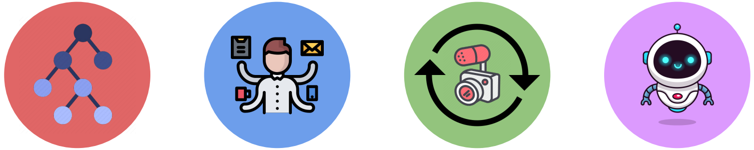
Incorporating Natural Structure into Transfer Learning Methods for Machine Listening



ECE PhD Dissertation Defense

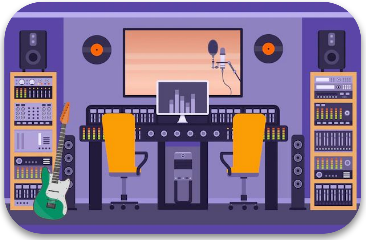
Aurora Cramer
(they/she)

Advised by
Prof. Juan Pablo Bello, PhD

NYU TANDON SCHOOL OF ENGINEERING

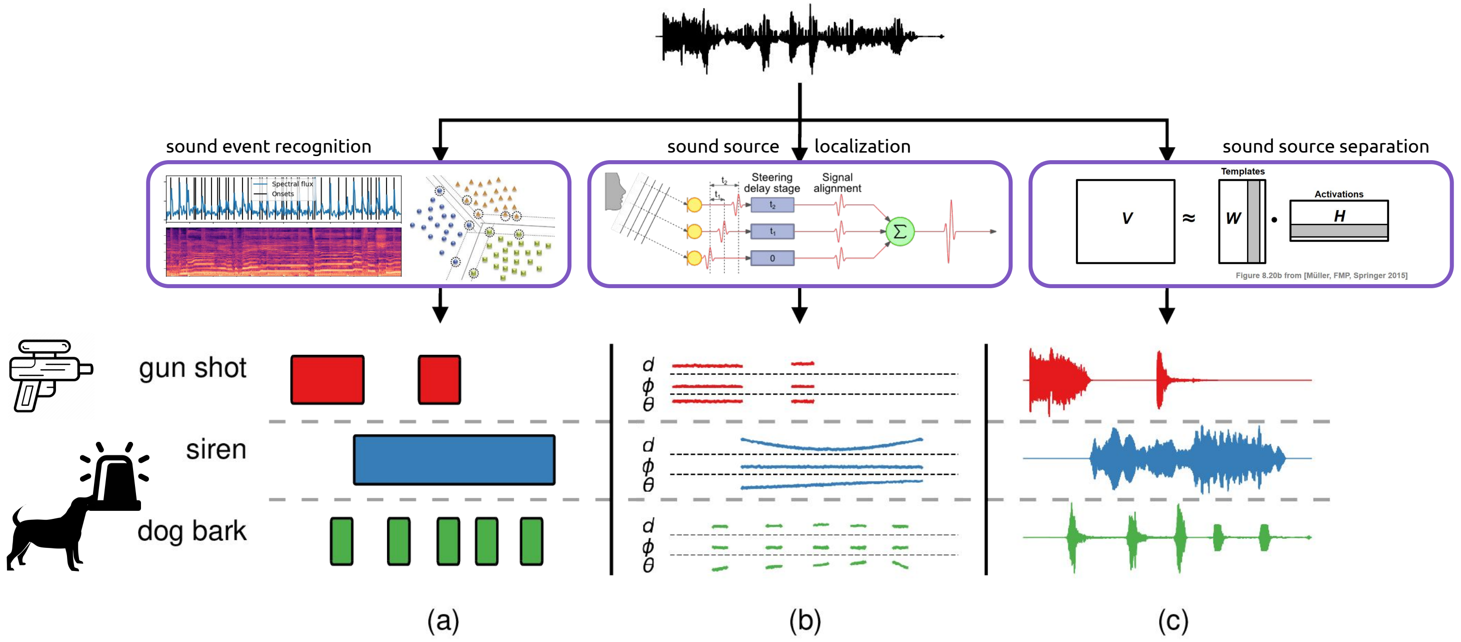
MARL

Machine listening is everywhere!

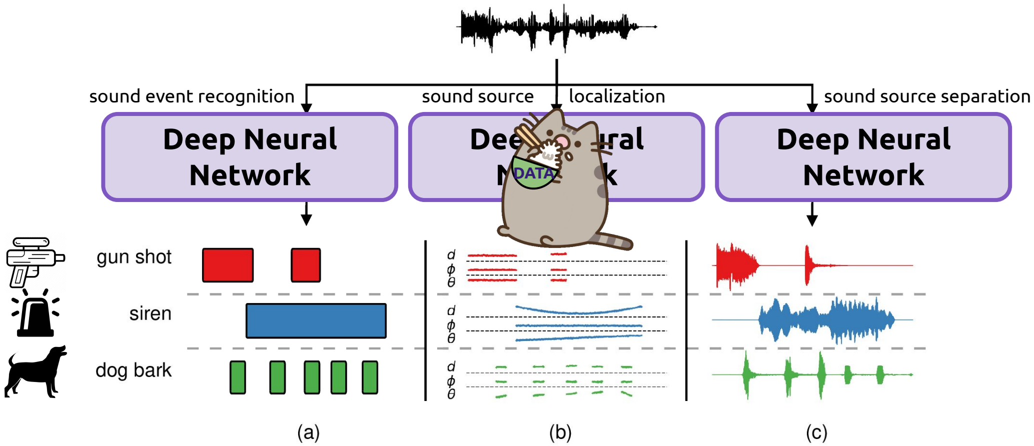


Traditional machine listening methods

Traditional methods use handcrafted signal processing features and **linear/Gaussian/Markov modeling methods**

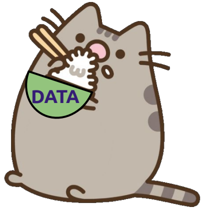


Modern machine listening methods



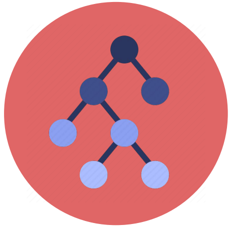
Fully-supervised DNN models have generally pushed the SOTA for machine listening, but suffer from **lack of abundant annotated audio data**

Audio annotation is **time-consuming**, **expensive**, and often **difficult**



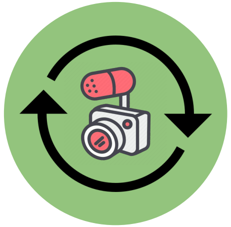
How do we make more effective use of the data we have?

