Comparison of Foundation Model Pre-Training Strategies and Architectures for Urban Garden Recordings

Parmenion Koutsogeorgos[1] *and Aki Härmä*[1]

[1]Department of Advanced Computing Sciences (DACS), Maastricht University, The Netherlands

***Abstract*—Environmental audio recordings captured via passive acoustic monitoring include various sounds such as bird vocalisations, weather phenomena, and human activity. Although abundant and easy to collect, these recordings often contain noise, are location-specific, and lack comprehensive annotations, posing challenges to traditional supervised methods. This paper compares self-supervised pre-training techniques and architectures for developing foundation models to learn transferable feature representations from environmental audio data. The reported experiments use the GardenFiles23 dataset, which consists of two years of stereo recordings and metadata from an urban garden. Pre-training tasks include masked spectrogram reconstruction, in which random patches of mel-spectrogram inputs are masked and the model learns to predict them, and a novel contrastive learning task, in which the model learns to align the two channels of stereo recordings that are masked in a complementary manner, meaning that the masked patches in one channel are unmasked in the other. Two architectures are compared: a Self-Supervised Audio Spectrogram Transformer (SSAST) and a State-Space Model variant (Mamba), which theoretically offers linear-time sequence modelling and improved efficiency. Embeddings are assessed on three downstream tasks: bird detection, time-of-day prediction, and weather metadata prediction. Results indicate that masked reconstruction provides stable convergence and superior bird detection performance, while contrastive learning generates richer embeddings that are beneficial for temporal and weather predictions. Overall, SSAST consistently outperforms Mamba with short input sequences.**

***Index Terms*—Foundation models, environmental audio analysis, bird detection, weather prediction, Mamba**

1. INTRODUCTION

Environmental audio recordings captured non-invasively via passive acoustic monitoring (PAM) encompass a broad spectrum of sounds such as animal vocalisations, weather phenomena, and human activities. Such recordings provide a valuable resource for ecological monitoring, biodiversity assessment, and the study of human impacts on natural habitats. Despite their abundance and ease of acquisition, PAM datasets are inherently noisy and unstructured, featuring overlapping sources, location-specific biases, and class imbalances dictated by the placement of the recording device. Moreover, the manual annotation of large-scale audio archives is labor-intensive, which further hinders the development of models that generalize reliably across diverse environments.

Foundation models (FMs) have been shown to mitigate analogous issues in various domains by learning generalizable feature representations from large unlabeled datasets via self-supervised learning. The main architecture used to create FMs is the Transformer [2], which has been highly effective in tasks including natural language processing [3], computer vision [4] and audio processing [5], [6]. However, Transformers have quadratic computational complexity with respect to the input sequence length and require large amounts of data to train effectively. Recent work has established State-Space Models (SSMs) as a powerful alternative to Transformers [7]–[10]. SSMs are a class of models that can capture long-range dependencies in sequential data, while being asymptotically more computationally efficient than Transformers.

In this paper, we explore the use of the Mamba SSM [9] variant for the analysis of environmental audio data and compare its performance to that of the Transformer on three downstream tasks. Our main contributions are the following:

- We establish a baseline for the development of FMs in the context of environmental audio data analysis using a dedicated dataset.
- We compare masked reconstruction and a novel contrastive learning task as self-supervised pre-training tasks for environmental audio data, showing that both approaches can be effective for different downstream tasks.
- We compare SSM and Transformer-based FMs for environmental audio data analysis, showing that the former offer comparable but generally inferior performance to the latter in this setup.
- We evaluate the learned representations on traditional tasks (bird detection) and novel downstream tasks (temporal and weather condition prediction), exploring the abilities and limitations of FMs in extracting complex and informative features.

2. RELATED WORK

In recent years, large-scale pre-trained FMs have been developed for spectrogram-based audio processing. Most notably, the Audio Spectrogram Transformer (AST) [5] re-purposes a Vision Transformer (ViT) backbone to operate on 2-D spectrogram patches, while the Self-Supervised Audio Spectrogram Transformer (SSAST) trains the same architecture using a joint masked reconstruction and contrastive pre-training recipe that remains a strong baseline in self-supervised spectrogram learning [6].

