluated classification performance using per-pixel precision, recall and F1 score, calculated with the scikit-learn library in Python. These metrics provided a comprehensive assessment of the model's ability to correctly identify acoustic sources within the soundscape at fine resolution.

2.5 | Acoustic space analysis

To demonstrate the ecological applicability of the conditional GAN framework, we applied the trained model to reconstruct purified spectrograms containing only target acoustic sources. These outputs served as filtered representations, from which source-specific acoustic information was quantitatively extracted.

Utilizing colour-coded masks, we delineated spectrogram regions for individual species or acoustic communities. For each segment, frequency distribution was computed by averaging pixel intensities along the time axis and highlighting the predominant frequency bands. Temporal occupancy was measured by calculating the total duration during which a species or community was acoustically active. Acoustic space (time-frequency) occupancy was assessed by summing all pixels attributed to a class across the spectrogram, indicating its acoustic prominence within the soundscape.

These operations were conducted for both the species-level and community-level models. The resulting metrics enabled precise and interpretable characterization of acoustic presence, offering insights into niche occupation, activity rhythms and relative abundance of vocalizing taxa within complex soundscapes.

2.6 | Model comparison

2.6.1 | Classifier

To evaluate the classification capabilities of our species-level GAN model, we treated it as a species classifier and benchmarked it against two widely used CNN architectures: ResNet50 and VGG16 (Gómez-Gómez et al., 2023; Nieto-Mora et al., 2023). Both have been extensively applied in ecological image and acoustic classification tasks, and thus served as baseline models. Importantly, this allowed us to use the same training data across all models to evaluate only the effect of the model architecture on classification performance. Comparisons with established models such as BirdNET (Kahl et al., 2021) or Perch (Hamer et al., 2023), use different and much larger training datasets which would confound the effect of training data and architecture.

Unlike the GAN-based approach, which utilized paired spectrogram-mask images, the baseline classifiers required only the original spectrograms as input. The same dataset as the species-level GAN was employed, split into 60% training, 20% validation and 20% testing, with the test set identical across models to ensure comparability.

In this multi-class classification task, each spectrogram segment was assigned a single bird species label. ResNet50 was trained for 400 epochs, while VGG16 converged after 200 epochs. All models were evaluated using standard image-level performance metrics, including precision, recall and F1 score, to assess their performance and generalization capabilities.

2.6.2 | Denoising

In addition to classification, we evaluated the GAN-based model for denoising by assessing its ability to reduce non-target acoustic interference and comparing it with two common noise reduction algorithms: spectral subtraction using Noisereduce, a Python-based toolkit designed for denoising time-domain signals such as speech, bioacoustics and physiological recordings; and Wiener filtering (Xie et al., 2021), a classical statistical method.

For this analysis, the GAN model reconstructed spectrograms from which non-target signals had been masked, serving as a colour-guided spectrogram restoration. To quantify denoising performance, we computed the mean squared error (MSE) (Liutkus et al., 2012; Sinha & Rajan, 2018) between reference spectrograms derived from manually annotated colour masks and the denoised outputs produced by each method. MSE captures the cumulative squared differences between corresponding pixel intensities, with lower values indicating closer resemblance to the manually cleaned reference spectrograms. This framework enabled objective comparison of the GAN model's noise suppression with conventional baseline approaches within ecoacoustic data.

2.6.3 | Sound source separation

To assess the effectiveness of the community-level GAN model in multi-source acoustic separation, we compared it with NMF (Virtanen, 2007), a well-established unsupervised source separation technique. NMF operates by iteratively decomposing a non-negative input spectrogram matrix into a set of basis functions and corresponding activation coefficients, enabling extraction of individual components through spectral dictionary learning.

Both methods aimed to isolate three major acoustic source types: avian vocalizations, insect sounds and anthropogenic noise. While NMF separated sources via matrix decomposition, the GAN model achieved isolation through colour-mask-guided spectrogram reconstruction.

To quantitatively evaluate separation fidelity, we computed the MSE between algorithm outputs and reference spectrograms from manually annotated colour masks. Lower MSE values indicated higher correspondence between the separated output and the reference source signals, thereby reflecting superior source separation performance. As NMF outputs are unlabelled, we tested all channel-class combinations and assigned labels based on the one with the lowest MSE.

3 | RESULTS

3.1 | Model evaluation

To evaluate the performance trajectory of the GAN during training, we conducted inference on the YNNR test set using models saved at different training epochs. For each checkpoint, SSIM and LPIPS metrics were computed to track image quality across training.

As illustrated in Figure 2, both metrics fluctuated during training. For the species-level model, optimal SSIM and LPIPS values were reached around epoch 18, indicating stable generative performance for multi-class acoustic sources at this stage.

In the community-level model, SSIM increased rapidly during the first 15 epochs, peaking at epoch 22, indicating that structural fidelity between generated and targeted images improved most significantly during early training. LPIPS decreased sharply early on and reached its minimum at epoch 16, reflecting enhanced perceptual similarity.

These trends offer valuable insights into the model's learning dynamics and inform the selection of optimal checkpoint epochs for downstream analysis.

In addition to quantitative evaluation, we conducted qualitative visual inspection of spectrogram outputs generated at various stages of model training. Figure 3 illustrates the species-level model's outputs at epochs: 1, 5, 10, 18, 50, 100 and 200. Visual