What kinds of structure do people use?

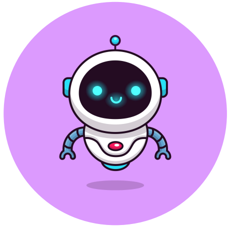


“It helps to know how concepts are related”

“similar tasks often require similar skills”



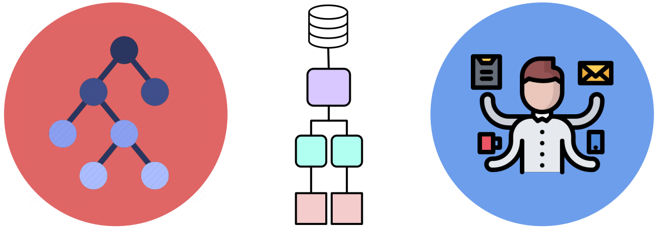
“many events can be perceived by multiple senses”



“learning can happen anywhere”

What kinds of structure can machines use?

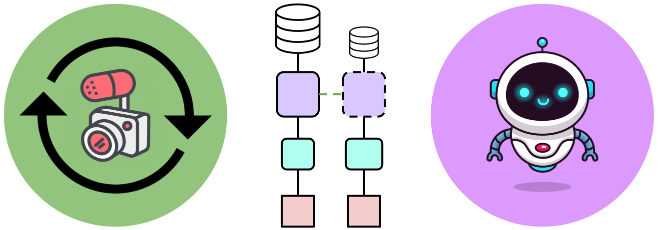
parallel transfer learning
(a.k.a. multi-task learning)



sound source hierarchies

multi-task output structure

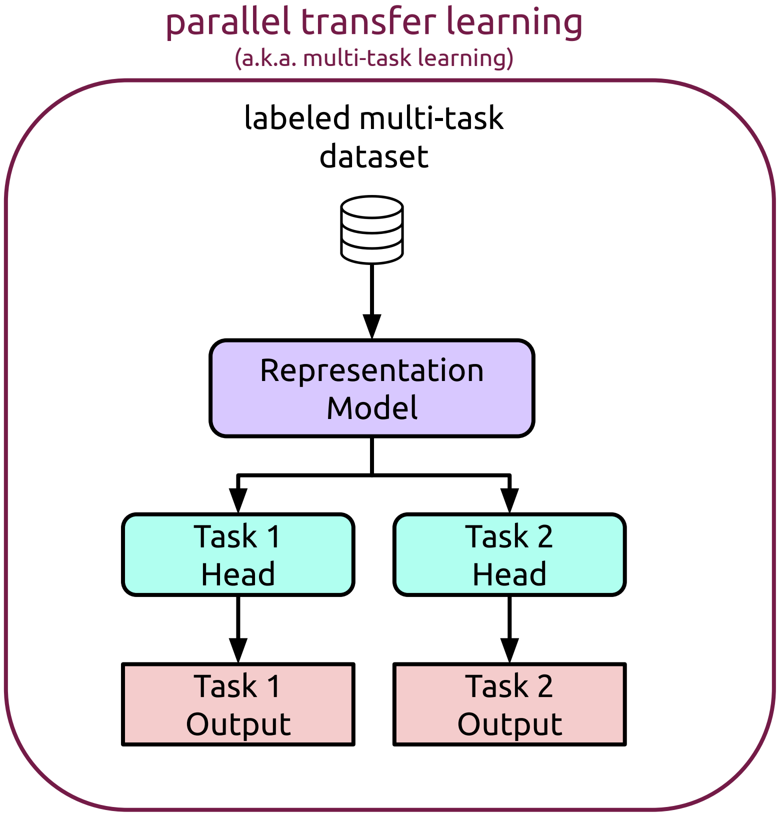
sequential transfer learning



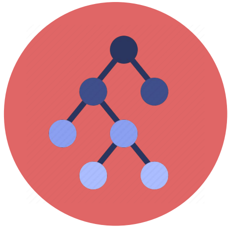
multi-modal self-supervision

embodied agents

Flavors of transfer learning



- **Shared representation is regularized** by encoding relevant information for each task
- Model can learn **more efficiently** from the **same input data**



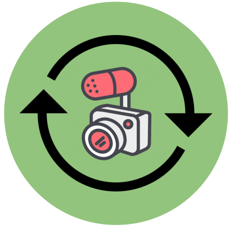
sound source hierarchies

multi-task output structure

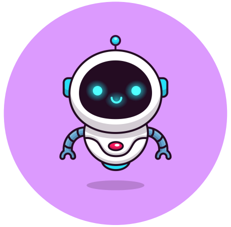
Flavors of transfer learning

- Models for niche applications can **take advantage of existing large datasets** despite small target datasets

- Sufficiently **general representation** can be used to easily bootstrap downstream models for variety of domains

