Project Final Report: CS 7643
Transformer-Based Masked Spectrogram Modeling on MBARI Underwater Acoustic Data

Ali Moughnieh* Yohanes Andre Setiawan Seongwon Yun

Georgia Institute of Technology

Abstract

To address the scarcity of labeled underwater acoustic data, we introduce **Sea2Vec** (`S2V`), a transformer-based Masked Autoencoder pre-trained on raw MBARI hydrophone recordings. `S2V` learns robust, domain-specific representations from unlabeled spectrograms, demonstrating remarkable efficiency and performance on a downstream marine mammal classification benchmark. Our 6-layer model achieves a competitive fine-tuning F1 score of 70.7 ± 0.7, while its linear probing F1 of 53.6 ± 1.1 surpasses all baselines. This was accomplished using just 144 hours of underwater raw audio recordings and nearly half the model parameters compared to the 12-layer `Wav2Vec2` and `ImageNet ViT` baselines, which were pre-trained on 960 hours of audio and 1.5 million images, respectively. `S2V` thus demonstrates that lightweight, domain-specific pre-training can unlock effective bioacoustic analysis with significantly less data and compute.

1. Introduction

Detecting and classifying underwater sound, such as ambient noise, vessel traffic, and biological vocalizations, is critical for oceanographers, conservationists, and naval operators; however, large annotated hydrophone datasets are scarce, and manual labeling is time-consuming, expensive, and inconsistent. Self-supervised learning (SSL) trains models to learn audio representations from unlabeled data through proxy tasks that take advantage of the data's natural structure. These representations can be fine-tuned with minimal annotation, allowing for faster deployment of accurate classifiers that can be used to monitor passive underwater acoustics. Several recent works apply SSL directly to audio spectrograms; we'll look at their designs next.

*This document includes a post-submission analysis, available on page 9, conducted independently by Ali Moughnieh to build upon the original project's findings.

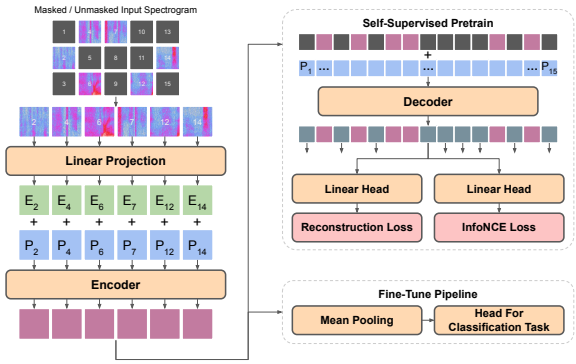


Figure 1: Architecture of the MAE-AST model. Image source: [2]

The Audio Spectrogram Transformer (AST) [5] was the first model to adapt the Vision Transformer (ViT) for audio classification. It treated audio spectrograms as images and used a standard transformer encoder to learn from them. As a fully supervised model, it required a large labeled dataset for pre-training and used a cross-entropy loss for classification.

To reduce the need for large amounts of labeled data, the Self-Supervised Audio Spectrogram Transformer (SSAST) [4] was developed. SSAST introduced a self-supervised approach where it masked random patches of a spectrogram and trained the model's encoder to predict the missing content. It used a joint loss function, combining a reconstruction loss to rebuild the patch with a contrastive loss (InfoNCE) [3] to learn discriminative features. Our work builds on these foundations, adopting the more computationally efficient architecture from the Masked Autoencoder (MAE) framework [7].

Building on this framework, we pre-train a transformer-based masked autoencoder, Sea2Vec, on 144 hours of unlabeled 2 kHz recordings from the MBARI Pacific Sound dataset, sampled across 38 days to cover diverse conditions. The quality of the learned acoustic representations is then evaluated by fine-tuning the model for a downstream classification

task on the Watkins Marine Mammal Sound (WMMS) dataset, using the BEANS split

2. Approach

2.1. Model Architecture

Our work implements the Masked Autoencoding Audio Spectrogram Transformer (MAE-AST) [2]. This model refines the SSAST framework by incorporating an asymmetric encoder-decoder architecture, a design adopted from the original Masked Autoencoder (MAE) [7]. This design significantly improves computational efficiency during pre-training.

The encoder, a standard Vision Transformer, processes only the unmasked, visible patches of the input spectrogram. A second, much lighter transformer decoder then takes the output of the encoder along with the masked tokens to reconstruct the full original spectrogram. This approach ensures that the bulk of the computation is performed on only a small subset of the input data.