Regarding SSM-based architectures, the Self-Supervised Audio Mamba (SSAM) model swaps each ViT block for a Mamba block and attains comparable downstream performance with fewer parameters [11]. To mitigate the inherently unidirectional view of 1-D SSMs on 2-D inputs, researchers have been experimenting with multi-directional analysis of spectrogram patch sequences. For example, AudioMamba scans spectrograms along four paths for supervised audio tagging [12], and simpler bidirectional variants including AuM [13], SPMamba [14] and SSAMBA [15] apply forward and backward SSM passes before merging the hidden states.

On the topic of environmental data analysis, dedicated CNNs [16], [17] and masked-prediction Transformers [18] have pushed bird species recognition. Regarding weather prediction and monitoring, small datasets have been assembled for rain-intensity classification via CNNs [19] or for rainfall-rate regression using a Transformer [20]. Wind noise is usually treated as a noise-removal problem rather than an environmental cue.

3. DATASET

The *GardenFiles23* dataset [21], [22] consists of stereo audio recordings collected from a PAM system installed in a residential back garden in the Netherlands. The system uses a two-microphone array

This work is based on the MSc thesis work of Mr. Koutsogeorgos [1]. Thesis and code can be found in https://github.com/pk-470/master-thesis.

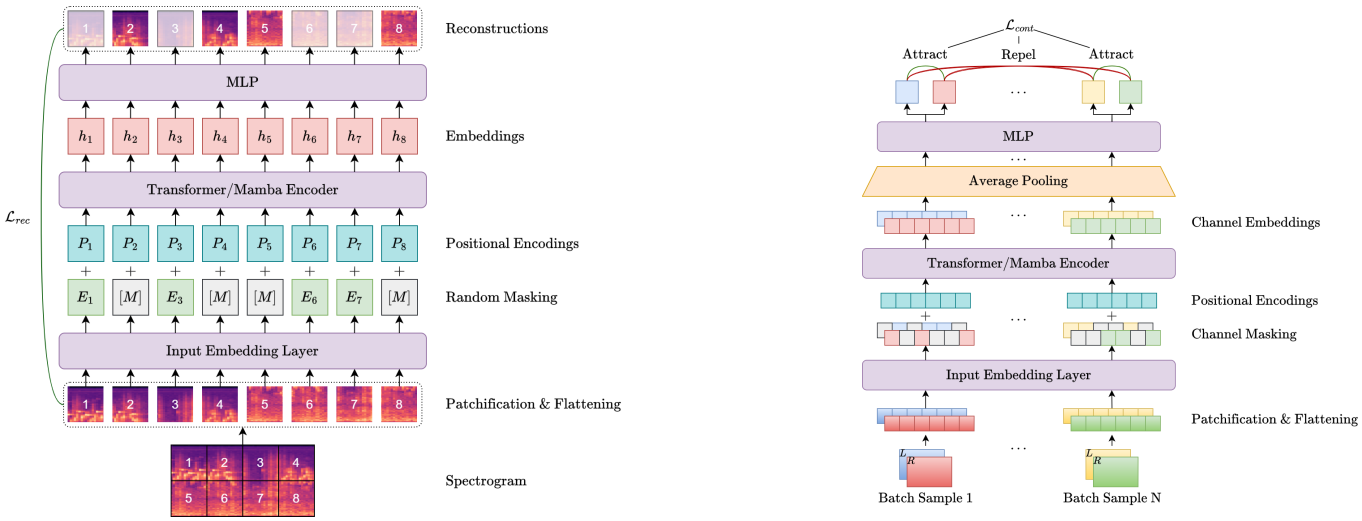


Fig. 1: Overview of pre-training pipelines. Left: Masked reconstruction pre-training. Right: Contrastive pre-training.

which captures 3-second stereo audio clips at a sampling frequency of 48 kHz in response to detected acoustic events above an adaptive background noise spectrum model.