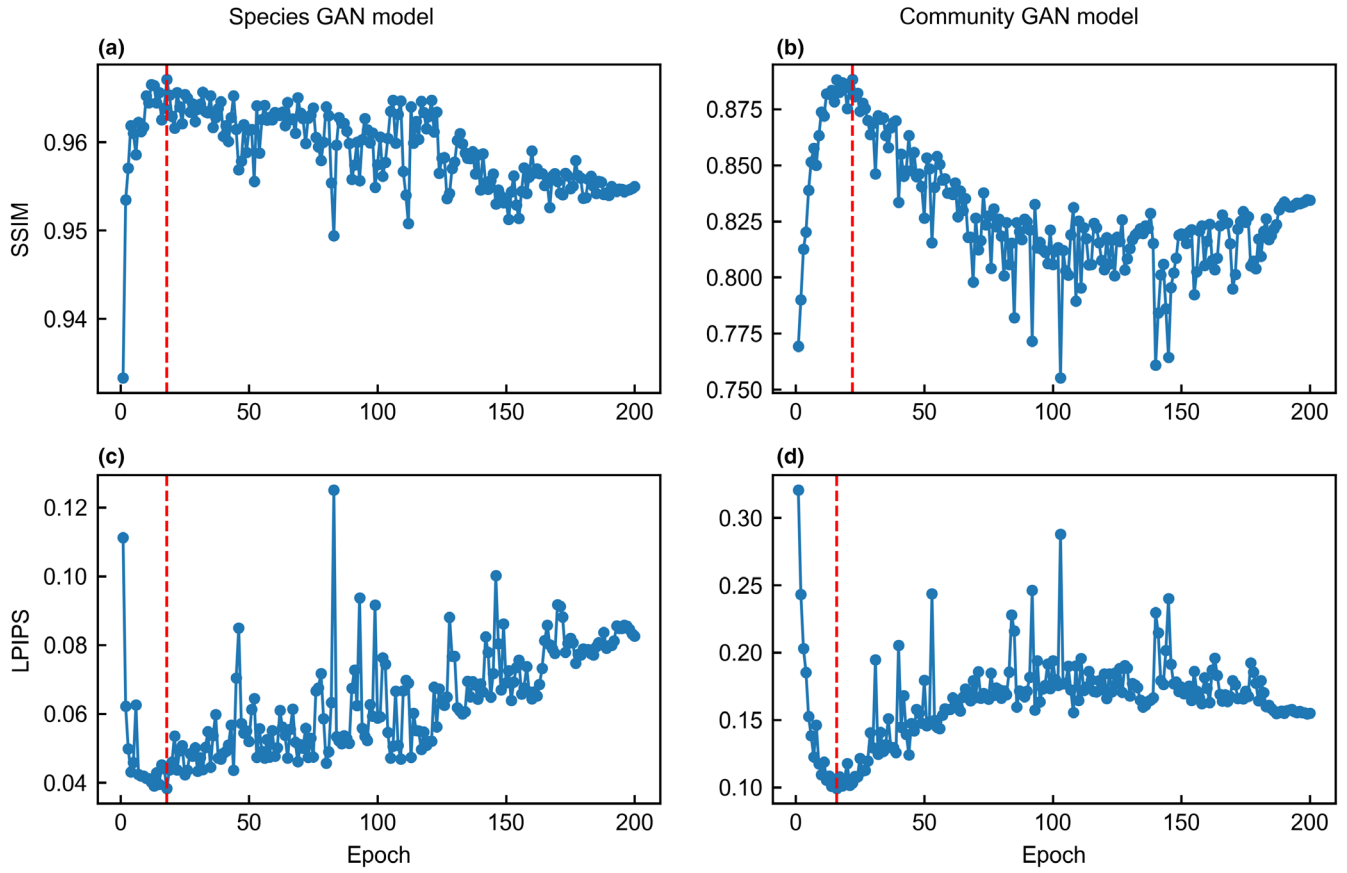

FIGURE 2 The structural similarity index measure (SSIM) and learned perceptual image patch similarity (LPIPS) during generative adversarial network (GAN) training. (a) The SSIM calculation results of the species GAN model. (b) The SSIM calculation results of the community GAN model. (c) The LPIPS calculation results of the species GAN model. (d) The LPIPS calculation results of the community GAN model. The red dashed line indicates the optimal training epoch for the model, corresponding to the maximum SSIM value and the minimum LPIPS value.

comparison reveals that epoch 18 produced spectrograms with the highest perceptual fidelity, exhibiting clearly species-specific features with minimal background noise. This aligns with the SSIM and LPIPS evaluations, indicating epoch 18 as the optimal model in both objective and visual evaluations.

Similarly, Figure 4 shows community-level model outputs at epochs 1, 5, 10, 16, 22, 100 and 200. Models from epochs 16 and 22 achieved superior separation of bird, insect and anthropogenic sounds, with well-defined class boundaries and minimal spectral overlap, corresponding to the lowest LPIPS and highest SSIM values. Both epochs 16 and 22 produced high-quality outputs, and we selected epoch 22 as the representative community-level model.

Overall, these visual comparisons reinforce the quantitative findings and validate the GAN's capacity to progressively improve acoustic source representation with continued training, while also identifying the training epochs that yield the most ecologically interpretable outputs.

Using the optimal models from training, we evaluated pixel-wise classification by comparing generated outputs with targeted images.

As presented in Table 1, the species-level model achieved an average F1 score of 0.76 across eight birds. Hartert's Leaf Warbler and Brownish-flanked Bush Warbler had the highest score, while Lesser Cuckoo had the lowest (0.69), indicating difficulty in distinguishing its acoustic features under complex soundscapes. Across all species, precision consistently exceeded recall, suggesting that while the GAN model localized vocalizations with high specificity in time-frequency space, it tended to underpredict the full extent of activity, possibly due to conservative masking near spectral boundaries.

For the community-level model (Table 2), the average F1 score reached 0.79 across the three vocal communities. Insect sounds achieved the highest accuracy, likely due to their distinct frequency patterns and temporal regularity, while anthropogenic sounds had the lowest performance, possibly due to their acoustic variability and overlap with biotic frequencies.

These results confirm the model's ability to effectively isolate and classify both species-specific and group-level vocalizations, while highlighting the challenges of distinguishing acoustically complex or diffuse sound sources.

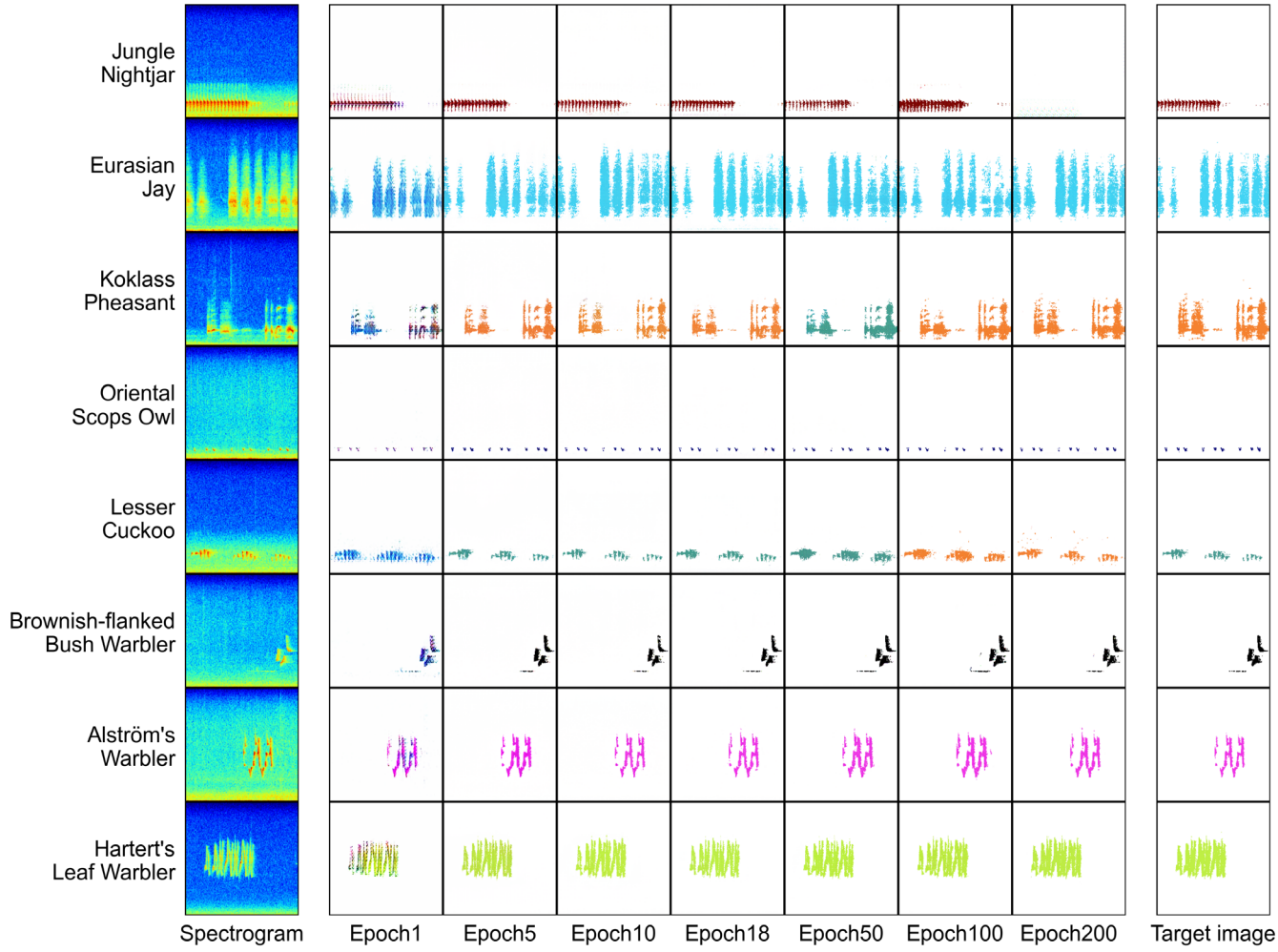

FIGURE 3 Visual comparison of images generated by the species generative adversarial network (GAN) model. From left to right are the original spectrogram, the generated images at different training epochs and the target image.