2.2. Self-Supervised Pre-training Process

The goal of the pre-training stage is to learn robust representations by reconstructing masked portions of input spectrograms. The process begins by tokenizing the input spectrogram into a sequence of non-overlapping 16x16 patches. Each patch is then mapped to an embedding via a trainable linear projection, and positional embeddings are added to retain spatial information.

A high masking ratio of 75% is then applied, removing the majority of the patch tokens. The small subset of remaining (visible) tokens is processed by the deep transformer encoder. At the decoder stage, the sequence is reconstructed using the encoded patch tokens from the encoder output, along with a shared, learnable "mask" token for every patch that was originally removed. A separate set of positional embeddings is added to this full sequence to inform the lightweight decoder of each token's original location before it reconstructs the spectrogram.

2.3. Loss Functions

The model's learning objective uses a joint loss function that combines a reconstruction loss and a contrastive InfoNCE loss [3]. The objective functions are as follow:

$$\mathcal{L}_{\text{infoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(c_i^\top x_i)}{\sum_{j=1}^{N}\exp(c_i^\top x_j)}\right)$$

Where N is the number of masked patches, x_i is the raw ith patch vector (pixels), and c_i is the corresponding output embedding from the decoder after a projection head. The InfoNCE loss treats the dot-product $c_i^\top x_i$ as the "positive" logit and uses the dot-products $c_i^\top x_j$ for all $j \neq i$ as "negative" logits. This encourages each projected embedding c_i to match its own true patch x_i more closely than any other patch in the spectrogram.

$$\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N}(r_i - x_i)^2$$

The reconstruction loss computes the average squared difference between each reconstructed patch r_i and its original x_i, driving the decoder to produce outputs that match the true pixel values.

The two loss functions are then summed up, where we used weights of $\lambda = 10$ and $\alpha = 1$:

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{recon}} + \alpha\mathcal{L}_{\text{infoNCE}}$$

2.4. Datasets

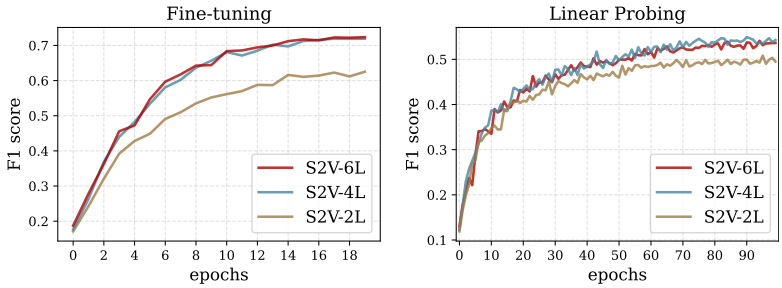
2.4.1 Pre-training Dataset: MBARI

For the self-supervised pre-training stage, we used the Monterey Bay Aquarium Research Institute (MBARI) Pacific Sound dataset [1]. This is an extensive archive of underwater sound recordings from the Pacific Ocean, publicly hosted on the Amazon Open Data Registry. The dataset provides recordings at multiple sampling rates (256, 16, and 2kHz). To ensure a feasible data curation workflow within the project's constraints, we focused exclusively on the 2kHz files. Their smaller size significantly streamlined the manual process of identifying and pre-processing acoustically rich segments. We downloaded over 2000 daily recordings, selected at random, and from these, curated our final 144-hour pre-training dataset. This final dataset was sourced from 38 distinct days, randomly selected from the period between 2015 and 2024 to ensure a diverse temporal sampling.

2.4.2 Downstream Task Dataset: WMMS

For our downstream evaluation, we used the Watkins Marine Mammal Sound (WMMS) dataset [8] as our classification benchmark. The dataset consists of 1017 audio samples distributed across 31 distinct classes of marine mammal sounds. Following the BEANS benchmark protocol, we use the sample split of 6:2:2 train/validation/test sets with stratification [12]. The low number of samples per class presented a significant fine-tuning challenge and increases the risk of overfitting.

For all downstream experiments, we attached a classification head consisting of a single linear layer with no activation function. To mitigate the significant risk of overfitting on the small WMMS dataset, we applied strong regularization: a 50% dropout rate on the classification head and a weight decay of 0.05. Training was limited to 20 epochs as a form of early stopping. We deliberately avoided spectrogram data augmentation to maintain a fair comparison with the waveform-based `Wav2Vec2` baseline.



Model	FT	LP	Params
S2V-6L	**70.7**±0.7	53.6±1.1	50.2M
S2V-4L	68.9±1.6	**54.2**±0.5	36.2M
S2V-2L	63.4±1.4	51.3±0.8	**21.9M**

Figure 2: Performance comparison of our three pre-trained models: S2V-6L, S2V-4L, and S2V-2L. The table summarizes the mean F1 scores for fine-tuning (FT) and linear probing (LP) on the test set.