In addition to the audio data, the dataset includes metadata such as precise timestamps of the recordings and environmental context, which is provided by a Froggit WH3000 SE weather station installed adjacent to the microphone array in the garden and connected to the WunderGround service. Moreover, each recording is automatically annotated using two pre-trained deep learning models, namely MIT-AST [5], [23], which provides a wide range of geophonic, biophonic and anthropophonic tags, and BirdNET [16], which is used for bird species classification. All recordings containing human vocalisations, based on MIT-AST detections, were removed from the dataset.

This work uses an expanded version of the GardenFiles23 dataset consisting of 1.3 million samples recorded between August 2023 and March 2025, out of which a random sample of 100,000 is held out for testing and the rest for pre-training and validation. We fine-tune on 550,000 samples, 50,000 of which are unseen during pre-training. The raw waveforms are converted to log-mel-spectrograms using 128 mel bands with a time resolution of 85.33 ms and a hop size of 42.67 ms.

4. METHODOLOGY

We follow a typical FM pipeline, where we first pre-train our models on a large dataset using self-supervised tasks and then fine-tune and evaluate them on a smaller dataset using supervised tasks. The input to all models consists of log-mel-spectrograms of size $T \times F = 65 \times 128$, which are split into 104 non-overlapping patches using a window size of $t \times f = 5 \times 16$. Our encoding pipeline follows the SSAST paradigm, consisting of the following sequential steps:

1) The sequence of patches passes through an embedding layer.
2) A ratio of the input patches is masked using the chosen masking strategy for each pre-training task.
3) Positional encodings are added to the input tokens.
4) Embeddings are extracted through a series of encoder blocks.
5) The extracted embeddings are processed by a task-specific head.

Encoder architectures. Our encoder consists of 12 stacked blocks. We compare two architectures for each encoder block:

- **SSAST** [6]: a model based on the Self-Supervised Audio Spectrogram Transformer architecture. The model is built using standard ViT blocks [4], using 3 attention heads and a hidden state dimension of 192.
- **Bi-directional Mamba (BiMamba)**: a model which replaces the standard ViT blocks with bi-directional Mamba blocks. The block consists of a single forward convolution which is applied to the input sequence, followed by two parallel SSM layers, one for the forward and one for the backward direction, as seen in [13]. We use a linear projection expansion factor of 3, a hidden state dimension of 192, a convolution kernel size of 4 and an SSM hidden state dimension of 24.

Pre-training tasks. The models are pre-trained using the following self-supervised tasks:

- **Masked reconstruction:** a standard self-supervised task where half of the input patches are randomly replaced by a trainable mask token and the model is trained to reconstruct the original input from the unmasked patches. For the reconstruction loss we use the mean absolute error (MAE) and the mean squared error (MSE). The loss is calculated only on the masked patches. The masked reconstruction pre-training pipeline is shown in Fig. 1 (left). In total, we conduct four masked reconstruction experiments: `bimamba-mae`, `bimamba-mse`, `ssast-mae`, `ssast-mse`.
- **Contrastive learning:** a self-supervised task where the model is trained to distinguish between positive and negative pairs of samples using the InfoNCE loss [24]. Instead of creating positive pairs by applying random augmentations on each sample, we use the left and right channels of our chosen dataset's stereo recordings. The two channels are typically highly correlated, but often exhibit differing background noise and recording artifacts, which can be used as a naturally occurring form of weak augmentation of the same signal. Additionally, two channels are randomly masked by random noise sampled from a truncated normal distribution at a ratio of 0.5 in a complementary manner, meaning that if a patch is masked in the left channel, it is unmasked in the right channel and vice versa. The contrastive pre-training pipeline is shown in Fig. 1 (right), and it consists of two experiments: `bimamba-cont` and `ssast-cont`.

Fine-tuning and evaluation. Finally, we fine-tune our models and evaluate them on the following downstream tasks:

- **Bird detection:** a binary classification task where the model is trained to detect the presence or absence of birds in a given audio clip. The task aims to evaluate each model's ability to extract

Table 1: Performance metrics on the downstream tasks. The best performing model for each task is shown in bold.