3.2 | Acoustic space qualification

Using the trained species-level GAN model, we reconstructed spectrograms containing only the vocalizations of individual target species. As shown in Figure 5, clear interspecific variation appeared across several acoustic dimensions.

The vocalizations of all eight species were within the frequency range below 12 kHz. Jungle Nightjar, Oriental Scops Owl and Lesser Cuckoo exhibited relatively low peak frequencies, whereas Hartert's Leaf Warbler presented the highest frequency peak. Oriental Scops Owl was confined to a narrow frequency band, while Eurasian Jay, Alström's Warbler and Hartert's Leaf Warbler demonstrated broader spectral ranges, indicating wider use of the acoustic frequency space.

Differences in acoustic amplitude were evident: Jungle Nightjar showed the highest intensity, whereas Alström's Warbler exhibited the lowest signal amplitude. With respect to temporal occupancy, Jungle Nightjar had the longest vocal activity, while Alström's Warbler was present for the shortest period. Interestingly, despite its relatively low amplitude and short duration, Alström's Warbler did not exhibit the smallest acoustic space (i.e. pixel-wise area) within the reconstructed spectrogram, suggesting broader spectral spread over shorter durations. Figure 6 shows similar results for the community-level GAN model: human-generated sounds occupied the lowest frequencies, followed by birds, while insects vocalized at the highest frequencies. Insect vocalizations also exhibited the highest acoustic intensity and longest duration among the three community types.

When considering overall acoustic space occupancy, insects covered the largest area in the reconstructed spectrograms, indicating their dominant presence in the soundscape. These findings highlight the considerable contribution of insect vocalizations to the temporal and spectral structure of the acoustic environment at YNNR.

3.3 | Model comparison

3.3.1 | Classifier

Table 3 compares species classification performance between our GAN-based model and two widely adopted CNN architectures, ResNet50 and VGG16. The average F1 score of eight birds was 0.97 for the GAN model, 0.95 for ResNet50 and 0.98 for VGG16.

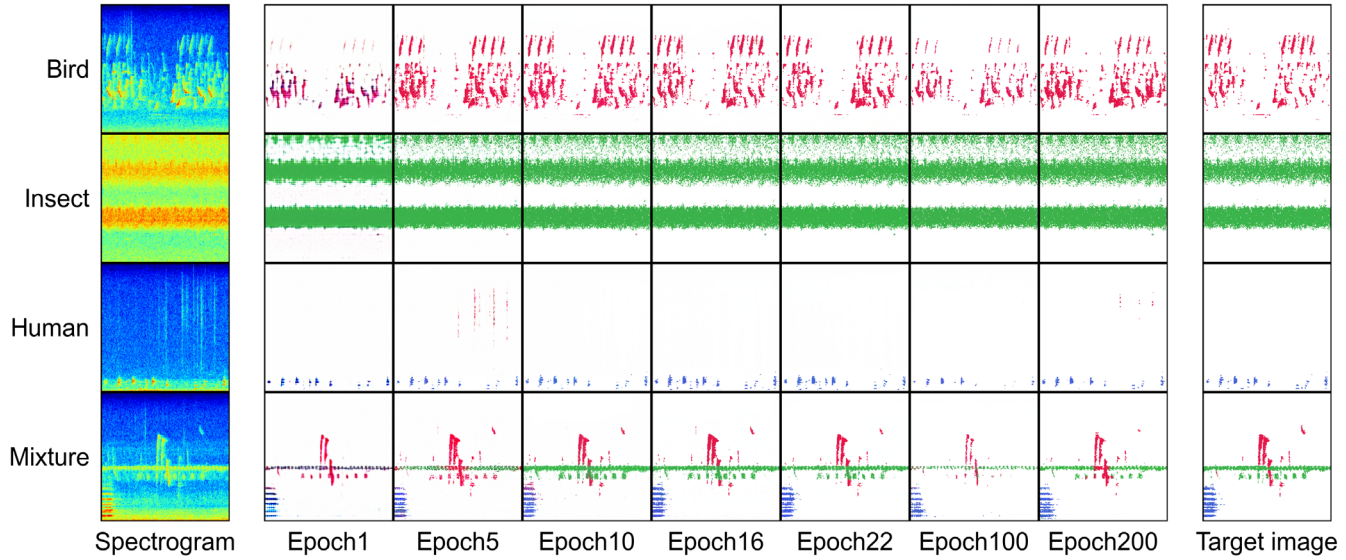

FIGURE 4 Visual comparison of images generated by the community generative adversarial network (GAN) model. From left to right are the original spectrogram, the generated images at different training epochs and the target image.

TABLE 1 Pixel-wise classification precision, recall and F1 score of the species generative adversarial network (GAN) model.

	Precision	Recall	F1 score
Jungle Nightjar	0.87	0.60	0.71
Eurasian Jay	0.89	0.76	0.82
Koklass Pheasant	0.94	0.62	0.75
Oriental Scops Owl	0.92	0.56	0.70
Lesser Cuckoo	0.91	0.56	0.69
Brownish-flanked Bush Warbler	0.91	0.79	0.85
Alström's Warbler	0.89	0.62	0.73
Hartert's Leaf Warbler	0.93	0.78	0.85
Background	0.99	1.00	0.99

Note: The values are calculated from the Yaoluoping National Nature Reserve test set. The pixels of the target images are the true values, and the pixels of the generated images are the predicted values.

TABLE 2 Pixel-wise classification precision, recall and F1 score of the community generative adversarial network (GAN) model.

	Precision	Recall	F1 score
Bird	0.89	0.67	0.77
Insect	0.96	0.92	0.94
Human	0.87	0.54	0.66
Background	0.97	0.99	0.98

Note: The values are calculated from the Yaoluoping National Nature Reserve test set. The pixels of the target images are the true values, and the pixels of the generated images are the predicted values.

These results indicate that all three models achieved high and comparable classification accuracy, with only marginal differences. Notably, the GAN model performed competitively with these well-established deep learning architectures, despite being originally designed for spectrogram generation rather than classification.

3.3.2 | Denoising

Figure 7 compares the GAN-based model's noise reduction performance with two conventional denoising methods: spectral subtraction and Wiener filtering and a non-denoised baseline. Performance was evaluated using the MSE between the denoised outputs and manually cleaned reference spectrograms.

The GAN model achieved the lowest MSE among all methods, indicating superior performance in suppressing background noise while preserving relevant acoustic features. In contrast, spectral subtraction and Wiener filtering exhibited higher residual error, suggesting greater loss of spectrogram detail or incomplete noise removal.