2.5. Pre-training and Implementation

We used the publicly available MAE-AST codebase [6] for our implementation. The model's data pipeline processes 10-second audio clips sampled at 2 kHz. Each clip is converted into a fixed-size 128x1000 pixel Mel spectrogram. This conversion uses a 1024-point Fast Fourier Transform (`n_fft=1024`), a 512-sample window (256 ms), and a 20-sample hop length (10 ms). The resulting spectrogram is then tokenized into a sequence of non-overlapping 16x16 patches.

During pre-training, we applied a high masking ratio of 75% to these patches. This strategy was chosen for two reasons: it significantly reduces the computational load on the encoder, and high masking ratios have been shown to yield more robust representations for downstream tasks in masked autoencoder models. We trained three different versions of our S2V model, varying only the encoder depth (2, 4, and 6 transformer layers) while using a lightweight single-layer decoder for all models. All pre-training was conducted on a single NVIDIA RTX 2060 GPU. We used the Adam optimizer with a learning rate of $1e^{-4}$ and a weight decay of 0.01, using gradient accumulation to achieve an effective batch size of 386. The pre-training loss curves for all models converged smoothly, as shown in Appendix A.

2.6. Evaluation Protocol

We assessed the performance of our pre-trained models on the WMMS classification task. We compared them against an identical architecture trained from scratch as well as two strong, general-purpose baselines:

- **`ImageNet ViT`:** A 12-layer Vision Transformer [10] pre-trained on the large-scale ImageNet dataset (14 million images). This tests the effectiveness of transfer learning from a massive, out-of-domain visual dataset.
- **`Wav2Vec2`:** A state-of-the-art audio model [9] with 4 convolutional layers followed by 12 transformer layers, pre-trained on 960 hours of general audio. This serves as a powerful self-supervised audio baseline.

To measure the benefit of our pre-training strategy, we evaluated our three S2V models of varying depths, S2V-2L, S2V-4L, and S2V-6L (hereafter 2L, 4L, and 6L), using the methods detailed below:

Fine-tuning: The entire pre-trained encoder was trained on the downstream task with a lightweight classification head. In this protocol, we used a differential learning rate strategy: the backbone used a learning rate of $1e^{-5}$, while the classification head used $1e^{-4}$. The learning rates followed a cosine decay schedule with an initial linear warm-up period, and models were trained for 20 epochs. Our baseline comparisons, Scratch and Scratch+, also followed this fine-tuning protocol. Scratch used the same learning rates as our pre-trained models, while Scratch+ used a more aggressive backbone learning rate of $1e^{-4}$ to see if optimized hyperparameter tuning on a randomly initialized model could match the performance gained from pre-training.

Linear Probing: Only the classification head was trained while the pre-trained encoder weights were kept frozen, a method that specifically tests the raw quality of the learned features. The head used the exact same hyperparameters as in fine-tuning: a learning rate of $1e^{-4}$ managed by the same cosine decay scheduler. As only the head was being trained, we extended the training duration to 100 epochs to thoroughly test the quality of the frozen features over a longer period.

All results are reported as the mean F1 score over five runs with different random seeds (specifically: 42, 1337, 2025, 31415, and 27182). The performance curves in our plots track the validation set score, while the scores reported in our tables are on the held-out test set.

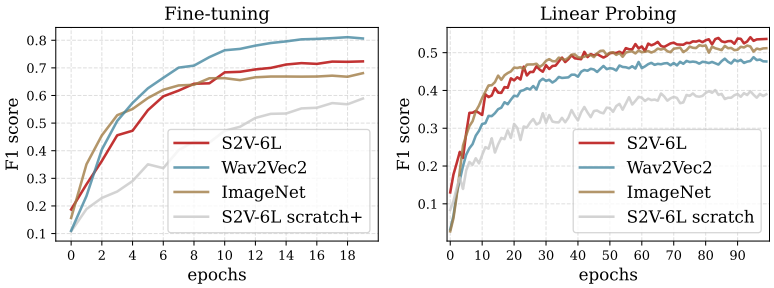
3. Experiments and Results

3.1. Effect of Pre-training Model Depth

A comparison of our three pre-trained models shows a clear trend related to model depth (Figure 2).

The 2L model consistently underperforms, while the 4L and 6L models achieve statistically comparable results. Their fine-tuning F1 scores (68.9 ± 1.6 and 70.7 ± 0.7) and linear probing scores (54.2 ± 0.5 and 53.6 ± 1.1) are similar, considering their standard deviations.