Model	Bird Detection			Time of Day		Precipitation Rate		Average Wind Speed	
	Precision	Recall	F1	MAE	STD	MAE	STD	MAE	STD
`bimamba-cont`	70.0 (69.5, 70.4)	87.3 (86.9, 87.6)	77.7 (77.3, 78.0)	0.322	0.225	0.193	0.282	0.381	0.318
`bimamba-mae`	78.3 (77.9, 78.7)	90.8 (90.5, 91.1)	84.1 (83.8, 84.4)	0.381	0.259	0.113	0.250	0.334	0.268
`bimamba-mse`	76.5 (76.1, 76.9)	91.0 (90.7, 91.3)	83.1 (82.8, 83.4)	0.284	0.219	0.121	0.250	0.350	0.273
`ssast-cont`	**78.9 (78.5, 79.3)**	91.0 (90.7, 91.3)	**84.5 (84.2, 84.8)**	**0.190**	**0.176**	**0.098**	**0.232**	**0.289**	**0.238**
`ssast-mae`	77.8 (77.4, 78.2)	**92.6 (92.3, 92.9)**	**84.6 (84.3, 84.9)**	0.257	0.204	0.109	0.249	0.338	0.263
`ssast-mse`	76.9 (76.5, 77.3)	92.2 (91.9, 92.5)	83.8 (83.6, 84.1)	0.278	0.217	0.116	0.254	0.350	0.274

features related to bird vocalisations and distinguish them from other sounds in the environment. We assign presence and absence labels by looking for agreement between MIT-AST's bird-related tags and BirdNET's prediction confidence score as follows. Clips where both models agree that no bird is present (MIT-AST detects no bird, BirdNET confidence < 0.5) or where MIT-AST finds a bird but BirdNET is very uncertain (confidence < 0.2) are labelled as "absence", while clips where BirdNET's confidence exceeds the pre-defined thresholds (0.2 when MIT-AST detects a bird and 0.5 otherwise) are labelled as "presence".

- **Temporal metadata prediction:** a regression task where the model is trained to predict the time of recording within the day from a given audio clip, which is treated as a cyclical variable and encoded by sine and cosine pairs. The goal of this task is to assess whether the embeddings extracted from each model can represent complex patterns in the daily activity of humans and animals.
- **Weather metadata prediction:** a regression task in which the model is trained to predict weather metadata associated with each audio clip, including precipitation rate and average wind speed. This task aims to evaluate each model's potential for weather-related feature extraction, which can be useful for monitoring climate change and its impact on the environment.

During fine-tuning the pre-trained encoder is frozen and no masking is applied to the input, while an MLP head is trained specifically for each of the three tasks.

Resources and training time. Training was conducted on an internal computing unit provided by Maastricht University, which is equipped with NVIDIA GeForce RTX 2080 Ti and Quadro RTX 6000 GPUs, as well as on the Snellius computing cluster provided by SURF [25], which includes NVIDIA A100 and NVIDIA H100 GPUs. Training generally took between 2 and 4 days per experiment for the pre-training phase and around 1 day per experiment for the fine-tuning phase.

5. RESULTS AND DISCUSSION

Embedding Analysis. We estimate the alignment and uniformity metrics [26] to assess the quality of the learned embedding spaces. Alignment measures how close similar samples are, while uniformity measures how well the embeddings are distributed on the unit hypersphere. To produce similar pairs for the alignment metric, we apply random transformations both on the waveform level (Gaussian noise, volume gain, pitch shift) as well as the spectrogram level (Gaussian noise, random roll, random masking). Figure 2 shows the alignment and negative (absolute) uniformity metrics for all pre-trained models. The Mamba models achieve lower alignment and lower absolute uniformity, meaning that the embeddings tend to cover a tightly clustered area of the unit sphere, while the SSAST models achieve higher alignment and higher absolute uniformity, indicating spread-out embeddings regardless of similarity.

