SSLAM: Enhancing Self-Supervised Models with Audio Mixtures for Polyphonic Soundscapes

Tony Alex, Sara Ahmed, Armin Mustafa, Muhammad Awais, Philip JB Jackson

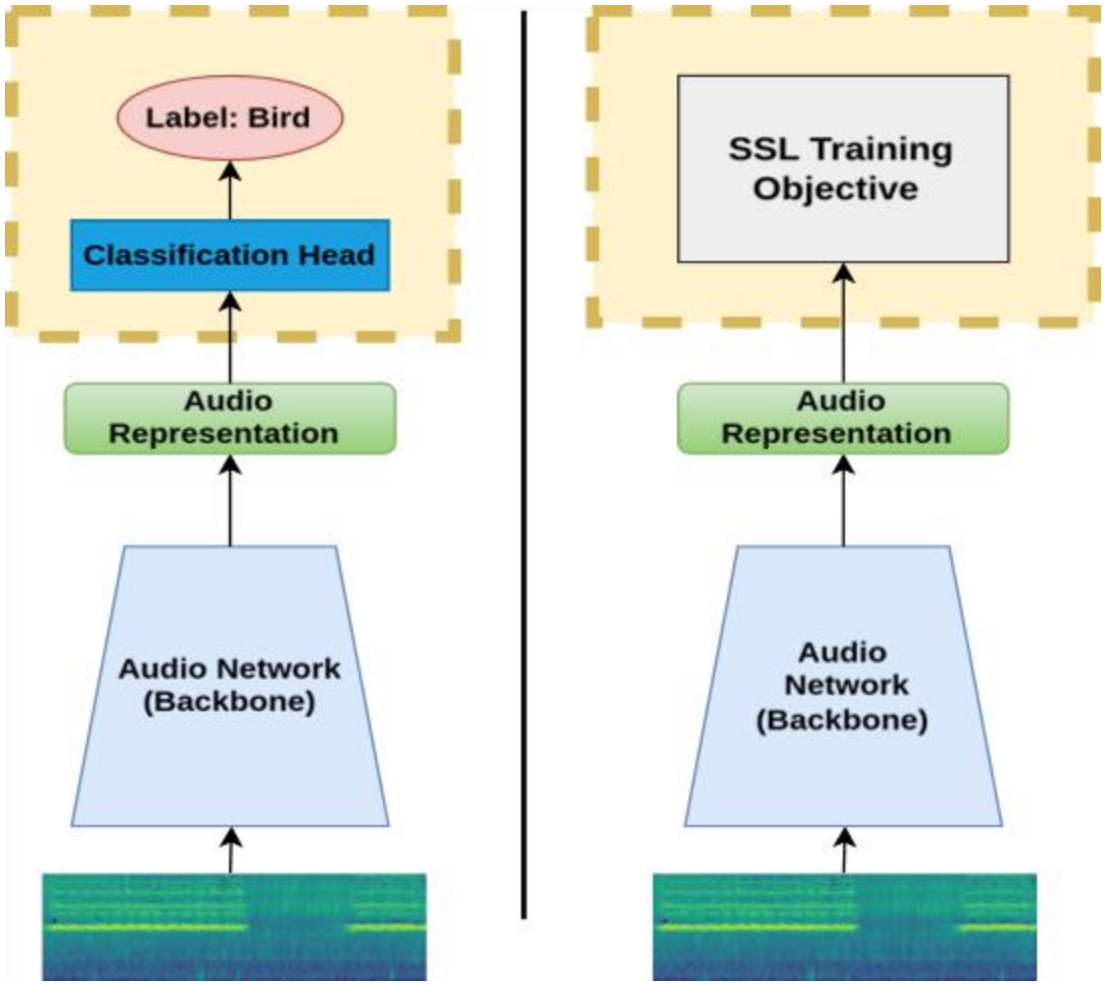
Surrey Institute for People-Centred AI(PAI),

Centre for Vision, Speech and Signal Processing(CVSSP),

University of Surrey

https://github.com/ta012/SSLAM

Supervised vs Self-Supervised Representation Learning?