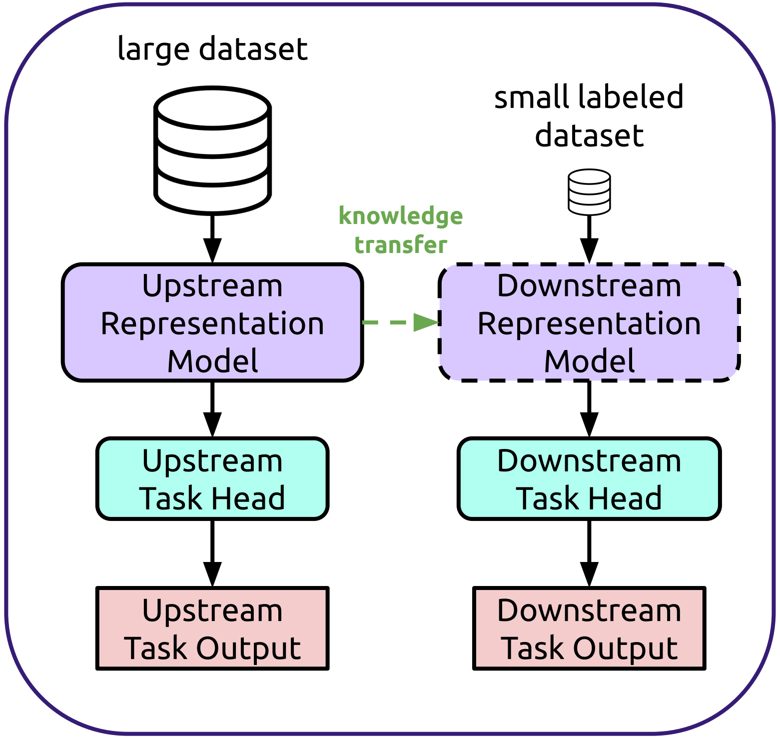


multi-modal
self-supervision



embodied agents

sequential transfer learning



Parallel transfer learning using sound source hierarchies

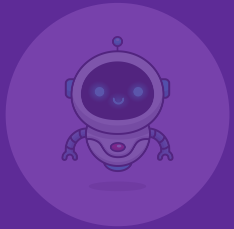
parallel transfer learning
(a.k.a. multi-task learning)

sequential transfer learning

sound source
hierarchies

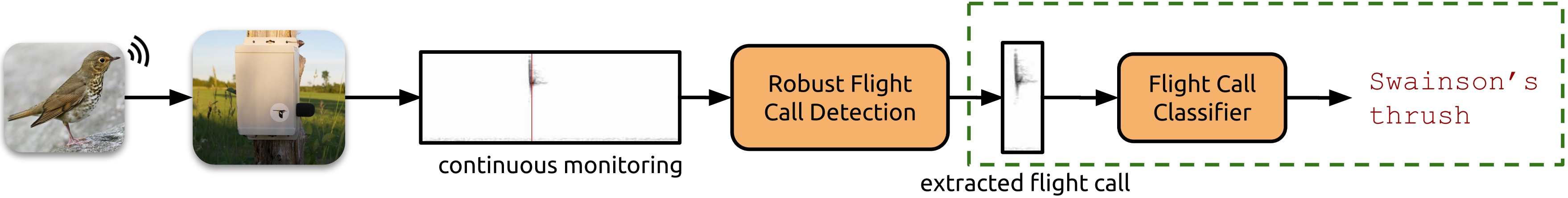
multi-task output
structure

multi-modal
self-supervision



embodied agents

Automated flight call classification

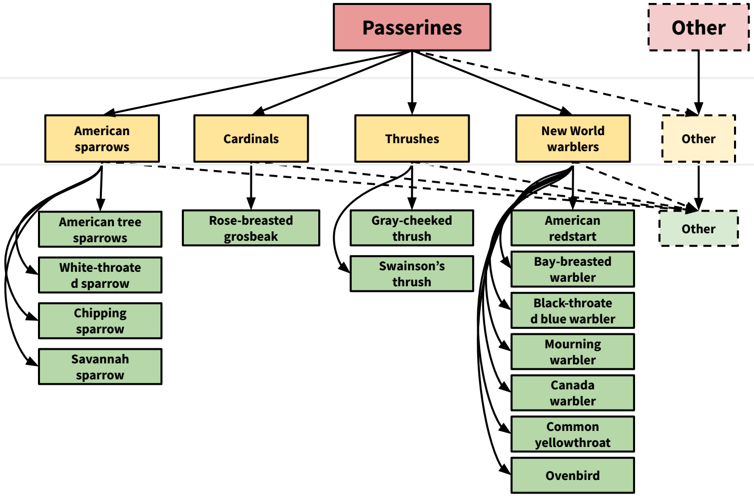


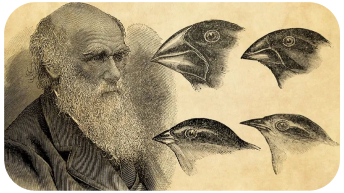
Incorporate hierarchical structure from **biological taxonomies** via **multi-task training** and **hierarchical model architectures**

Coarse Level: Order

Medium Level: Family

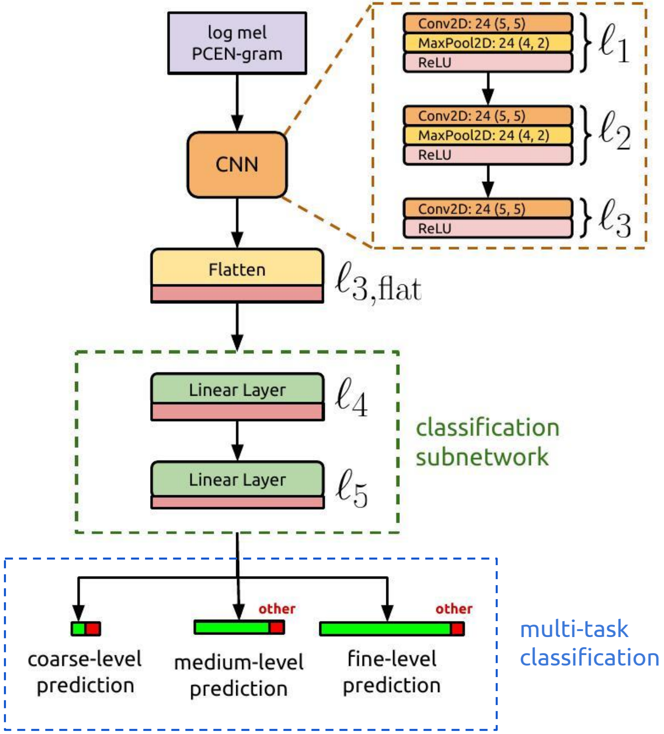
Fine Level: Species





V. Lostanlen, **A. Cramer**, J. Salamon, et al., "BirdVox: machine listening for bird migration monitoring," bioRxiv preprint 2022
A. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon, and J. P. Bello, "**Chirping up the right tree: incorporating biological taxonomies into bioacoustic classifiers**," ICASSP 2020