These results highlight the GAN model's effectiveness in complex acoustic environments and demonstrate its suitability for ecoacoustic applications requiring both signal fidelity and noise robustness.

3.3.3 | Sound source separation

As illustrated in Figure 8, the GAN-based model consistently achieved lower MSE values than NMF across all target acoustic sources, including bird, insect and human vocalizations. The performance difference was particularly notable for insect sound separation, where NMF exhibited substantially higher spectral reconstruction errors.

This suggests that the GAN model better preserves fine-grained temporal-spectral structures characteristic of insect signals, which

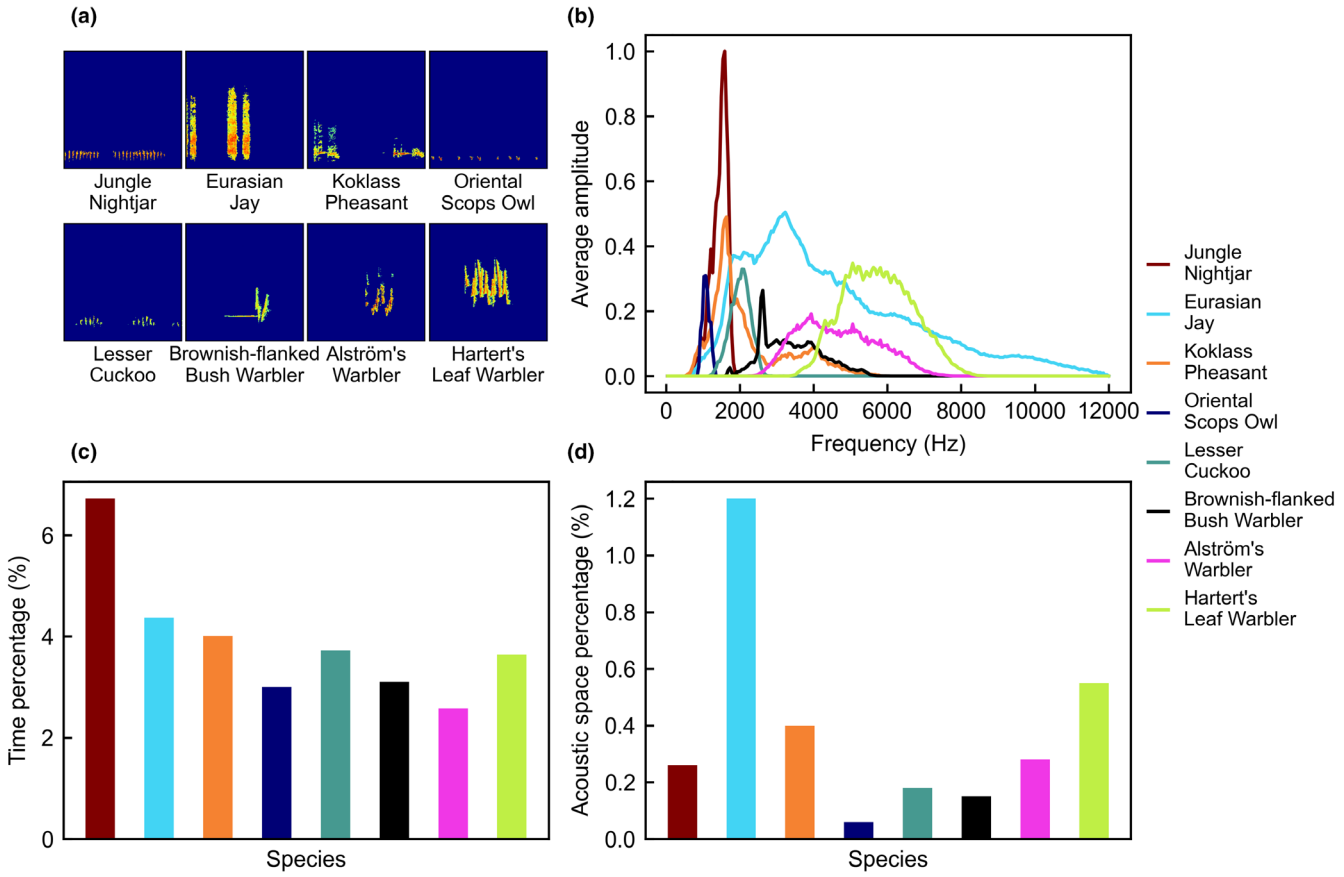

FIGURE 5 Acoustic space analysis of the species generative adversarial network (GAN) model. (a) Species-only spectrograms reconstructed from the GAN output and original spectrograms. (b) Frequency distributions derived from pixel intensities in the spectrograms. (c) Temporal occupation calculated as the total duration of pixels in the spectrograms. (d) Acoustic space occupation estimated by summing all pixels attributed to the spectrograms.

often feature high-frequency content and repetition patterns. The results reinforce the GAN's advantage in handling complex, overlapping acoustic sources, especially in ecologically rich and acoustically dense environments.

4 | DISCUSSION

4.1 | Contributions and key findings

This study presents a novel application of GANs for animal sound identification and acoustic space quantification in complex, natural soundscapes. By generating colour-encoded spectrogram-like images, our framework enables the extraction of ecologically relevant information from complex audio recordings. Two GAN-based models were developed and trained using paired spectrogram datasets: a species-level model targeting eight avian taxa, and a community-level model distinguishing among bird, insect and anthropogenic sounds. Both models successfully identified acoustic sources from mixed recordings and quantified them in terms of temporal activity, frequency range, signal intensity and overall acoustic space occupation, offering fine-resolution insights into soniferous biodiversity.

Compared to conventional classifiers (e.g. ResNet50 and VGG16), our GAN achieved comparable species recognition, while additionally supporting pixel-wise spectrogram translation. Unlike discriminative models, GANs learn the data distribution itself, allowing fine-scale mapping between raw and labelled acoustic representations (Metri & Mamatha, 2021). The model's ability to retain spectro-temporal fidelity enables more precise localization of short duration or narrowband vocalizations in ecoacoustics (LeBien et al., 2020; Malavasi & Farina, 2012).

Our results show that pixel-wise mapping facilitates precise identification of how sound-producing organisms occupy acoustic space. In both species and community models, time-frequency annotations revealed unique acoustic niches across taxa and community types. These distinctions provide valuable information on acoustic resource partitioning, especially under interspecific competition or anthropogenic disturbance. Previous research has shown that animals adjust their vocal behaviour in response to overlapping signals or environmental noise (Chronister et al., 2023; Sueur et al., 2019), and our approach offers a scalable means to quantify such responses.