The performance difference between the 2L and



Model	FT	LP	Params
S2V-6L	70.7±0.7	**53.6**±1.1	**50.2M**
Wav2Vec2	**77.5**±5.2	42.6±0.5	94.4M
ImageNet	67.6±1.6	50.3±1.3	85.8M
S2V-6L Sc+	56.2±1.5	40.7±1.2	**50.2M**

Figure 3: Performance of our best model (S2V-6L) compared against Wav2Vec2 and ImageNet ViT baselines. The table summarizes the resulting mean F1 scores for fine-tuning and linear probing on the test set.

the larger models is not surprising, as depth is usually needed to learn more complex relationships. However, the performance plateau between the 4L and 6L models could be attributed to the limited size of the pre-training dataset. Raghu et al. [11] found that for larger models, larger datasets can help in learning high quality intermediate representations. If the dataset is not sufficiently large, it is likely to represent a bottleneck for learning high quality features. Refer to Appendix B for full fine-tuning comparison curves.

3.2. Comparison with Baselines

Our analysis compares the performance of our best model, 6L, against two powerful, general-purpose baselines and an identical architecture trained from scratch. The results, presented in Figure 3, highlight the profound impact of our domain-specific pre-training strategy.

The value of pre-training is immediately evident when comparing the 6L model to its counterpart trained from scratch. The pre-trained model shows a greater performance gain, outperforming the Scratch+ model by approximately 14.5 points in the fine-tuning F1 score (70.7 ± 0.7 vs. 56.2 ± 1.5) and 12.9 points in the linear probing score (53.6 ± 1.1 vs. 40.7 ± 1.2). The validation curves in Figure 3 clearly illustrate this gap, confirming that our self-supervised approach successfully learned robust and meaningful representations.

When the comparison comes to the other pre-trained baselines, the Wav2Vec2 model achieves the highest fine-tuning F1 score (77.5 ± 5.2). This is not surprising, as it is a large model pre-trained on an extensive 960 hours of audio data. Its success underscores the importance of the training modality; even though it was trained on general audio, its inherent understanding of audio waveforms provides a stronger foundation for this downstream task than the ImageNet ViT model, which was pre-trained on an entirely different modality (images). However, our 6L model still demonstrates a clear advantage of domain specificity, outperforming the larger ImageNet ViT (70.7 ± 0.7 vs. 67.6 ± 1.6) despite being pre-trained on a fraction of the data and having nearly half the parameters.

The advantage of our domain-specific approach becomes even more stark in the linear probing evaluation. This test freezes the backbone weights to assess the raw quality of the learned features. Here, 6L emerges as the clear winner with an F1 score of 53.6 ± 1.1, surpassing both the ImageNet ViT and Wav2Vec2 baselines. Notably, the Wav2Vec2 model's performance collapses in this setting to an F1 score of 42.6 ± 0.5, which is much closer to our Scratch model (40.7 ± 1.2) than to the other pre-trained models. While Wav2Vec2's representations are effective when fully fine-tuned, they are not as readily transferable to this specific downstream task without adaptation. The superior performance of our 6L model, despite its disadvantages in model size and pre-training data volume, is a powerful testament to the value of domain-specific pre-training for learning highly effective and transferable features.

Beyond these quantitative metrics, Figure 4 provides a qualitative validation, showing the 6L model's ability to reconstruct "plausible" spectrograms even from highly masked inputs.

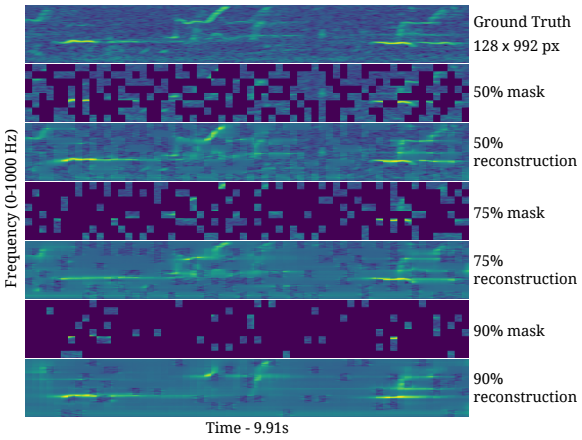


Figure 4: Reconstructions of a spectrogram by the S2V-6L model at 50%, 75%, and 90% masking ratios. The model demonstrates its ability to reconstruct plausible acoustic patterns from highly incomplete inputs. Results are in line with examples shown in [7]

Limitations

While this project successfully demonstrates the effectiveness of domain-specific pre-training, it is important to acknowledge its limitations.

