# AudioSphere: Towards Self-Supervised Spatial Audio Representation Models

Goksenin Yuksel, Marcel van Gerven, Kiki van der Heijden
https://arxiv.org/abs/2611.09275

Radboud Universiteit

# OVERVIEW

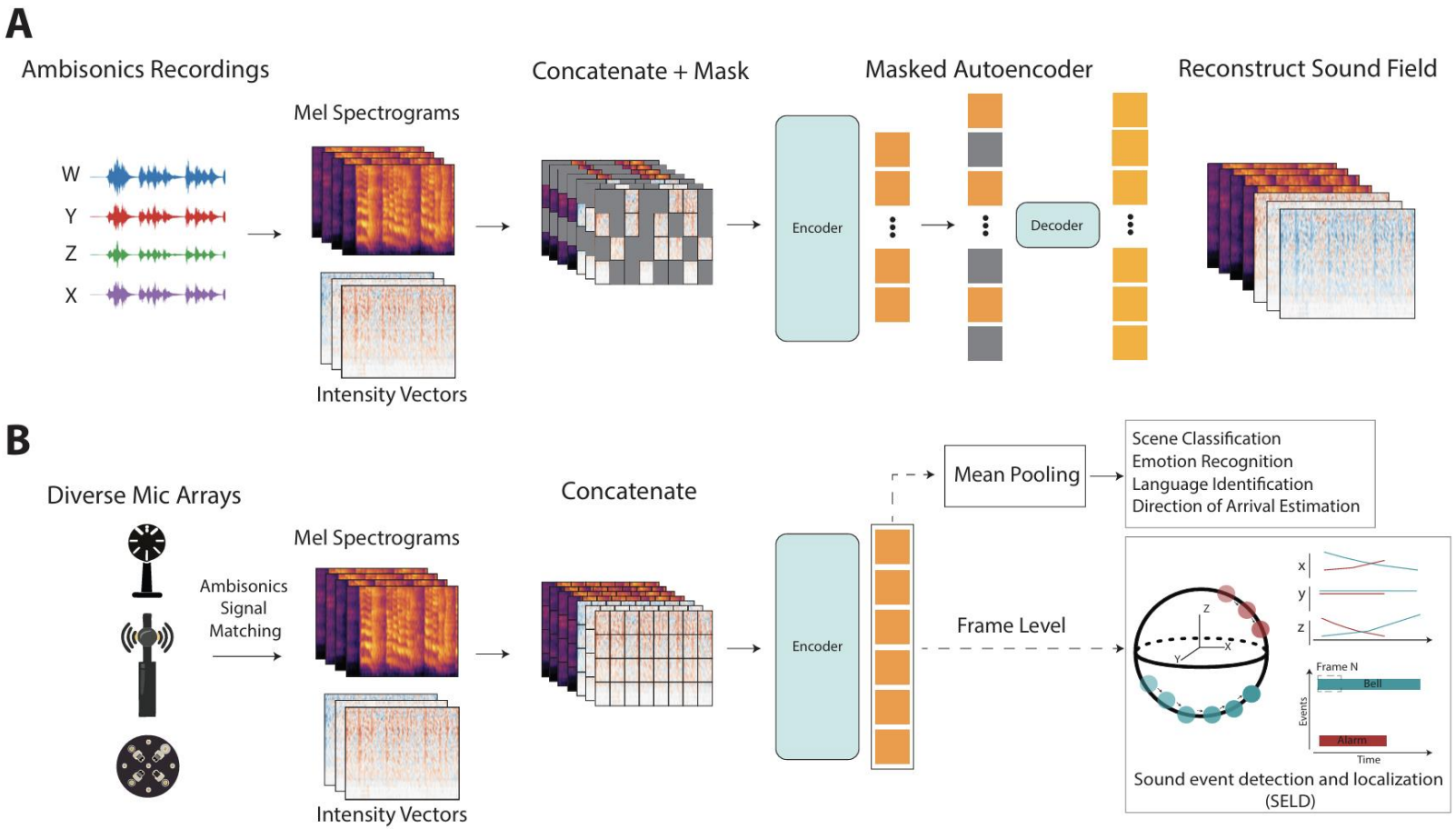
Audio representation models learn **representations that transfer to a variety of downstream tasks**.