The community identification of vocal organisms represents the cumulative signals from numerous species, reflecting internal structural changes within a community and the interactions between communities

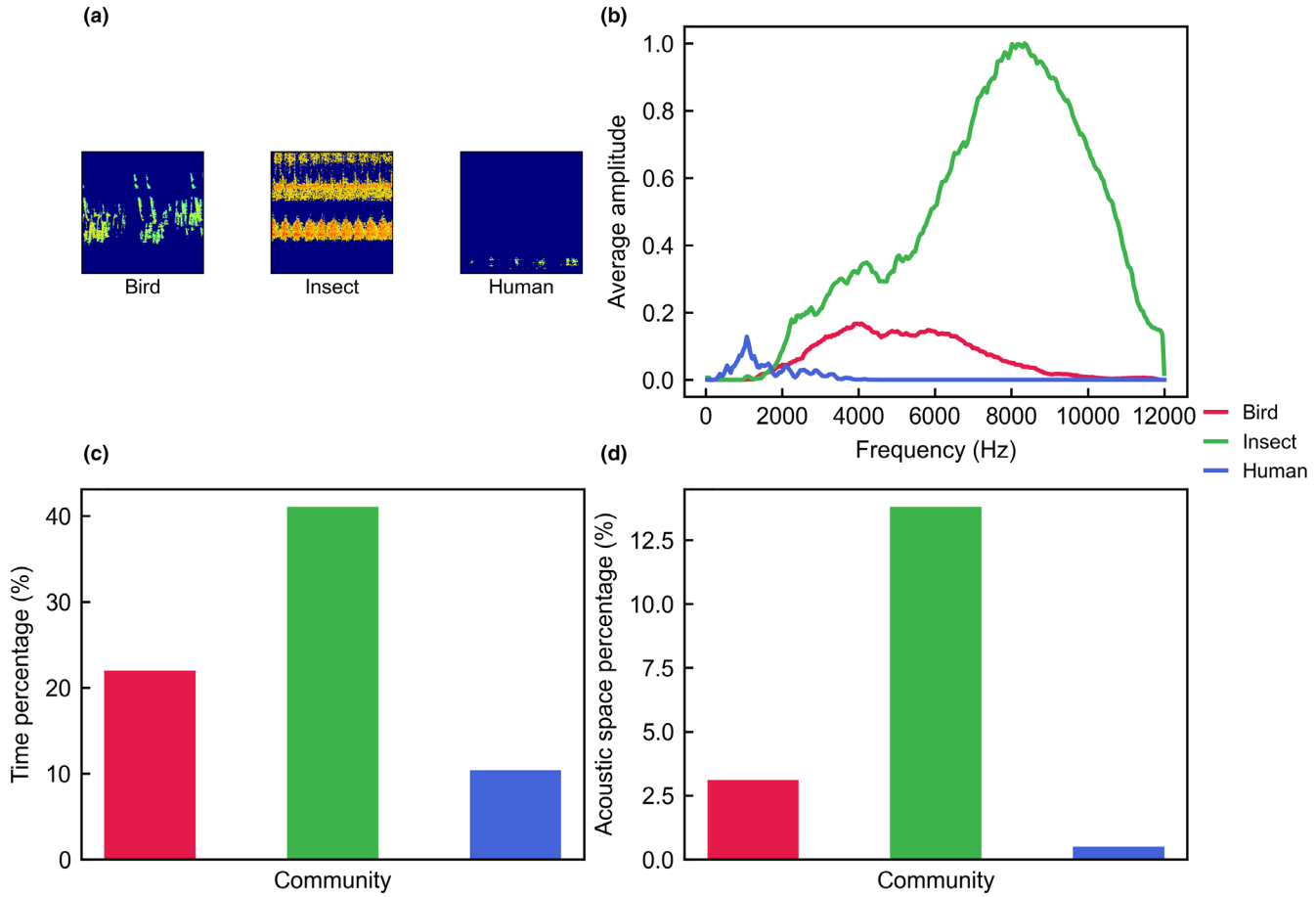

FIGURE 6 Acoustic space analysis of the community generative adversarial network (GAN) model. (a) Community-only spectrograms reconstructed from the GAN output and original spectrograms. (b) Frequency distributions derived from pixel intensities in the spectrograms. (c) Temporal occupation calculated as the total duration of pixels in the spectrograms. (d) Acoustic space occupation estimated by summing all pixels attributed in the spectrograms.

TABLE 3 Comparison of image-level classification precision, recall and F1 score of the species generative adversarial network (GAN) model, ResNet50 and VGG16.

	GAN			ResNet50			VGG16		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
Jungle Nightjar	1.00	0.98	0.99	0.92	1.00	0.96	0.99	1.00	1.00
Eurasian Jay	0.97	0.98	0.98	0.98	0.95	0.96	0.97	0.97	0.97
Koklass Pheasant	0.98	0.95	0.96	0.95	0.96	0.96	0.99	0.95	0.97
Oriental Scops Owl	1.00	0.99	0.99	0.91	0.99	0.95	1.00	0.99	0.99
Lesser Cuckoo	0.93	0.99	0.96	1.00	0.95	0.97	1.00	0.94	0.97
Brownish-flanked Bush Warbler	0.99	0.96	0.97	0.95	0.97	0.96	0.93	1.00	0.96
Alström's Warbler	0.96	0.95	0.95	0.98	0.86	0.91	0.96	0.99	0.98
Hartert's Leaf Warbler	0.97	0.97	0.97	0.91	0.93	0.92	0.97	0.98	0.98
Background	0.67	0.73	0.70	0.63	0.57	0.60	0.86	0.80	0.83

Note: The values are calculated from the Yaoluoping National Nature Reserve test set. True species labels were derived from the spectrograms, and predicted labels correspond to the model outputs.

(Sueur et al., 2019). Community-level analysis further revealed that vocal activity at the population scale, which is aggregated across multiple species, can serve as a proxy for ecological structure. Our model potentially enables exploration of temporal and spectral partition, dominance and niche dynamics within and between acoustic communities, enhancing understanding of biotic interactions in soundscapes.

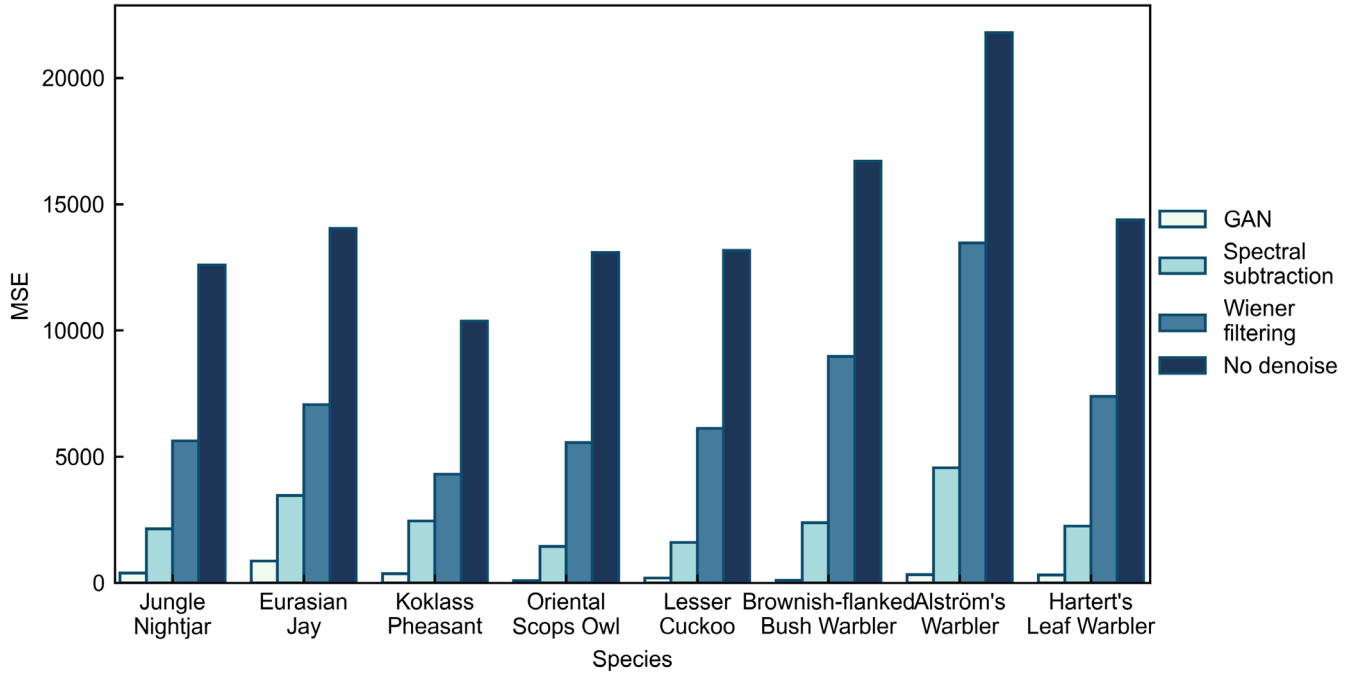

FIGURE 7 Noise reduction comparison between generative adversarial network (GAN) and conventional methods. The mean squared error (MSE) was calculated between the output spectrograms generated by each method and the manually processed reference spectrograms.