Hierarchical flight call classification

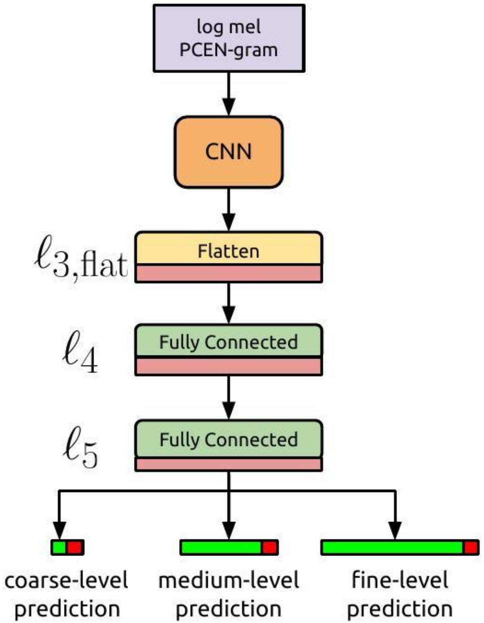


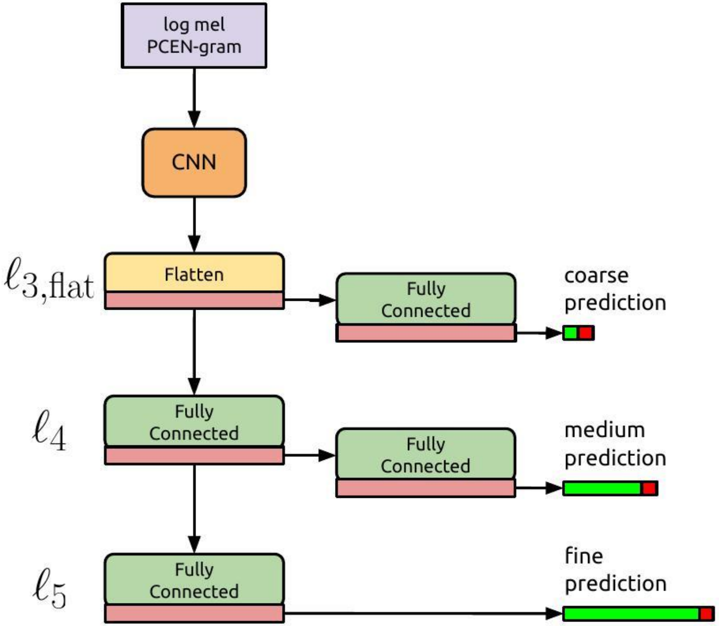
CNN architecture based on prior bird flight call classification model (Salamon et al. 2017)

NYU

A. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon, and J. P. Bello, "**Chirping up the right tree: incorporating biological taxonomies into bioacoustic classifiers**," ICASSP 2020

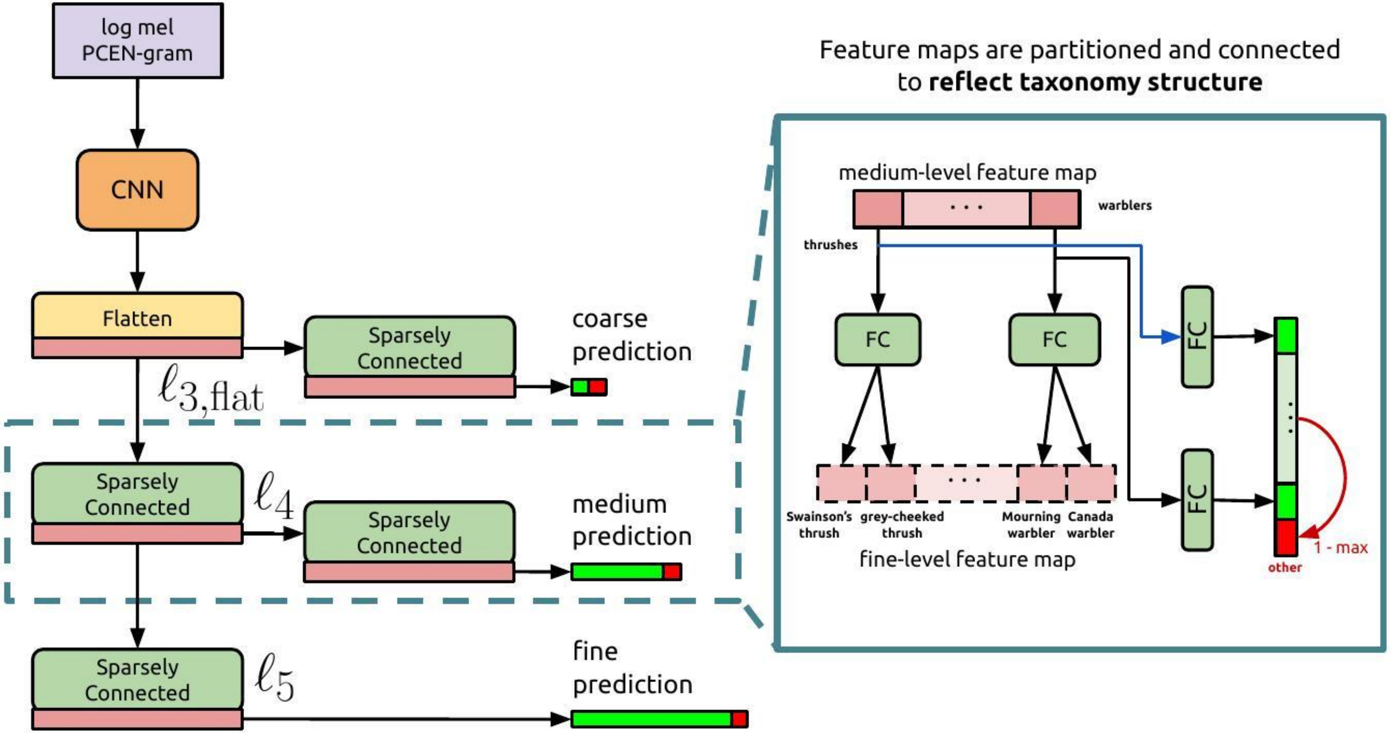
Multi-task prediction (d) and hierarchical outputs (e)