Manual Data Curation and Potential Bias: The most significant limitation stems from the manual data curation process. Our 144-hour pre-training dataset was curated by manually downloading 24-hour recordings and visually inspecting their spectrograms in `Audacity` to identify segments with acoustic signals. This process was extremely time-consuming, which led to a practical focus on files with high signal activity. This likely introduced a seasonal bias into the dataset, as periods of high acoustic activity often correspond to specific marine events like migration or mating seasons. Consequently, the resulting dataset may be unbalanced and over-represent a limited number of species, potentially impacting the generalizability of the learned features.

Geographical Bias: This is related to the previous point. Our model was pre-trained on data collected from one geographical location, which potentially limits the diversity of the data.

Computational Constraints: All pre-training was conducted on a single NVIDIA RTX 2060 GPU with 6 GB of VRAM. This hardware limited the feasible batch size, necessitating the use of gradient accumulation to achieve a larger effective batch size for stable training.

Future Work

Building on the promising results of this study, several avenues for future work could be explored to address the current limitations and further advance this approach.

Automated Data Curation: A key direction is the replacement of the manual curation process with an automated pipeline to build a larger, more diverse, and balanced pre-training dataset. A recent paper by Hummel et al. [13] developed a process that uses a pre-trained model to extract embeddings from raw audio, applies hierarchical k-means clustering to identify acoustic patterns, and then performs balanced sampling. This approach could mitigate the selection bias present in our current dataset.

Scaling Models and Data: This research could also be extended by scaling both the models and the data. With an automated curation pipeline, an expanded dataset including thousands of hours of recordings from diverse hydrophones could be used to pre-train deeper S2V models (e.g., 8 and 12-layer architectures), which would likely lead to significant performance gains.

Broader Downstream Evaluation: Finally, to further validate the quality of the learned representations, the S2V models could be evaluated on a broader range of downstream bioacoustic and underwater acoustic tasks, such as vessel classification, noise-level monitoring, and detection of specific geophysical events.

Conclusion

In this work, we addressed the challenge of data scarcity in underwater bioacoustics by introducing `Sea2Vec` (S2V), a transformer-based masked autoencoder pre-trained on a small, 144-hour dataset of unlabeled hydrophone recordings. Our results demonstrate that S2V learns highly effective, domain-specific representations. Despite the model's smaller size and significantly smaller pre-training dataset, S2V outperformed a general-purpose `ImageNet ViT` baseline in fine-tuning and, more notably, surpassed both the `ImageNet ViT` and the powerful `Wav2Vec2` baselines in linear probing evaluations. This latter result highlights the superior quality and transferability of the features learned through domain-specific self-supervision.

The success of `S2V` provides a compelling case for the value of smaller, domain-focused pre-training in specialized scientific fields. Our findings suggest that access to massive, general-purpose datasets and large-scale computational resources are not prerequisites for developing high-performing models. Instead, this work demonstrates a viable, resource-efficient pathway for achieving strong results in niche domains like underwater acoustics. Ultimately, **Sea2Vec** serves as a strong proof-of-concept that establishes a solid foundation and points toward clear future research directions to further unlock the potential of self-supervised learning in the marine sciences.

References

[1] Registry of Open Data on AWS, "Pacific Ocean Sound Recordings," 2025. Accessed: June, 2025. https://registry.opendata.aws/pacific-sound 2

[2] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked Autoencoding Audio Spectrogram Transformer," in *Proc. ICASSP*, 2022. https://arxiv.org/abs/2203.16691 1, 2, 9

[3] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," 2019. https://arxiv.org/abs/1807.03748 1, 2

[4] Y. Gong, C.-I. J. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," in *Proc. AAAI*, 2022. https://arxiv.org/abs/2110.09784 1

[5] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021. https://arxiv.org/abs/2104.01778 1

[6] A. Baade, "MAE-AST-Public: Code for Masked Autoencoding Audio Spectrogram Transformer," GitHub, 2022. https://github.com/AlanBaade/MAE-AST-Public 3

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, "Masked Autoencoders Are Scalable Vision Learners," 2021. https://arxiv.org/abs/2111.06377 1, 2, 4

[8] Woods Hole Oceanographic Institution, "Watkins Marine Mammal Sound Database," Woods Hole, MA, USA. Accessed: June, 2025. https://whoicf2.whoi.edu/science/B/whalesounds/index.cfm 2

[9] Facebook, "Wav2Vec2-Base," Hugging Face Model Hub, 2020. https://huggingface.co/facebook/wav2vec2-base 3

[10] Google, "Vision Transformer (ViT-Base-Patch16-224)," Hugging Face Model Hub, 2020. https://huggingface.co/google/vit-base-patch16-224 3

[11] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?," in *Proc. NeurIPS*, 2021. https://arxiv.org/abs/2108.08810 4