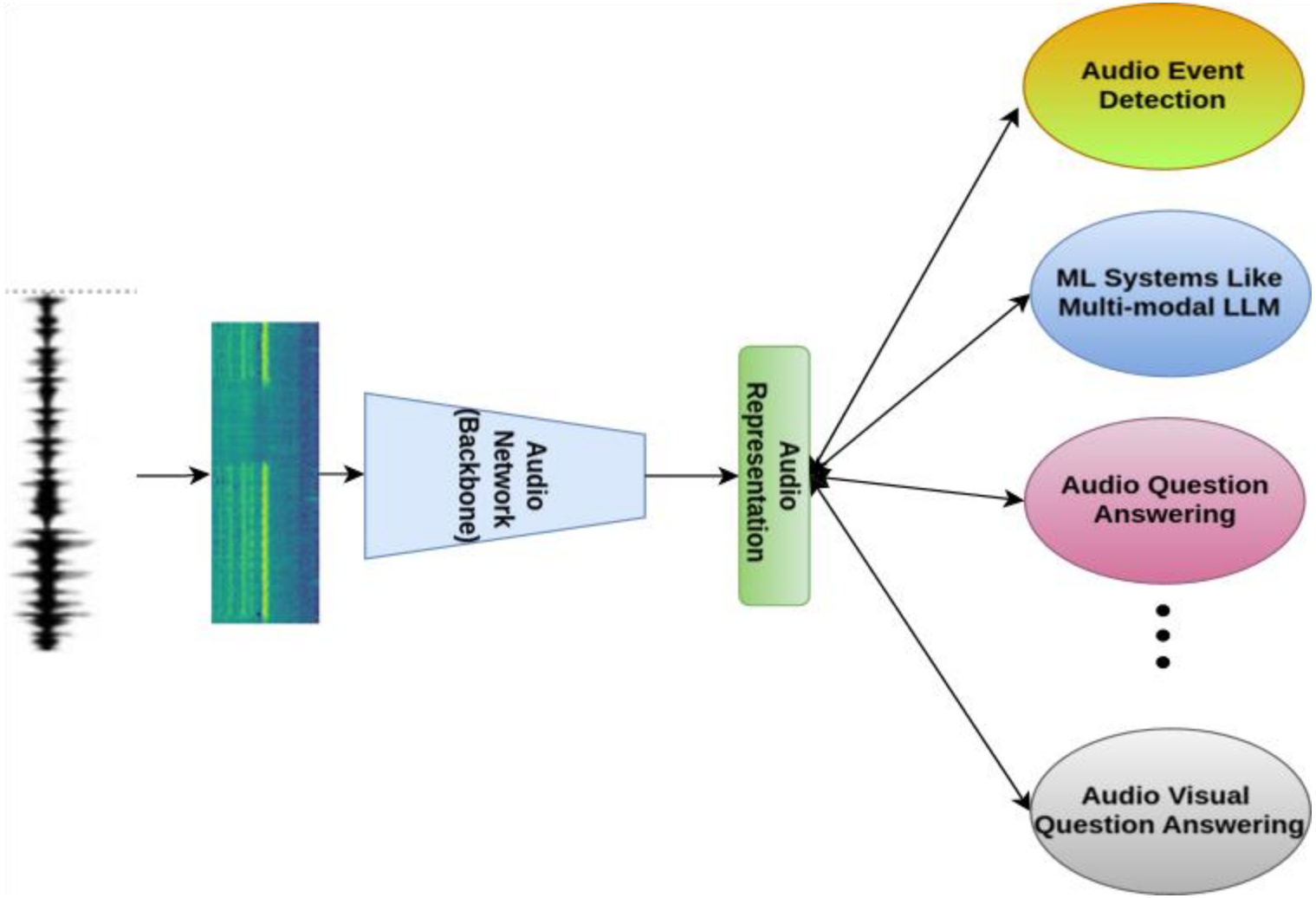
- Audio-MAE[1]
- BEATs[2]
- EAT[3]

[1] Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., & Feichtenhofer, C. (2022). Masked autoencoders that listen. Advances in Neural Information Processing Systems, 35, 28708–28720.