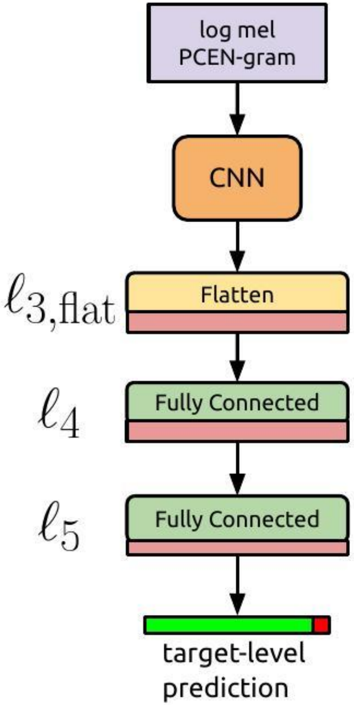
A. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon, and J. P. Bello, "**Chirping up the right tree: incorporating biological taxonomies into bioacoustic classifiers**," ICASSP 2020

Hierarchical partitioning (f)



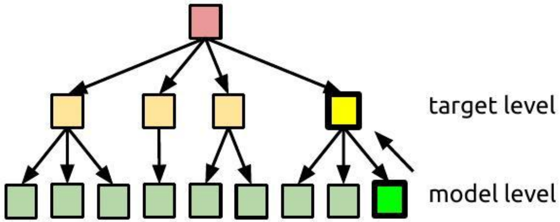
A. **Cramer**, V. Lostanlen, A. Farnsworth, J. Salamon, and J. P. Bello, "**Chirping up the right tree: incorporating biological taxonomies into bioacoustic classifiers**," ICASSP 2020

Single-task baseline (a) - (c)



Specialist strategy: use model trained at target level

Coarsening strategy: use target-level ancestor of predicted taxa



A. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon, and J. P. Bello, "**Chirping up the right tree: incorporating biological taxonomies into bioacoustic classifiers**," ICASSP 2020

Training and evaluation

- Train on **heterogeneous dataset** of flight calls from various sources (ANAFCC★)
- Input to model: 150 ms **log-scale mel-frequency spectrogram** with **per-channel energy normalization (PCEN)** applied
- **Data augmentation:** pitch shifting, time-stretching, additive background noise
- Evaluate at each taxonomic level using annotated clips from **full season of sensor network recordings** (BirdVox-14SD‡)

★ ANAFCC: https://doi.org/10.5281/zenodo.3666782
‡ BirdVox-14SD: https://doi.org/10.5281/zenodo.3667094

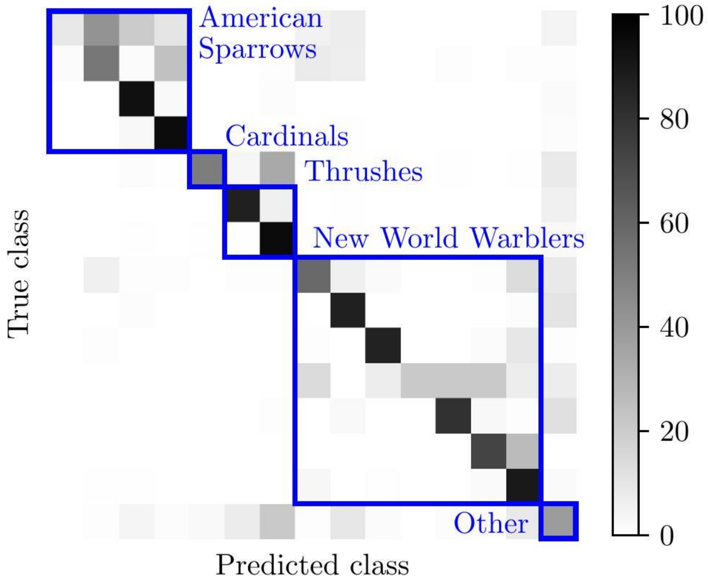
A. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon, and J. P. Bello, “**Chirping up the right tree: incorporating biological taxonomies into bioacoustic classifiers**,” ICASSP 2020

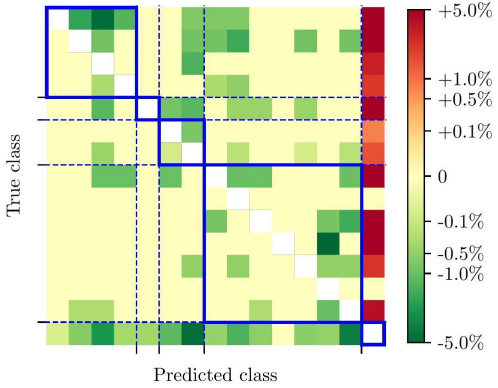
Experimental Results

	Model					# Trained Params	Fine Micro Acc.	Fine Macro Acc.	Medium Micro Acc.	Medium Macro Acc.	Coarse Acc.
	Single-Task Model										
Multi-task models match specialized and outperform coarsened!	(a)	Fine Level				641K	61.13	54.80	64.61	50.40	77.72
	(b)	Medium Level				640K	-	-	73.80	56.04	**94.75**
	(c)	Coarse Level				640K	-	-	-	-	93.85
	TaxoNet Model										
		Layer Partioning	Hier. Outputs	Classifier Activation	Classifier Projection						
safe, simple choice	(d)	No	No	sigmoid	Trainable	641K	61.82	55.83	75.10	55.87	94.39
better for variety of species	(e)	No	Yes	sigmoid	Trainable	650K	58.74	**58.06**	75.83	60.04	94.54
better for common species	(f)	Yes	Yes	sigmoid	Trainable	649K	**66.33**	55.69	76.50	61.60	94.69
	(g)	Yes	Yes	softmax	Trainable	649K	60.39	52.30	75.94	56.96	94.67
better for coarser taxa	(h)	Yes	Yes	tanh	Mean	640K	63.47	41.46	**79.36**	**65.08**	**94.75**