[12] M. Hagiwara, B. Hoffman, J.-Y. Liu, M. Cusimano, F. Effenberger, and K. Zacarian, "BEANS: The Benchmark of Animal Sounds," Earth Species Project, 2022. https://arxiv.org/abs/2210.12300 2

[13] H. I. Hummel, S. Bhulai, B. Ghani, and R. van der Mei, "Automated data curation for self-supervised learning in underwater acoustic analysis," in *Proc. Forum Acusticum*, 2025. https://arxiv.org/abs/2505.20066 5

APPENDIX A: Pre-training Loss Curves

The following figures show the pre-training loss curves for each model. The total loss is calculated using the formula:

$$\mathcal{L}_{\text{total}} = 10 \times \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{infoNCE}}$$

All pre-training was conducted on a single NVIDIA RTX 2060 GPU with 6 GB of VRAM. The GPU's memory limited the batch size for a single forward pass to approximately 48 spectrograms.

To simulate a larger batch for more stable gradients, we used gradient accumulation with an update frequency of 8. This technique averages the gradients over these 8 passes to compute a single update. This resulted in an effective batch size of approximately 386 samples per optimizer update (∼48 samples × 8 passes).

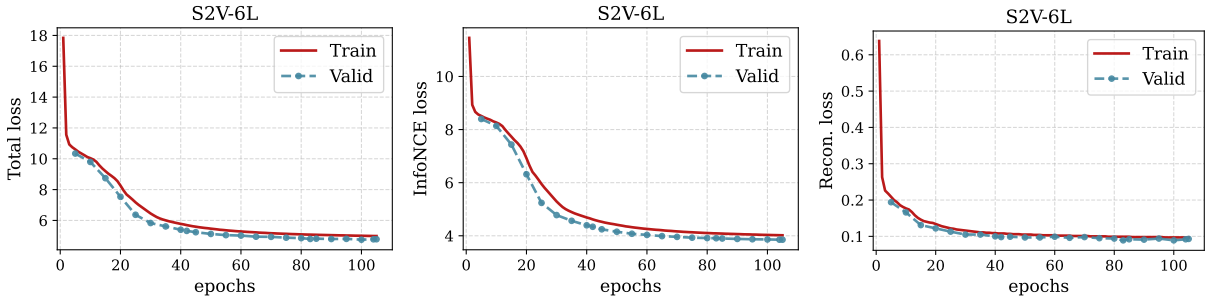


Figure 5: Loss curves for the `S2V-6L` model. Training time: 21.07 hours. Number of epochs: 105.

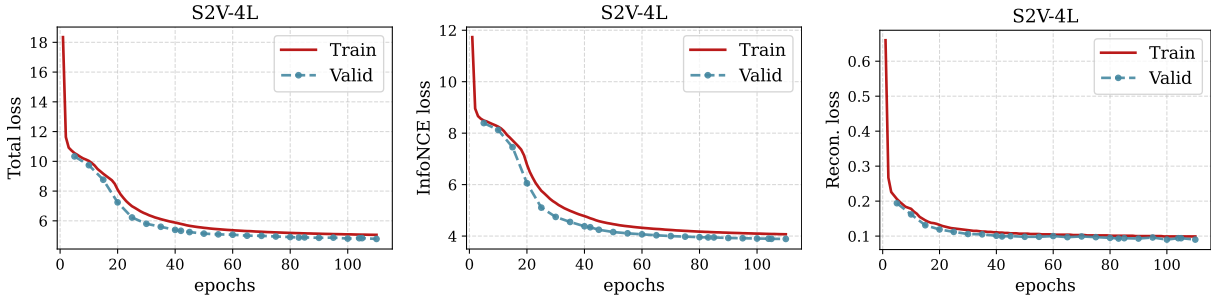


Figure 6: Loss curves for the `S2V-4L` model. Training time: 19.28 hours. Number of epochs: 110.

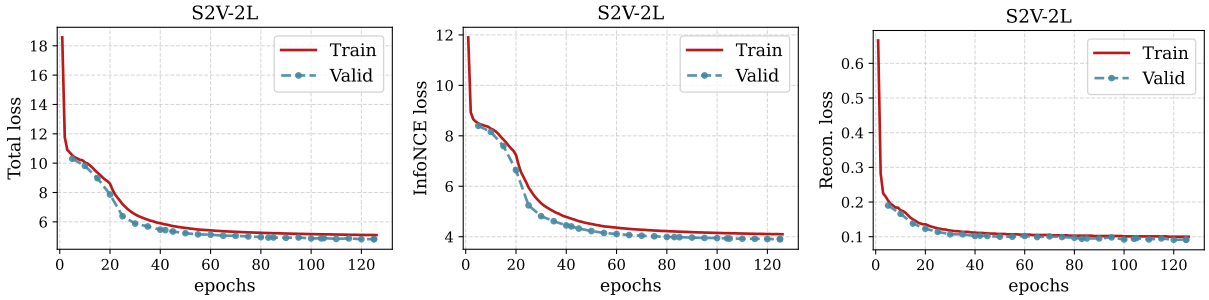


Figure 7: Loss curves for the `S2V-2L` model. Training time: 22.49 hours. Number of epochs: 126.