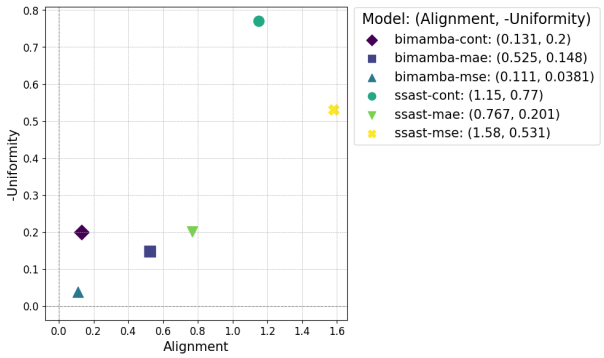


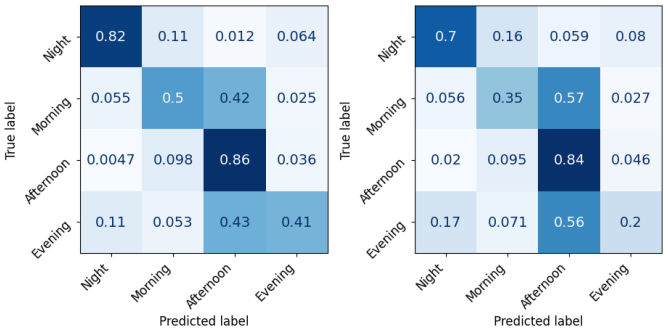
Fig. 2: Alignment and absolute uniformity of the embeddings for all pre-trained models.

Table 1 shows the performance of the pre-trained and fine-tuned models on the three downstream tasks.

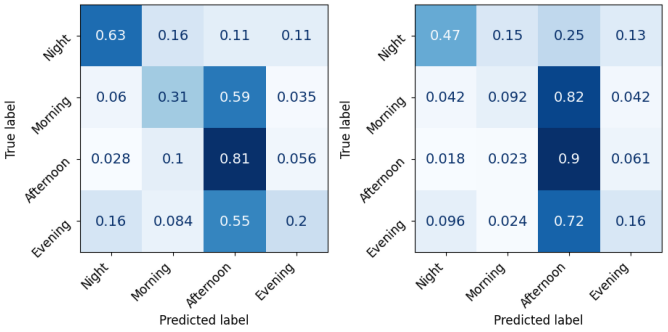
Bird detection. We evaluate the performance of the fine-tuned models on the bird detection task using the standard accuracy, precision, recall and F1 score metrics. Confidence intervals are computed at a level of 95% using 1,000 bootstrap samples. The results suggest `ssast-mae` as the best performing model in terms of recall, while `ssast-cont` achieves the highest precision, with both models achieving similar F1 scores. We observe that the Mamba models generally perform worse than their corresponding SSAST models for each pre-training technique, while the MAE pre-training task outperforms the MSE task across both models.

Time of day prediction. To assess performance on the time-of-day prediction task, we compute the minimum angle difference between the predicted and ground-truth angles. This difference is then normalized to the range $[0, 1)$, corresponding to fractions of a 12-hour period. We calculate the mean absolute error (MAE) and its standard deviation to quantify the prediction error. The `ssast-mae` model achieves the lowest mean absolute error, along with the lowest standard deviation.

We further examine the ability of the models to make predictions that indicate insights about the acoustic patterns that may appear throughout a 24-hour period. To this end, we bin the true and predicted values into 6-hour intervals (Night, Morning, Afternoon, Evening). We then compute the confusion matrix for the true and predicted values, which is shown for the two best performing models of each type in Figure 3. The results show that the models tend to classify the night hours correctly, which is most likely attributed to the decreased acoustic activity during the night-time. All models except for `ssast-cont` tend to predict "Afternoon" for most samples outside the night hours. We hypothesise that all models apart from `ssast-cont` essentially collapse the task to a binary classification problem, with one class corresponding to "inactive" hours, which are predicted during night-time, and the other class to "active" hours, which are predicted near the dataset mean in the opposite direction,