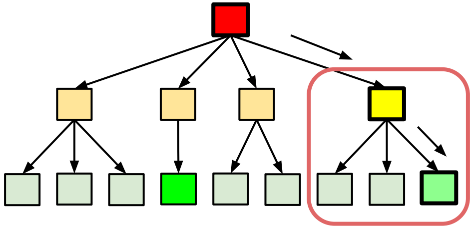
A. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon, and J. P. Bello, "**Chirping up the right tree: incorporating biological taxonomies into bioacoustic classifiers**," ICASSP 2020

Model makes reasonable errors



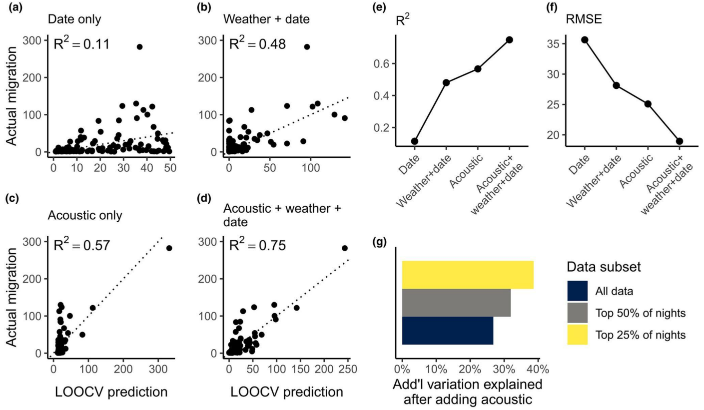


Enforcing hierarchical consistency procedure reduces confusion further

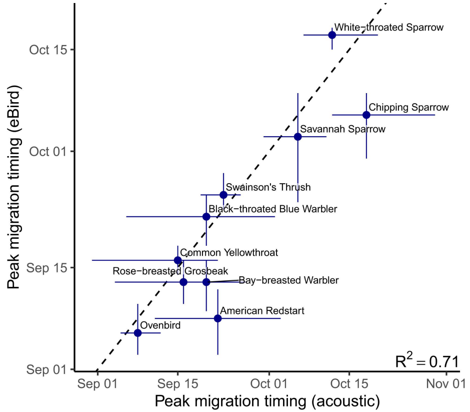


V. Lostanlen, **A. Cramer**, J. Salamon, et al., "**BirdVox: machine listening for bird migration monitoring**," bioRxiv preprint 2022

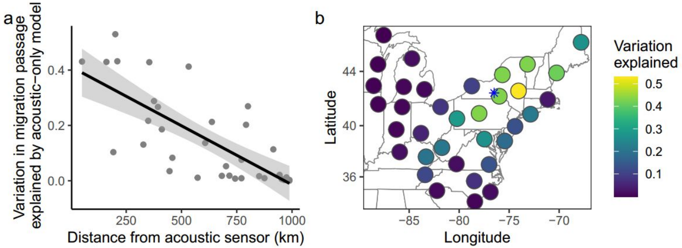
Automated acoustic monitoring captures intensity and timing of bird migration!



generalized linear model predicts migration intensity as measured by weather radar



generalized additive model predictions of peak migration time correlate with birdwatcher estimates



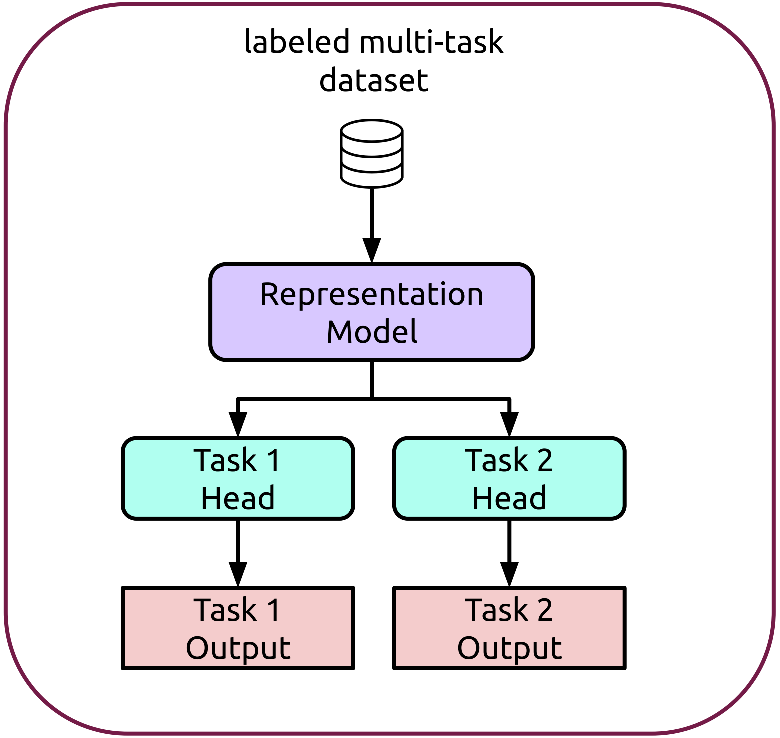
model also captures regional intensity at nearby radar locations

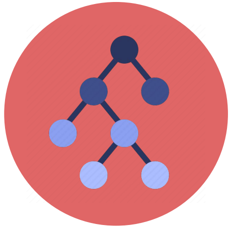
```
$ pip install birdvoxclassify
$ pip install birdvoxdetect
```

B. M. Van Doren, V. Lostanlen, **A. Cramer**, et al., "**Automated acoustic monitoring captures timing and intensity of bird migration**," Journal of Applied Ecology 2022

Flavors of transfer learning

parallel transfer learning
(a.k.a. multi-task learning)