APPENDIX B: Ablation Study on `S2V` Model Depth

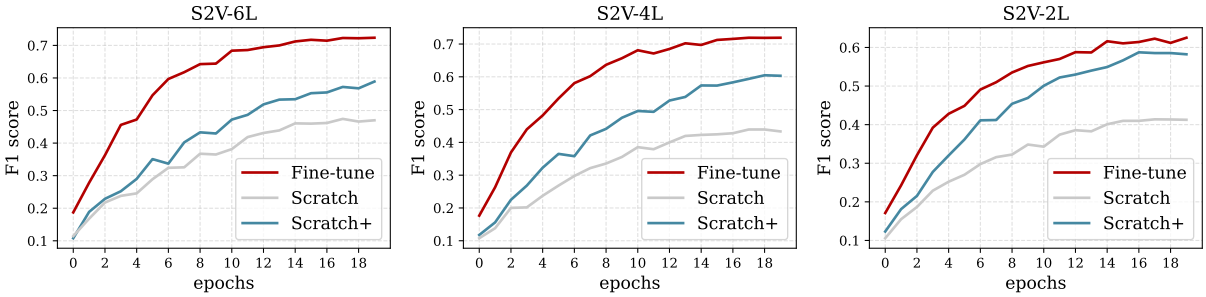


Figure 8: Fine-tuning performance (F1 score) on the downstream task for `S2V` models with 6, 4, and 2 layers. Each plot compares the pre-trained model (`Fine-tune`) against the same architecture trained from scratch. `Scratch` with backbone lr $= 1e^{-5}$ and `Scratch+` with backbone lr $= 1e^{-4}$. All models had a linear classifier with lr $= 1e^{-4}$. The results demonstrate a clear and consistent performance benefit from pre-training across all tested model depths.

Post-Submission Analysis: The Critical Role of Data Diversity and Spectral Range

This follow-up analysis was conducted by Ali Moughnieh after the submission of the original project to investigate pathways for model improvement.

Introduction and New Benchmarks

This analysis introduces two key changes: a new `S2V` model trained on more diverse data, and the inclusion of a new state-of-the-art baseline, `MAE-AST` [2].

The `MAE-AST` model is a 12-layer transformer pre-trained on over 5000 hours of general audio. It was included as it shares the same architecture as `S2V` and provides a more direct benchmark. While a 6-layer version would be an ideal comparison, its checkpoint was not publicly released. However, the original `MAE-AST` paper reports comparable performance between their 6L and 12L variants, making the 12L model a reasonable proxy for state-of-the-art performance.

Revisiting Data Scaling: A Surprising Initial Result

Following the initial project, an investigation was launched to improve the `S2V` model's performance. The first hypothesis was that simply increasing the quantity of in-domain data would improve representations. To test this, the 2kHz MBARI pre-training dataset was expanded from 144 hours to 778 hours—a more than five-fold increase. A new `S2V-6L` model was pre-trained on this larger dataset.

Surprisingly, as shown in Figure 9, this massive increase in data quantity resulted in no discernible performance improvement. The model trained on 778 hours performed identically to the one trained on the original 144 hours. This finding yields two critical insights:

- **Diversity is more crucial than quantity.** The model had likely extracted all unique patterns available in the 2kHz MBARI data from the smaller 144-hour dataset. Adding more data of the same kind offered no new information, suggesting that data diversity is the real bottleneck.
- The Masked Autoencoder framework is an exceptional regularizer. These results confirm our earlier discussion; the high masking ratio prevents the model from simply memorizing the training set, even when the dataset size is very small.

The Impact of Diversity and Spectral Range

Based on the previous result, the strategy shifted from scaling quantity to increasing diversity and spectral richness. A new pre-training dataset of approximately 700 hours was curated using higher-fidelity 16kHz recordings. This dataset included not only data from MBARI but also from three new, geographically distinct locations, introducing novel acoustic signatures from different species (e.g., dolphins, seals) and environments.