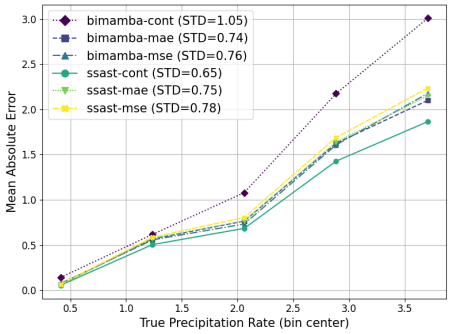


(a) ssast-cont (b) ssast-mae

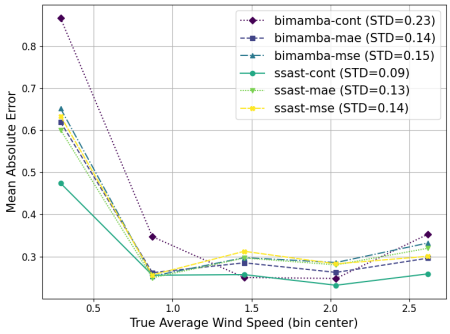


(c) bimamba-mse (d) bimamba-cont

Fig. 3: Confusion matrices for the true and predicted time-of-day values for all pre-trained models.



(a) Precipitation rate performance per model across intervals.



(b) Average wind speed performance per model across intervals.

Fig. 4: Mean absolute error and standard deviation for the precipitation rate and average wind speed tasks across 5 equally sized bins.

i.e., “Afternoon”. Conversely, the `ssast-cont` model is able to learn more complex patterns, with most of its misclassifications occurring between consecutive time intervals of similar acoustic activity. For example, morning and afternoon hours are expected to present with similar acoustic patterns in terms of animal and human activity.

Weather metadata prediction. We calculate the mean and standard deviation of the absolute error between the predicted and true values for the precipitation rate and average wind speed tasks. We find that the `ssast-cont` model achieves the lowest mean absolute error and standard deviation for both tasks, while the SSAST-based models outperform the Bi-directional Mamba-based models for each pre-training technique.

To assess model performance across varying weather conditions, we divide the precipitation rate and average wind speed values into five equally sized bins and compute the mean absolute error for each bin, along with its standard deviation. As shown in Figure 4, the `ssast-cont` model achieves the most consistent performance, with a binned MAE standard deviation of 0.65 for precipitation rate and 0.09 for wind speed. Regarding precipitation, MAE increases monotonically with rain intensity for all models. We attribute this pattern both to the scarcity of heavy rain samples (over 90% of observations record negligible precipitation) and the fact that intense rainfall produces overwhelming broadband, high-energy acoustic noise that obscures finer spectral features. In contrast, wind speed MAE exhibits a sharp decline after the first bin. Under calm conditions, estimation error may be elevated because biotic and anthropogenic sounds dominate the spectrogram, some of which may share similar acoustic patterns with stronger winds (e.g. distant exhaust sounds), or because other weather cues may falsely hint at strong winds (e.g. heavy rain). On the other hand, stronger wind is typically easier to recognise as it presents with more acoustic cues such as rustling of leaves or a distinctive broadband turbulent airflow (“whoosh”) sound.

6. CONCLUSION

Our results highlight distinct strengths and weaknesses of each pre-training strategy and model architecture across downstream tasks. On the one hand, masked reconstruction generally led to better performance in bird detection, likely due to its focus on reconstructing structured spectral content, which encourages the model to capture high-level auditory features such as harmonics and time-frequency patterns. Across models of the same architecture, the MAE loss generally leads to better performance than the MSE loss, with the exception of `bimamba-mae` versus `bimamba-mse` on the temporal prediction task. We attribute this to the fact that MAE is more robust to outliers as it is not dominated by high-energy pixels, pushing the models to learn better reconstructions throughout the entire spectra, which translates to finer feature extraction.