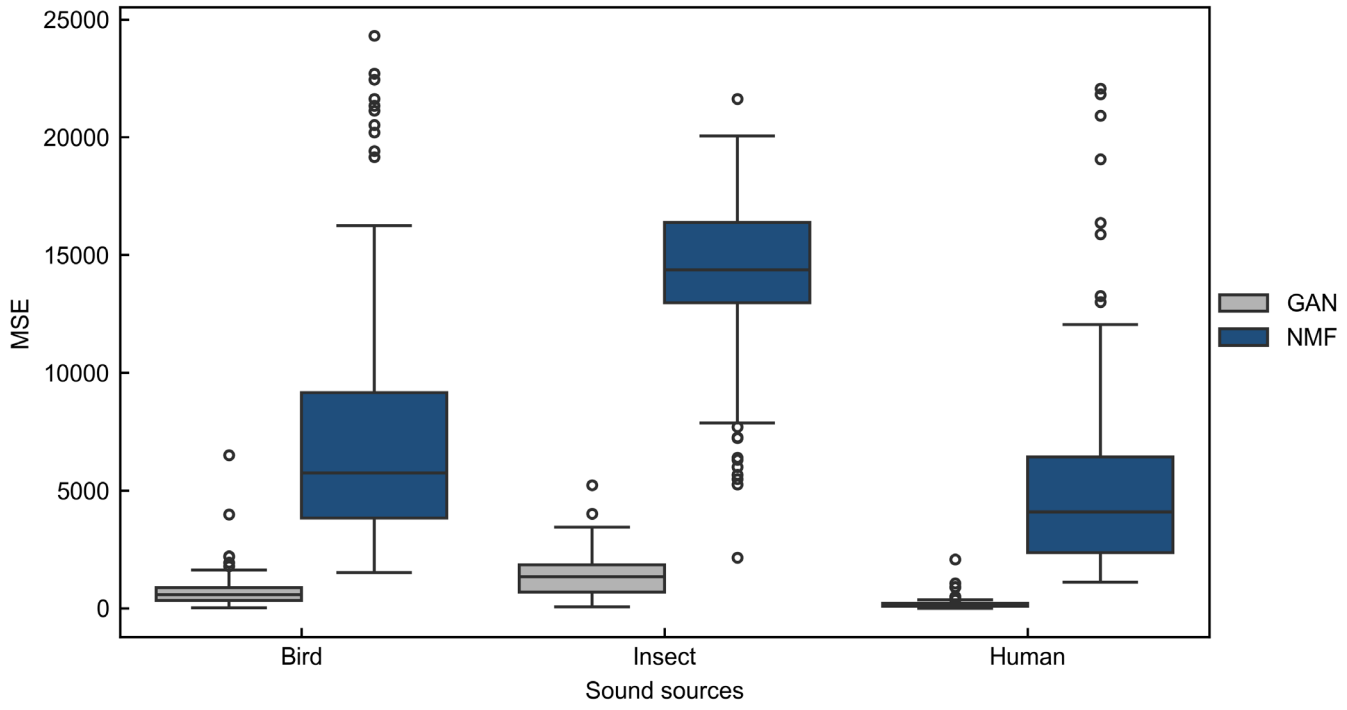

FIGURE 8 Sound source separation comparison between generative adversarial network (GAN) and non-negative matrix factorization (NMF). The mean squared error (MSE) was calculated between the output spectrograms generated by each method and the manually processed reference spectrograms.

Moreover, by isolating sound sources and removing non-target background signals, the GAN model serves as a pre-processing tool. In comparison with conventional denoising algorithms (e.g. Spectral subtraction, Wiener filtering) and unsupervised source separation methods (e.g. NMF), the GAN achieved lower reconstruction error, reinforcing its utility in acoustically cluttered environments.

4.2 | GAN particularities

Training stability continues to be a critical challenge for GAN-based models, as demonstrated both in our experiments and in prior research (Abdelmotaal et al., 2021). In our experiments, both species-level and community-level models performed best around 20 epochs, after which output quality declined and class flipping occurred, indicating unstable class attribution. This instability may arise from adversarial dynamics: an overly strong discriminator causes vanishing gradients, while a weak one leads to low-diversity outputs. Future studies could address this by adopting alternative architectures or adding perceptual loss to improve robustness (de Souza et al., 2023).

A structural characteristic of the current GAN framework is that each pixel is assigned only one class label, which limits its ability to capture overlapping vocalizations. When multiple sources overlap both in time and frequency, the model can only assign a single label per time-frequency bin, inevitably leading to loss or distortion of source information. This distortion becomes more severe as the degree of overlap increases. To address this, false-colour (Towsey et al., 2014) techniques could be applied to pixels based on the overlap degree of different sound sources, or other methods for handling source superposition could be explored.

4.3 | Study limitations

A notable limitation of our model is the imbalance between recall and precision, particularly for species with low vocal activity. This is likely due to pixel-wise class imbalance, as background noise dominates most recordings. Mitigating false negatives through targeted data augmentation or weighted loss functions represents a potential improvement (Sun, Maeda, et al., 2022). In addition, increasing the colour threshold to allow more ambiguous predictions can improve recall across classes. This is conceptually similar to adjusting decision boundaries in probabilistic classifiers (Stowell et al., 2018), and threshold tuning appears theoretically feasible. Such strategies can be adapted to specific research objectives or evaluated according to application needs.

During the GAN model training, the proposed method requires manually annotating each pixel to provide target images, which is significantly time-consuming and labour-intensive. Combining image processing and audio analysis technology to develop efficient automatic labelling tools based on thresholding and automatic selection of contiguous pixels is a recommended future direction.

Despite promising results, the generalizability and transferability of our GAN-based model to broader ecological contexts remain uncertain due to potential data leakage and limited comparisons with globally recognized ecoacoustic models. In particular, data leakage may occur because some training, validation and test samples originate from recordings made at similar times and on the same recorder. This could lead to overly optimistic performance, since the model may have seen very similar data during training (Kaufman et al., 2012). Future work could consider splitting data by recorder or recording date to reduce this risk, which is especially important for long-term regional acoustic monitoring, where spatiotemporal autocorrelation is more likely (Gibb et al., 2018). In addition, expanding the current framework beyond the eight bird species and three acoustic communities analysed here, incorporating additional sound-producing taxa such as amphibians and mammals, would further improve ecological applicability. Moreover, although this study highlights the versatility of GAN-based methods, it only includes comparisons with a few baseline models. Extensive evaluations with state-of-the-art approaches such as BirdNET (Kahl et al., 2021), Perch (Hamer et al., 2023) and MixIT (Denton et al., 2022) would enhance both performance evaluation and theoretical insight into ecoacoustic analysis, representing an important direction for future research.