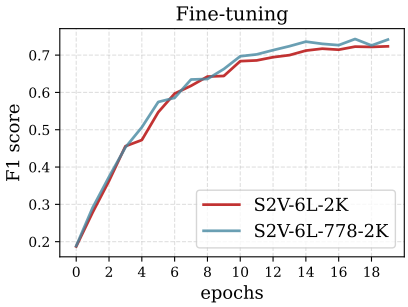


Figure 9: Fine-tuning performance of the `S2V-6L` model pre-trained on 144 hours vs. 778 hours of 2kHz MBARI data. Performance is identical, showing no benefit from increased data quantity alone.

The new `S2V-6L-16K` model pre-trained on this dataset showed a remarkable jump in performance (Figures 10 and 11). It now clearly outperforms the `Wav2Vec2` baseline in the full fine-tuning setting and further widens its lead in the linear probing evaluation. This confirms that the performance gains were primarily driven by the richer spectral information (16kHz vs. 2kHz) and the increased diversity of acoustic sources, not simply data volume.

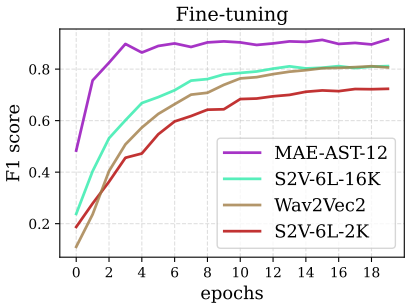


Figure 10: The new `S2V-6L-16K` model's fine-tuning performance compared to baselines. The inclusion of diverse, high-resolution data enabled it to outperform `Wav2Vec2`.

Architectural Insights and Future Directions

While `MAE-AST`'s top performance can be attributed to its massive and diverse pre-training data, the results also suggest a more fundamental insight. Both `S2V` and `MAE-AST` treat audio spectrograms as images and use a Vision Transformer architecture. In contrast, `Wav2Vec2` applies a CNN to the 1D raw waveform before its transformer layers.

The superior linear probing performance of both `S2V-6L-16K` and `MAE-AST` over `Wav2Vec2` suggests that for this bioacoustics task, the ViT-on-spectrogram approach learns more effective and readily transferable features.

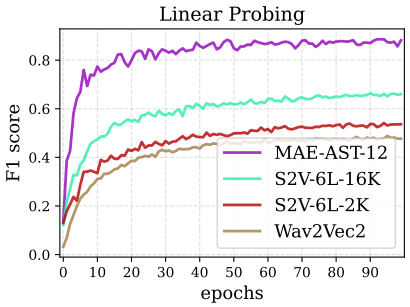


Figure 11: Linear probing results for the `S2V-6L-16K` model, showing a significant improvement in raw feature quality over all other models except the much larger MAE-AST.

This points to a clear path forward: a data-centric strategy. To create a state-of-the-art specialist model, future work should focus on curating a large-scale pre-training dataset that is not just large, but also diverse and balanced across a wide range of global geographies, species, and acoustic conditions.

Team Contribution

Student Name	Contributed Aspect	Details
Ali Moughnieh (903952855)	Pre-training Data Curation	Downloaded over 2000 daily recordings, randomly sampled 103 days, and curated 144 hours of acoustically rich periods from 38 days.
	Pre-training Data Preprocessing	Used Audacity, applied various type of spectral filters to flatten background noise, manually removed high-amplitude outliers and normalized recordings.
	`S2V` Model Pre-training	Modified MAE-AST codebase, adapted hyperparameters for the dataset and GPU constraints, implemented gradient accumulation, and pre-trained the three `S2V` models until loss plateau.
	Downstream Task Experiments	Optimized the initial downstream pipeline by simplifying the classifier head, adding stronger regularization (dropout, weight decay), and implementing separate learning rates for the backbone and head.
	Experiment Design	Defined the experimental setup, including training protocols such as budgeted epochs for fine-tuning and longer runs for linear probing. Ran over 90 experiments across all seeds, models, and baselines.
	Figures and Analysis	Produced all figures, graphs, and tables for the final report.
	Final Report Authorship	Adapted proposal and wrote the abstract, introduction, approach, evaluation protocol, experiments & results, limitations, future work, and conclusion.
Yohanes A. Setiawan (903964616)	Pre-training Data Pipeline	Created the audio preprocessing pipeline for the pre-training task on the MBARI dataset.
	Initial Downstream Task Implementation	Created the initial downstream pipeline, including dataset loading and pre-processing, pre-trained backbone and classifier-head definition, training loop with early stopping, class-imbalance handling, and evaluation-metric computation (later optimized and perfected by Ali Moughnieh).
Seongwon Yun (903618548)	GitHub Management & Optimization	Organized repository and optimized pipelines.