They **overlook the spatial structure,** ruling out **tasks such as sound event localization and detection (SELD)**.

We introduce **AudioSphere**, a masked autoencoder that learns joint spatial-acoustic representations.

We also introduce **RealSELD**, a benchmark that evaluates frozen spatial-acoustic representations on real-world SELD tasks.

# AUDIOSPHERE PIPELINE



# TRAINING AND EVALUATION

We pre-train AudioSphere on AudioSet.

We render room impulse responses (RIRs) with the SoundSpaces 2.0 simulator to generate generate first-order Ambisonics (FOA) signals.

For evaluation, we use the **HEAR benchmark** and propose the **RealSELD benchmark**.

We use the HEAR benchmark to answer the following question:

- **Does spatial pre-training transfer to single-channel audio understanding?**

# REALSELD

RealSELD consists of tasks with real-world recorded RIRs and real-world microphone recordings. It covers both first-order Ambisonics and tetrahedral microphone arrays.

| Name | Motion | Impulse Response | Max Overlap | Noise |
|---|---|---|---|---|
| TUT-2018 Real (ov1, 2, 3) [Adavanne et al., 2019a] | Static | Real | 1, 2, 3 | Ambient 30 dB |
| TAU-2019 [Adavanne et al., 2019b] | Static | Real | 2 | Ambient 30 dB |
| TAU-2020 [Politis et al., 2020a] | Dynamic | Real | 2 | Ambient [30-6 dB] |
| TAU-2021 [Politis et al., 2021] | Dynamic | Real | 3 | Dynamic + Ambient |
| STARSS23 [Shimada et al., 2023] | Dynamic | – | 6 | Dynamic + Ambient |

We also evaluate on RSL2019-dev, a real-world sound localization task with a planar microphone array.

# REALSELD: RESEARCH QUESTIONS

We use this benchmark to answer the following questions:

1. **Do the learned embeddings carry frame-resolved spatial information that supports real-world SELD, in static and in dynamic scenes?**
2. **Does this hold across microphone geometries?**

# RESULTS ON HEAR

AudioSphere **preserves single-channel audio performance**.

Mixing spatial and single-channel audio **yields the best performance**.

Table 2: Aggregate HEAR score (mean over the 11 tasks of Appendix Table 13; per-task results in Appendix Table 14), ± SD over three downstream seeds. Grey: supervised. **AS**: AudioSet.

| Model | Corpus | HEAR |
|---|---|---|
| BEATs | AS | 76.6 |
| SSAM | AS | 79.6 |
| ATST-Clip | AS | 79.8 |
| MW-MAE | AS | 80.0 |
| ATST-Frame | AS | 81.9 |
| Dasheng | Mix (272k h) | 82.8 |
| PaSST | AS | 68.1 |
| Spatial-AST | AS | 74.6 |
| AudioSphere ($\alpha$=0.0) | AS | 81.1 |
| AudioSphere ($\alpha$=0.75) | AS | **83.2** ± 0.1 |
| AudioSphere-Clean | AS | 83.0 ± 0.2 |

# RESULTS ON DYNAMIC REALSELD

AudioSphere performs dynamic sound event localization and detection. **Fine-tuning with a temporal head (TM) yields the best results.**

Table 5: **Dynamic real-world SELD.** Frozen probing and fine-tuning on TAU-2020/2021 with the shallow MLP probe and with the lightweight temporal head (TM) of Section 4. **Bold** = best per column.