[2] Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., & Wei, F. (2022). Beats: Audio pre-training with acoustic tokenizers. ArXiv Preprint ArXiv:2212.09058.

[3] Chen, W., Liang, Y., Ma, Z., Zheng, Z., & Chen, X. (2024). EAT: Self-supervised pre-training with efficient audio transformer. ArXiv Preprint ArXiv:2401.03497.

Real-World Applications of Learned Audio Representations

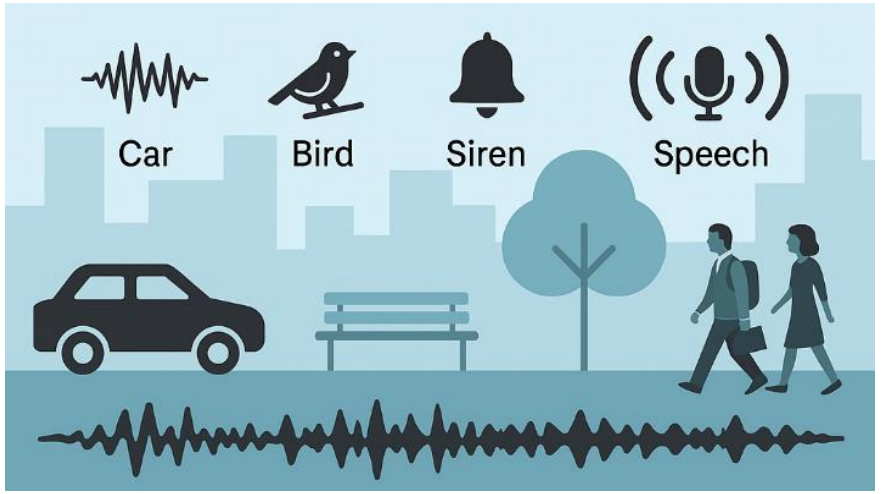


Why SSLAM?

- Real-world audio is **polyphonic**—multiple overlapping sound sources are common in everyday environments.

- Existing SSL models are **benchmarked** primarily on monophonic audio, and their **pre-training does not consider polyphonic** environments. As a result, their ability to generalize to real-world, multi-source audio scenarios is limited.

- SSLAM bridges this gap by introducing self-supervised learning from **audio mixtures**, enabling robust learning across both **monophonic** and **polyphonic** soundscapes.