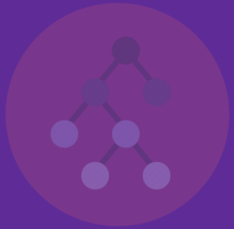
sound source hierarchies

hierarchical structure
improves **model robustness!**

Parallel transfer learning using multi-task output structure

parallel transfer learning
(a.k.a. multi-task learning)

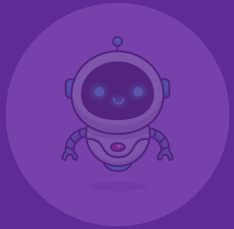
sequential transfer learning



sound source hierarchies

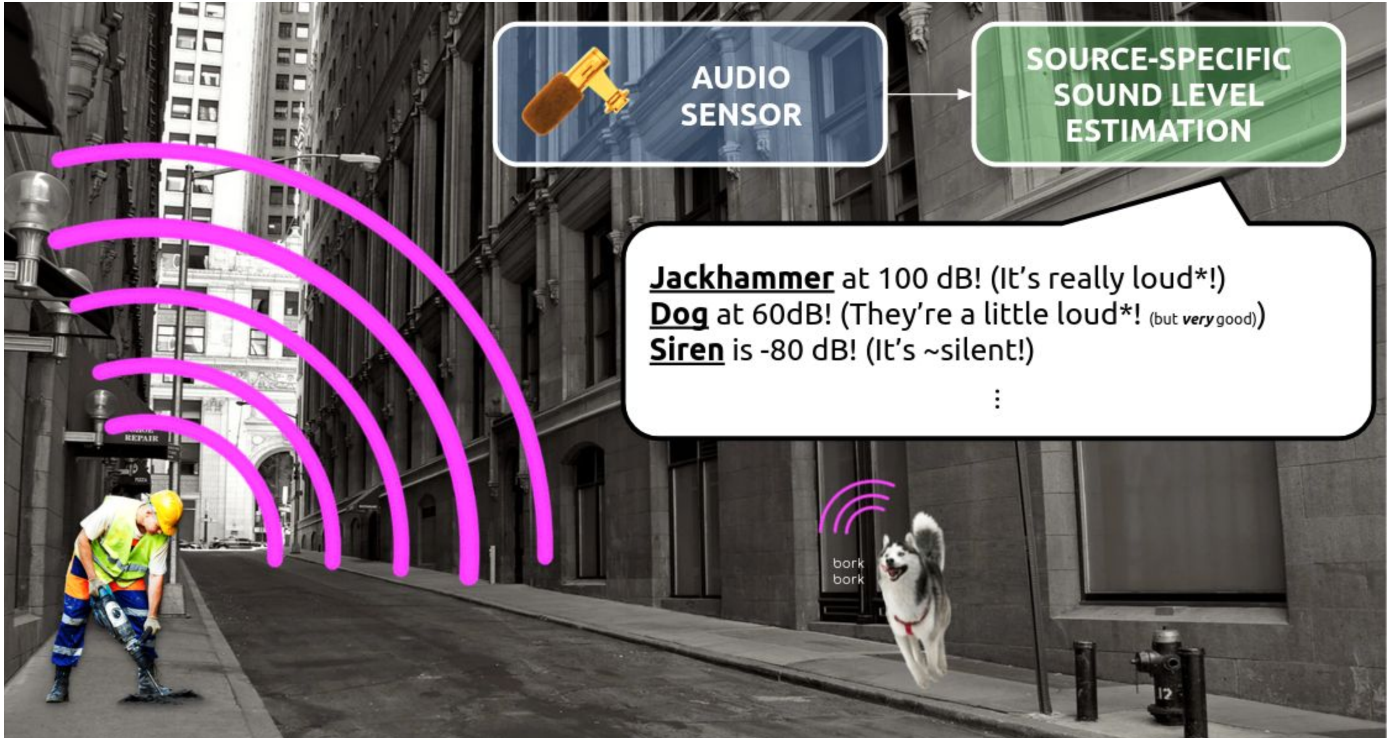
multi-task output structure

multi-modal self-supervision

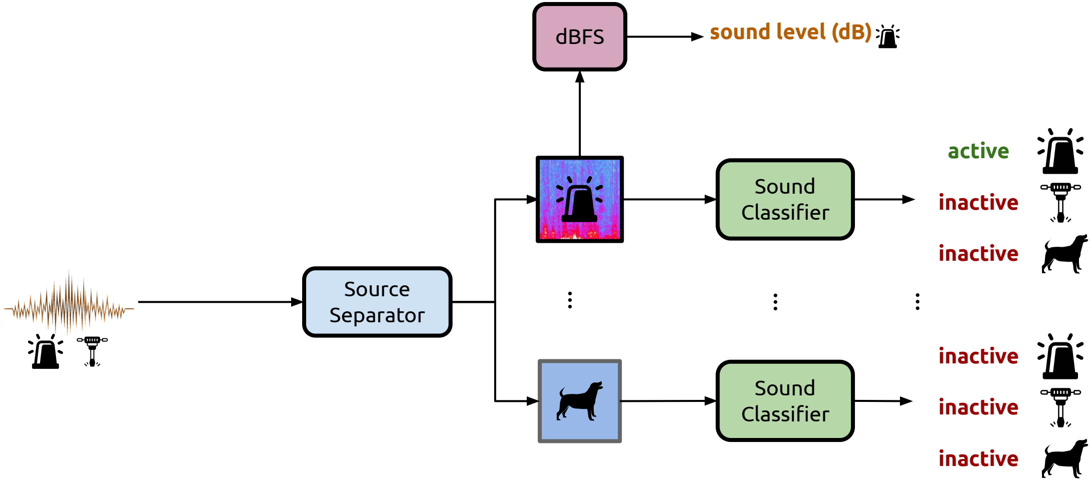


embodied agents

Source specific sound level estimation is difficult in practice



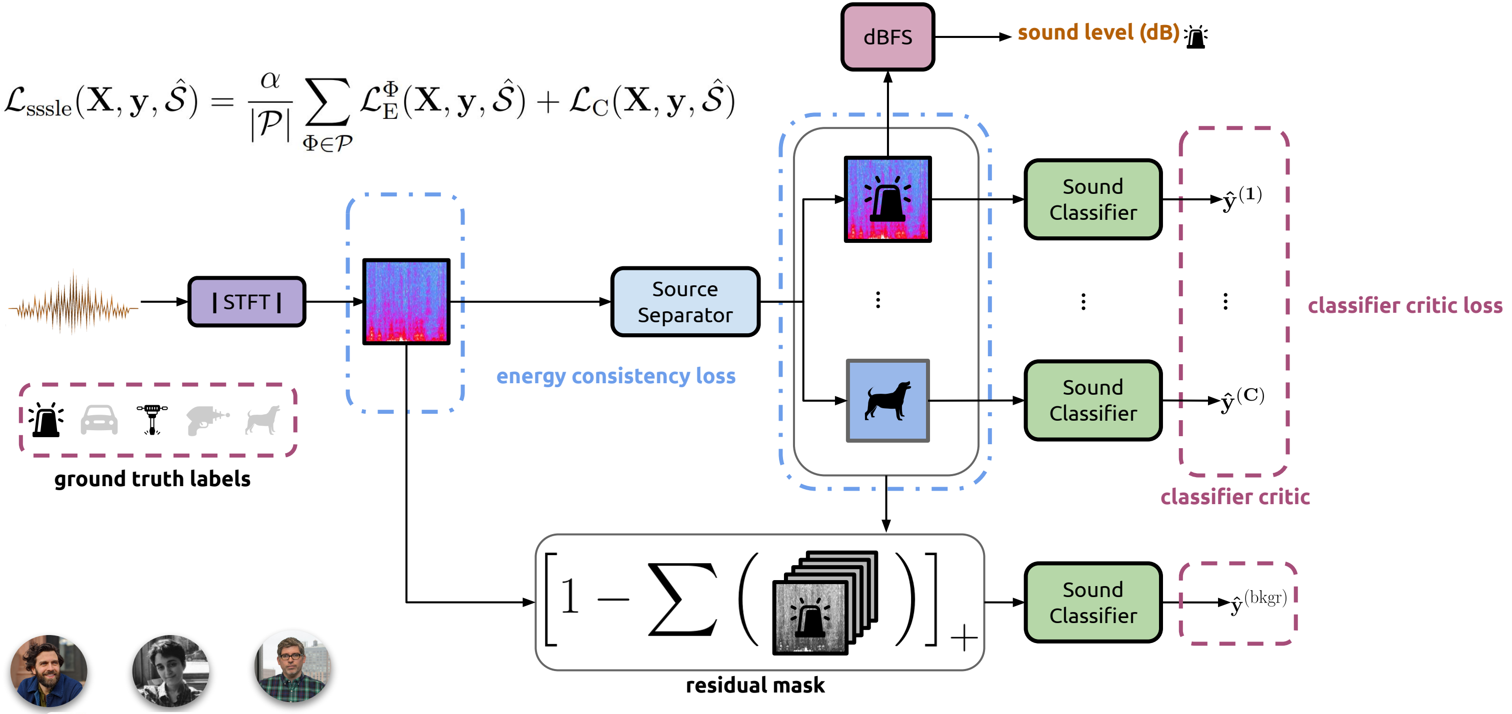
Source specific sound level estimation with multi-task learning



A. Cramer, M. Cartwright, F. Pishdadian, and J. P. Bello, "**Weakly supervised source-specific sound level estimation in noisy soundscapes**," WASPAA 2021