| | TAU-2020 | | | | TAU-2021 | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | $\mathbf{ER}_{20^\circ}\downarrow$ | $\mathbf{F}_{20^\circ}\uparrow$ | $\mathbf{LE}_{\mathbf{CD}}\downarrow$ | $\mathbf{LR}_{\mathbf{CD}}\uparrow$ | $\mathbf{ER}_{20^\circ}\downarrow$ | $\mathbf{F}_{20^\circ}\uparrow$ | $\mathbf{LE}_{\mathbf{CD}}\downarrow$ | $\mathbf{LR}_{\mathbf{CD}}\uparrow$ |
| **Supervised baseline** | | | | | | | | |
| SELDNet | 0.72 | 37.4 | 22.8 | 60.7 | 0.73 | 30.7 | 24.5 | 40.5 |
| **Probing (frozen backbone)** | | | | | | | | |
| Frozen + MLP | 0.57 | 46.6 | 23.5 | 63.2 | 0.75 | 21.0 | 36.9 | 49.1 |
| Frozen + TM | 0.41 | 63.6 | 15.0 | 75.0 | 0.63 | 37.2 | 24.3 | 61.3 |
| **Fine-tuning** | | | | | | | | |
| Fine-tune + MLP | 0.42 | 65.8 | 13.4 | **78.2** | 0.58 | 47.9 | 18.2 | 66.3 |
| Fine-tune + TM | **0.40** | **68.4** | **11.1** | 77.8 | **0.51** | **57.7** | **14.6** | **67.4** |

# PRE-TRAINING VS. TRAINING FROM SCRATCH

Our pre-trained weights supply crucial information: **pre-training helps AudioSphere train faster and better** than training the model in-domain from scratch.

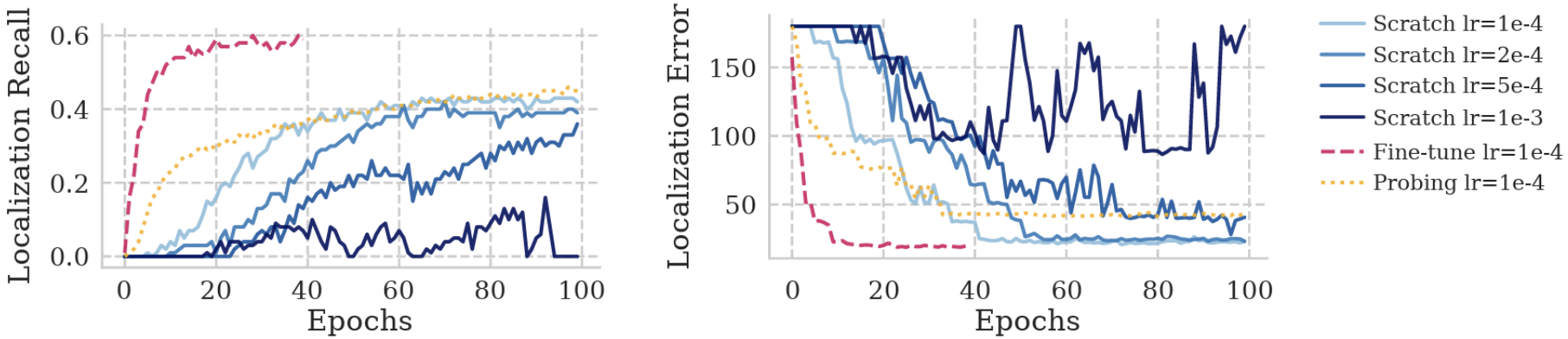



Figure 2: **Training dynamics on STARSS23.** Validation localization recall and localization error across epochs with a shallow MLP head attached to the backbone.

# RESULTS ON MICROPHONE GEOMETRIES: STATIC SCENES

AudioSphere works with diverse microphone geometries in static scenes, using Ambisonics Signal Matching (ASM) to map each array to FOA.

AudioSphere **localizes sound sources from a planar microphone array** without re-training the backbone.