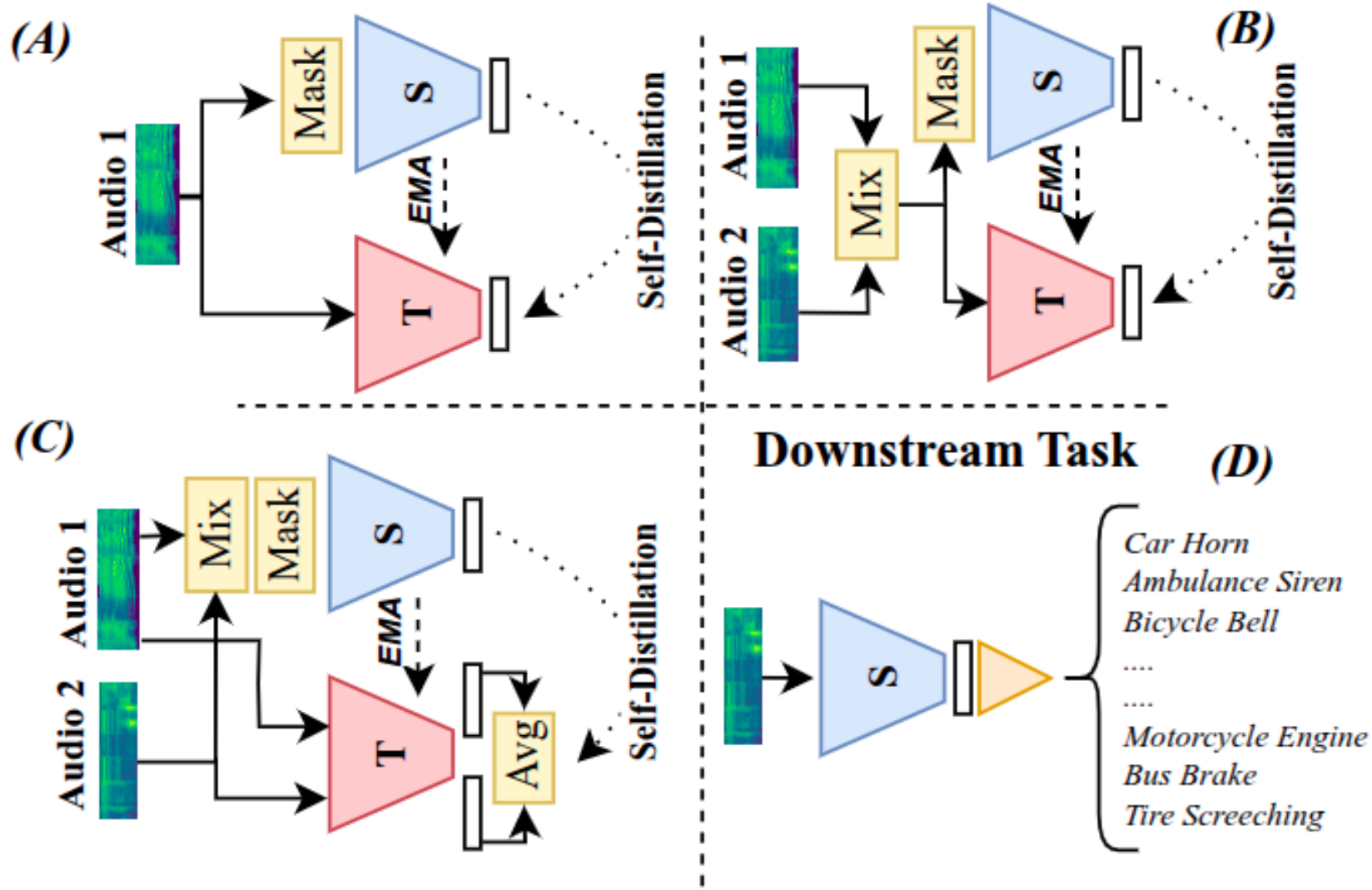
Our Contributions

- **Pre-Training**
  - **Self-Supervised Pre-training on Audio Mixtures**
    Enables better adaptation to real-world polyphonic audio through diverse source combinations.
  - **Source Retention Loss(SRL)**
    Preserves individual source characteristics within mixtures, ensuring source integrity in learned representations.

- **Benchmarking**
  - **SOTA Performance on audio-SSL Benchmarks**
    Achieves state-of-the-art performance on general audio and speech SSL tasks, with up to a +3.9% mAP improvement on AudioSet.
  - **Extended Benchmarking with Polyphonic Datasets**
    Added polyphonic datasets to the benchmarking suite and achieved substantial improvements in real-world polyphonic audio handling, with gains of up to +9.1%.

Pretraining objectives Overview

- Our pre-training is based on Masked Latent bootstrapping.