5 | CONCLUSIONS

Overall, this study demonstrates the potential of generative models—originally developed in the field of computer vision—for application in ecoacoustics and soundscape analysis. By converting noisy, mixed soundscapes into clear, source-specific representations, our GAN-based approach offers a scalable and generalizable tool for ecoacoustic analysis, enabling advances in biodiversity monitoring, community ecology and soundscape-level inference. Leveraging pixel-wise image-to-image translation, GANs enable the identification, extraction and quantification of acoustic signals from diverse biological species and communities, thereby facilitating the representation of both species-level and community-level acoustic structures in complex natural soundscapes, even in the absence of prior taxonomic knowledge.

Future research should aim to expand the ecological applications of generative models by integrating them with large-scale monitoring systems, real-time acoustic sensing networks and multimodal ecological datasets. Continued advances in these techniques have the potential to transform biodiversity assessment, enabling high-resolution, scalable and non-invasive monitoring across extensive spatial and temporal domains. Ultimately, such tools could provide critical insights for conservation planning and support proactive strategies to safeguard global biodiversity under increasing environmental pressures.

AUTHOR CONTRIBUTIONS

Mei Wang was responsible for conceptualization, data curation, formal analysis, investigation, methodology, visualization and writing—original draft. Kevin F. A. Darras contributed to investigation, methodology, supervision and writing—review and editing. Renjie Xue contributed to data curation, investigation and methodology. Fanglin Liu provided conceptualization, investigation, methodology, supervision and writing—review and editing. All authors critically revised the manuscript and approved the final version for publication.

ACKNOWLEDGEMENTS

Field surveys were conducted under the instruction of Jun Chu, to whom we are the most grateful. We would like to thank Muhammad Zahid Sharif for his helpful suggestions on the manuscript. We also thank Jinjuan Mei and Sabah Mushtaq Puswal for their assistance with bird species identification.

FUNDING INFORMATION

This research did not receive any specific grant from funding agencies.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70148.

DATA AVAILABILITY STATEMENT

Data available via the Dryad Digital Repository https://doi.org/10.5061/dryad.vhhmgqp6k (Wang et al., 2025).

ORCID

Mei Wang https://orcid.org/0000-0001-9783-6995
Kevin F. A. Darras https://orcid.org/0000-0002-9013-3784
Fanglin Liu https://orcid.org/0000-0002-8371-6316

REFERENCES

Abdelmotaal, H., Abdou, A. A., Omar, A. F., El-Sebaity, D. M., & Abdelazeem, K. (2021). Pix2pix conditional generative adversarial networks for scheimpflug camera color-coded corneal tomography image generation. *Translational Vision Science & Technology*, *10*(7), 21. https://doi.org/10.1167/tvst.10.7.21

Bahmei, B., Birmingham, E., & Arzanpour, S. (2022). CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification. *IEEE Signal Processing Letters*, *29*, 682–686. https://doi.org/10.1109/LSP.2022.3150258

Chronister, L. M., Rhinehart, T. A., & Kitzes, J. (2023). When birds sing at the same pitch, they avoid singing at the same time. *Ibis*, *165*(3), 1047–1053. https://doi.org/10.1111/ibi.13192

Darras, K. F., Rountree, R. A., Van Wilgenburg, S. L., Cord, A. F., Pitz, F., Chen, Y., Dong, L., Rocquencourt, A., Desjonquères, C., Diaz, P. M., Lin, T. H., Turco, T., Emmerson, L., Bradfer-Lawrence, T., Gasc, A., Marley, S., Salton, M., Schillé, L., Wensveen, P. J., … Wanger, T. C. (2025). Worldwide soundscapes: A synthesis of passive acoustic monitoring across realms. *Global Ecology and Biogeography*, *34*(5), e70021. https://doi.org/10.1111/geb.70021

de Souza, V. L. T., Marques, B. A. D., Batagelo, H. C., & Gois, J. P. (2023). A review on generative adversarial networks for image generation. *Computers & Graphics*, *114*, 13–25. https://doi.org/10.1016/j.cag.2023.05.010

Denton, T., Wisdom, S., & Hershey, J. R. (2022). Improving bird classification with unsupervised sound separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 636–640). IEEE. https://doi.org/10.1109/icassp43922.2022.9747202

Fu, Y., Yu, C., Zhang, Y., Lv, D., Yin, Y., Lu, J., & Lv, D. (2023). Classification of birdsong spectrograms based on DR-ACGAN and dynamic convolution. *Ecological Informatics*, *77*, 102250. https://doi.org/10.1016/j.ecoinf.2023.102250

Gasc, A., Sueur, J., Jiguet, F., Devictor, V., Grandcolas, P., Burrow, C., Depraetere, M., & Pavoine, S. (2013). Assessing biodiversity with sound: Do acoustic diversity indices reflect phylogenetic and functional diversities of bird communities? *Ecological Indicators*, *25*, 279–287. https://doi.org/10.1016/j.ecolind.2012.10.009

Gibb, R., Browning, E., Glover-Kapfer, P., & Jones, K. E. (2018). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, *10*(2), 169–185. https://doi.org/10.1111/2041-210X.13101

Gómez-Gómez, J., Vidaña-Vila, E., & Sevillano, X. (2023). Western Mediterranean Wetland Birds dataset: A new annotated dataset for acoustic bird species classification. *Ecological Informatics*, *75*, 102014. https://doi.org/10.1016/j.ecoinf.2023.102014

Hamer, J., Triantafillou, E., Van Merriënboer, B., Kahl, S., Klinck, H., Denton, T., & Dumoulin, V. (2023). *Birb: A generalization benchmark for information retrieval in bioacoustics*. arXiv preprint arXiv:2312.07439. https://doi.org/10.48550/arXiv.2312.07439

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5967–5976). IEEE. https://doi.org/10.1109/CVPR.2017.632

K. Lisa Yang Center for Conservation Bioacoustics. (2014). *Raven Pro: Interactive sound analysis software* (Version 1.5) [Computer software]. The Cornell Lab of Ornithology. https://ravensoundsoftware.com/

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), Article 15. https://doi.org/10.1145/2382577.2382579

Keen, S. C., Odom, K. J., Webster, M. S., Kohn, G. M., Wright, T. F., & Araya-Salas, M. (2021). A machine learning approach for classifying and quantifying acoustic diversity. *Methods in Ecology and Evolution*, *12*(7), 1213–1225. https://doi.org/10.1111/2041-210X.13599

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., & Aide, T. M. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, *59*, 101113. https://doi.org/10.1016/j.ecoinf.2020.101113

Li, L., Cui, P., Xu, H., Wan, Y., Yong, F., Hou, X., Ma, H., & Yu, L. (2017). A comparative study of bird species diversity in breeding season at Anhui Yaoluoping National Nature Reserve. *Chinese Journal of Wildlife*, *38*(1), 52–62. https://doi.org/10.19711/j.cnki.issn2310-1490.2017.01.009

Lin, T. H., & Tsao, Y. (2020). Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval. *Remote Sensing in Ecology and Conservation*, *6*(3), 236–247. https://doi.org/10.1002/rse2.141