Table 6: **Array geometry.** (a) Native Ambisonics vs. ASM-encoded tetrahedral input on TAU2019 for SELDNet and frozen AUDIOSPHERE. (b) DoA error on RSL2019-dev (median ± IQR). **Bold** = best per column.

| Input | ER↓ | F↑ | LE↓ | LR↑ |
|---|---|---|---|---|
| **SELDNet (supervised)** | | | | |
| Ambisonics | 0.34 | 79.9 | 28.5 | 85.4 |
| Mic. array | 0.35 | 80.0 | 30.8 | 84.0 |
| **AUDIOSPHERE (self-supervised, frozen)** | | | | |
| Ambisonics | 0.28±0.05 | 84.6±2.1 | 14.1±1.3 | 78.5±0.6 |
| Mic. array | 0.25 | 84.8 | 21.2 | 73.6 |

| Method | DoA error (°)↓ |
|---|---|
| **Baselines** | |
| BSM + Spatial-AST (BAT) | 13.1 ± 29.0 |
| MUSIC | 5.6 ± 11.6 |
| **Ours** | |
| ASM + AUDIOSPHERE | **1.9** ± 6.6 |

# RESULTS ON MICROPHONE GEOMETRIES: DYNAMIC SCENES

AudioSphere works with diverse microphone geometries in dynamic scenes, and fine-tuning improves performance on ASM-mapped FOA signals.

| Method | TAU2020 | | | | TAU2021 | | | |
|---|---|---|---|---|---|---|---|---|
| | $ER_{20^\circ}\downarrow$ | $F_{20^\circ}\uparrow$ | $LE_{CD}\downarrow$ | $LR_{CD}\uparrow$ | $ER_{20^\circ}\downarrow$ | $F_{20^\circ}\uparrow$ | $LE_{CD}\downarrow$ | $LR_{CD}\uparrow$ |
| **Baseline (SELDNet)** | | | | | | | | |
| Baseline FOA | 0.72 | 37.4 | 22.8 | 60.7 | 0.73 | 30.7 | 24.5 | 40.5 |
| Baseline ASM-FOA | 0.77 | 32.2 | 28.0 | 58.3 | 0.72 | 20.5 | 29.4 | 39.7 |
| **AUDIOSPHERE (frozen)** | | | | | | | | |
| Frozen + TM, FOA | 0.41 | 63.6 | 15.0 | 75.0 | 0.63 | 37.2 | 24.3 | 61.3 |
| Frozen + TM, ASM-FOA | 0.49 | 55.5 | 17.7 | 74.9 | 0.71 | 27.0 | 28.8 | 60.3 |
| **AUDIOSPHERE (fine-tuned)** | | | | | | | | |
| Fine-tune + TM, FOA | **0.40** | **68.4** | **11.1** | **77.8** | **0.51** | **57.7** | **14.6** | **67.4** |
| Fine-tune + TM, ASM-FOA | 0.44 | 63.7 | 14.5 | 77.2 | 0.58 | 46.4 | 20.9 | 66.7 |

# LIMITATIONS

We use SoundSpaces 2.0 to generate a sufficiently large training corpus, but **only with static scenes**. Training on real-world Ambisonics recordings from 360° video remains open.

We rely on ASM for microphone-agnostic scene analysis, but the results depend on the spatial sampling capacity of the array. A native multi-microphone model could yield better results.

Our spatial evaluation covers sound event detection and direction-of-arrival (DoA) estimation only; it does not evaluate distance estimation, source separation, dereverberation, or room-acoustic inference.

# SUMMARY

We present **AudioSphere**, a self-supervised multichannel masked autoencoder that learns joint spatial-acoustic representations.

Frozen AudioSphere transfers to single-channel audio tasks and to real-world SELD.

We release **RealSELD**, the first probing benchmark for SELD: seven real-world tasks from five datasets.

Code: github.com/labhamlet/AudioSphere
Model: huggingface.co/labhamlet/AudioSphere
Benchmark: huggingface.co/datasets/labhamlet/RealSELD