Experiments: General Audio-SSL Benchmark

Model	#Param	Pre-training Data	Audio			Speech	
			AS-2M	AS-20K	ESC-50	KS2	KS1
SS-AST (Gong et al., 2022)	89M	AS+LS	-	31.0	88.8	98.0	96.0
MAE-AST (Baade et al., 2022)	86M	AS+LS	-	30.6	90.0	97.9	95.8
MaskSpec (Chong et al., 2023)	86M	AS	47.1	32.3	89.6	97.7	-
MSM-MAE (Niizumi et al., 2022)	86M	AS	-	-	85.6	87.3	-
data2vec (Baevski et al., 2022)	94M	AS	-	34.5	-	-	-
Audio-MAE (Huang et al., 2022a)	86M	AS	47.3	37.1	94.1	98.3	96.9
BEATs$_{iter3}$ (Chen et al., 2022)	90M	AS	48.0	38.3	95.6	98.3	97.7
BEATs$_{iter3+}$	90M	AS	48.6	38.9	98.1	98.1	98.1
ASiT (Ahmed et al., 2024)	86M	AS	48.0	38.6	95.3	**98.9**	98.2
A-JEPA (Fei et al., 2024)	86M	AS	48.6	38.4	**96.3**	98.5	97.7
EAT (Chen et al., 2024)	88M	AS	48.6	40.2	95.9	98.3	-
SSLAM (Ours)	88M	AS	**50.2**	**40.9**	96.2	98.1	**98.8**

Experiments: Different degrees of Polyphony

Table 3: Evaluation on the *Degrees of polyphony* dataset: Assessing the impact of various individual contributions across different polyphony levels. $\{a,b\}$ denotes a data subset where audio files contain a or b distinct sound events. All performances are reported in mAP. For more details about the datasets refer to Appendix B.0.2.

Model	Unmixed Data	Partial Mixed Data	SRL	Number of Distinct sound events {2,3}	{4,5}	{6,7}	{8,9}	{10,11}	{12,13}	{14+}
Linear Evaluation										
MB-UA	✓	✗	✗	**61.5**	69.4	45.8	53.5	58.3	61.6	66.7
MB-PMA (Ours)	✗	✓	✗	58.6	70.0	50.7	57.2	61.3	64.8	67.6
MB-UA-PMA (Ours)	✓	✓	✗	58.2	70.0	49.8	56.9	61.1	64.7	67.9
SSLAM (Ours)	✓	✓	✓	60.6	**70.6**	**53.2**	**58.7**	**63.0**	**66.1**	**69.7**
Fine-tuning										
MB-UA	✓	✗	✗	87.3	86.5	69.5	81.5	82.5	80.7	78.1
MB-PMA (Ours)	✗	✓	✗	87.3	**86.9**	71.4	83.0	83.4	82.0	79.3
MB-UA-PMA (Ours)	✓	✓	✗	87.2	86.4	70.3	82.7	83.4	81.8	78.8
SSLAM (Ours)	✓	✓	✓	**87.7**	**86.9**	**71.9**	**83.3**	**83.8**	**82.2**	**79.4**

Experiments: Different Polyphonic datasets

Table 2: Impact of individual novel contributions evaluated across various polyphonic datasets. All performances are reported in mAP. For more details about the datasets refer to Appendix B.0.2.

Model	SPASS					IDMT	URBAN	AS-20K
	Square	Park	Waterfront	Street	Market	DESED	SED	
Linear Evaluation								
MB-UA	60.1	59.7	55.2	63.7	62.8	75.8	71.3	13.9
MB-PMA (Ours)	63.1	63.5	58.5	66.5	67.4	78.4	70.9	16.1
MB-UA-PMA (Ours)	62.7	63.5	58.2	66.6	66.6	77.7	70.9	15.2
SSLAM (Ours)	**64.2**	**64.2**	**59.5**	**67.4**	**68.5**	**77.8**	**71.4**	**16.9**
Fine-tuning								
MB-UA	84.4	78.4	80.1	81.4	89.7	94.4	**90.9**	40.4
MB-PMA (Ours)	85.1	80.0	82.0	82.2	90.8	94.4	**90.9**	40.6
MB-UA-PMA (Ours)	85.0	79.7	82.0	82.2	90.5	94.4	**90.9**	40.7
SSLAM (Ours)	**85.6**	**80.5**	**82.6**	**82.2**	**90.2**	**94.5**	**90.9**	**40.9**

Thank You!