Liutkus, A., Pinel, J., Badeau, R., Girin, L., & Richard, G. (2012). Informed source separation through spectrogram coding and data embedding. *Signal Processing*, *92*(8), 1937–1949. https://doi.org/10.1016/j.sigpro.2011.09.016

Malavasi, R., & Farina, A. (2012). Neighbours' talk: Interspecific choruses among songbirds. *Bioacoustics*, *22*(1), 33–48. https://doi.org/10.1080/09524622.2012.710395

Metri, O., & Mamatha, H. (2021). Image generation using generative adversarial networks. In A. Solanki, A. Nayyar, & M. Naved (Eds.), *Generative adversarial networks for image-to-image translation* (pp. 235–262). Elsevier. https://doi.org/10.1016/B978-0-12-823519-5.00007-5

Mikula, P., Valcu, M., Brumm, H., Bulla, M., Forstmeier, W., Petrusková, T., Kempenaers, B., & Albrecht, T. (2020). A global analysis of song

frequency in passerines provides no support for the acoustic adaptation hypothesis but suggests a role for sexual selection. *Ecology Letters*, *24*(3), 477–486. https://doi.org/10.1111/ele.13662

Molnár, C., Kaplan, F., Roy, P., Pachet, F., Pongrácz, P., Dóka, A., & Miklósi, Á. (2008). Classification of dog barks: A machine learning approach. *Animal Cognition*, *11*(3), 389–400. https://doi.org/10.1007/s10071-007-0129-9

Napier, T., Ahn, E., Allen-Ankins, S., Schwarzkopf, L., & Lee, I. (2024). Advancements in preprocessing, detection and classification techniques for ecoacoustic data: A comprehensive review for large-scale Passive Acoustic Monitoring. *Expert Systems with Applications*, *252*, 124220. https://doi.org/10.1016/j.eswa.2024.124220

Nieto-Mora, D. A., Rodríguez-Buritica, S., Rodríguez-Marín, P., Martínez-Vargaz, J. D., & Isaza-Narváez, C. (2023). Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. *Heliyon*, *9*(10), e20275. https://doi.org/10.1016/j.heliyon.2023.e20275

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., Gage, S. H., & Pieretti, N. (2011). Soundscape ecology: The science of sound in the landscape. *BioScience*, *61*(3), 203–216. https://doi.org/10.1525/bio.2011.61.3.6

Ronneberger, O., Fischer, P., & Brox, T. (2015). Medical image computing and computer-assisted intervention – MICCAI 2015 (Lecture Notes in Computer Science, Vol. 9351). In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *U-Net: Convolutional networks for biomedical image segmentation* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Shailesh, K. R., Kurian, C. P., & Kini, S. G. (2016). Study of color space transformation techniques for converting spectrographs to spectrograms. In *Proceedings of the 2016 IEEE International Conference on Current Trends in Advanced Computing* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCTAC.2016.7567343

Sinha, R., & Rajan, P. (2018). A deep autoencoder approach to bird call enhancement. In *Proceedings of the 2018 IEEE 13th International Conference on Industrial and Information Systems* (pp. 22–26). IEEE. https://doi.org/10.1109/ICIINFS.2018.8721406

Stowell, D. (2022). Computational bioacoustics with deep learning: A review and roadmap. *PeerJ*, *10*, e13152. https://doi.org/10.7717/peerj.13152

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., & Glotin, H. (2018). Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution*, *10*(3), 368–380. https://doi.org/10.1111/2041-210X.13103

Sueur, J., Krause, B., & Farina, A. (2019). Climate change is breaking earth's beat. *Trends in Ecology & Evolution*, *34*(11), 971–973. https://doi.org/10.1016/j.tree.2019.07.014

Sun, Y., Maeda, T. M., Solís-Lemus, C., Pimentel-Alarcón, D., & Buřivalová, Z. (2022). Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation. *Ecological Indicators*, *145*, 109621. https://doi.org/10.1016/j.ecolind.2022.109621

Sun, Y. J., Yen, S. C., & Lin, T. H. (2022). soundscape_IR: A source separation toolbox for exploring acoustic diversity in soundscapes. *Methods in Ecology and Evolution*, *13*(11), 2347–2355. https://doi.org/10.1111/2041-210X.13960

Towsey, M., Truskinger, A., Cottman-Fields, M., & Roe, P. (2018). *Ecoacoustics audio analysis software* (version v18.03.0.41) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.1188744

Towsey, M., Zhang, L., Cottman-Fields, M., Wimmer, J., Zhang, J., & Roe, P. (2014). Visualization of long-duration acoustic recordings of the environment. *Procedia Computer Science*, *29*, 703–712. https://doi.org/10.1016/j.procs.2014.05.063

Ulloa, J. S., Haupert, S., Latorre, J. F., Aubin, T., & Sueur, J. (2021). Scikit-maad: An open-source and modular toolbox for quantitative soundscape analysis in python. *Methods in Ecology and Evolution*, *12*(12), 2334–2340. https://doi.org/10.1111/2041-210X.13711

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(3), 1066–1074. https://doi.org/10.1109/TASL.2006.885253

Wang, M., Darras, K. F., Xue, R., & Liu, F. (2025). Data from: Animal acoustic identification, denoising, and source separation using generative adversarial networks. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.vhhmgqp6k

Wang, M., Mei, J., Darras, K. F., & Liu, F. (2023). VGGish-based detection of biological sound components and their spatio-temporal variations in a subtropical forest in eastern China. *PeerJ*, *11*, e16462. https://doi.org/10.7717/peerj.16462

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (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

Wu, Y., Guo, J., & Liu, G. (2009). Improved discriminative training for generative model. *The Journal of China Universities of Posts and Telecommunications*, *16*(3), 126–130. https://doi.org/10.1016/S1005-8885(08)60238-1

Xie, J., Colonna, J. G., & Zhang, J. (2021). Bioacoustic signal denoising: A review. *Artificial Intelligence Review*, *54*(5), 3575–3597. https://doi.org/10.1007/s10462-020-09932-4

Xie, J., Hu, K., Zhu, M., & Guo, Y. (2020). Data-driven analysis of global research trends in bioacoustics and ecoacoustics from 1991 to 2018. *Ecological Informatics*, *57*, 101068. https://doi.org/10.1016/j.ecoinf.2020.101068

Yip, D. A., Mahon, C. L., MacPhail, A. G., & Bayne, E. M. (2021). Automated classification of avian vocal activity using acoustic indices in regional and heterogeneous datasets. *Methods in Ecology and Evolution*, *12*(4), 707–719. https://doi.org/10.1111/2041-210X.13548

Zhang, L., Huang, H.-N., Yin, L., Li, B.-Q., Wu, D., Liu, H.-R., Li, X.-F., & Xie, Y.-L. (2022). Dolphin vocal sound generation via deep WaveGAN. *Journal of Electronic Science and Technology*, *20*(3), 100171. https://doi.org/10.1016/j.jnlest.2022.100171

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 586–595). IEEE. https://doi.org/10.1109/CVPR.2018.00068

How to cite this article: Wang, M., Darras, K. F. A., Xue, R., & Liu, F. (2025). Animal acoustic identification, denoising and source separation using generative adversarial networks. *Methods in Ecology and Evolution*, *16*, 2472–2486. https://doi.org/10.1111/2041-210X.70148