On the other hand, contrastive learning yielded comparable results in the bird detection task, while producing superior results in the temporal and weather prediction tasks. These findings suggest that stereo contrastive learning encourages the extraction of low-level acoustic cues that can carry information about acoustic patterns throughout the day and correlate with weather changes. This effect was most prominent in the Transformer-based models, where contrastive pre-training produced diverse and well-structured embeddings, leading to the best overall performance in regression tasks. In comparison, SSM-based models tended to produce more collapsed representations and were less stable and performative during fine-tuning, particularly under contrastive learning. Despite this, SSMs may still offer benefits for modelling long-range dependencies in future work involving longer input sequences.

REFERENCES

[1] P. Koutsogeorgos, "Towards a Foundation Model for the Analysis of Environmental Audio Data Using the Transformer and Mamba Architectures," Master's thesis, Department of Advanced Computing Sciences, Faculty of Science and Engineering, Maastricht University, 2025.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. ICLR*, 2021.

[5] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[6] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

[7] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *Proc. ICLR*, 2022.

[8] J. T. Smith, A. Warrington, and S. Linderman, "Simplified state space layers for sequence modeling," in *Proc. ICLR*, 2023. [Online]. Available: https://openreview.net/forum?id=Ai8Hw3AXqks

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[10] T. Dao and A. Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," in *Proc. ICML*, 2024.

[11] S. Yadav and Z.-H. Tan, "Audio Mamba: Selective state spaces for self-supervised audio representations," in *Proc. Interspeech*, 09 2024, pp. 552–556.

[12] J. Lin and H. Hu, "Audio Mamba: Pretrained audio state space model for audio tagging," 2024. [Online]. Available: https://arxiv.org/abs/2405.13636

[13] M. H. Erol, A. Senocak, J. Feng, and J. S. Chung, "Audio Mamba: Bidirectional state space model for audio representation learning," *IEEE Signal Process. Lett.*, vol. 31, pp. 2975–2979, 2024.

[14] K. Li, G. Chen, R. Yang, and X. Hu, "SPMamba: State-space model is all you need in speech separation," 2024. [Online]. Available: https://arxiv.org/abs/2404.02063

[15] S. Shams, S. S. Dindar, X. Jiang, and N. Mesgarani, "SSAMBA: Self-supervised audio representation learning with Mamba state space model," *arXiv preprint arXiv:2405.11831*, 2024.

[16] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954121000273

[17] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Scientific Reports*, vol. 13, no. 1, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41598-023-49989-z

[18] G. Vengrovski, M. R. Hulsey-Vincent, M. A. Bemrose, and T. J. Gardner, "TweetyBERT: Automated parsing of birdsong through self-supervised machine learning," *bioRxiv*, 2025. [Online]. Available: https://www.biorxiv.org/content/early/2025/04/10/2025.04.09.648029

[19] R. Avanzato and F. Beritelli, "An innovative acoustic rain gauge based on convolutional neural networks," *Information*, vol. 11, no. 4, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/4/183

[20] M. Wang, M. Chen, Z. Wang, Y. Guo, Y. Wu, W. Zhao, and X. Liu, "Estimating rainfall intensity based on surveillance audio and deep-learning," *Environmental Science and Ecotechnology*, vol. 22, p. 100450, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666498424000644

[21] A. Härmä and E. Nazarenko, "Bird activities in a residential back garden," in *32nd European Signal Processing Conference, EUSIPCO 2024 - Proceedings*, ser. European Signal Processing Conference. United States: IEEE, 2024, pp. 1262–1266. [Online]. Available: https://eusipcolyon.sciencesconf.org/

[22] A. Härmä, "GardenFiles23," 2024. [Online]. Available: https://doi.org/10.34894/HPLUCH

[23] Y. Gong, Y.-A. Chung, and J. Glass, "PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2021.

[24] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[25] SURF, "Snellius: The national supercomputer," 2025, accessed: 2025-09-19. [Online]. Available: https://www.surf.nl/en/services/compute/snellius-the-national-supercomputer

[26] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," *CoRR*, vol. abs/2005.10242, 2020. [Online]. Available: https://arxiv.org/abs/2005.10242