F. Pishdadian, G. Wichern, and J. Le Roux, "**Finding strength in weakness: Learning to separate sounds with weak supervision,**" TASLP, 2020.

Source specific sound level estimation with multi-task learning

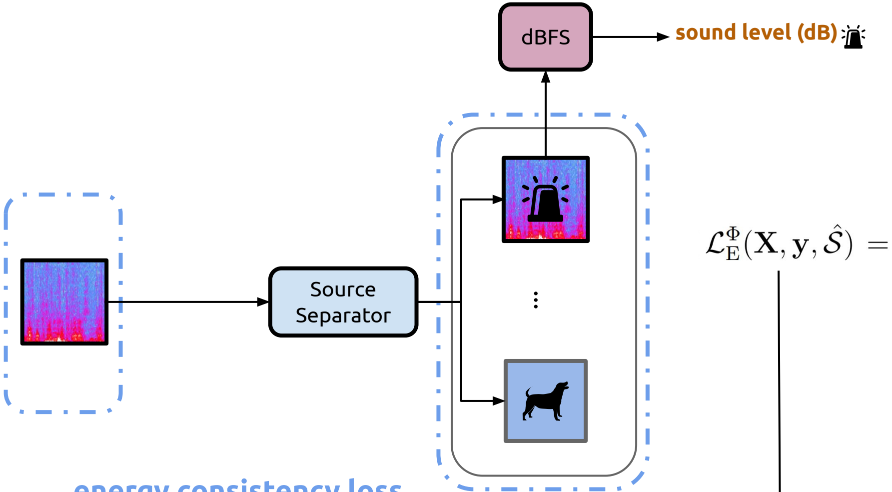
$$\mathcal{L}_{\mathrm{sssle}}(\mathbf{X}, \mathbf{y}, \hat{\mathcal{S}}) = \frac{\alpha}{|\mathcal{P}|} \sum_{\Phi \in \mathcal{P}} \mathcal{L}_{\mathrm{E}}^{\Phi}(\mathbf{X}, \mathbf{y}, \hat{\mathcal{S}}) + \mathcal{L}_{\mathrm{C}}(\mathbf{X}, \mathbf{y}, \hat{\mathcal{S}})$$



NYU TANDON SCHOOL OF ENGINEERING MARL NJIT New Jersey Institute of Technology Northwestern University

A. Cramer, M. Cartwright, F. Pishdadian, and J. P. Bello, "**Weakly supervised source-specific sound level estimation in noisy soundscapes**," WASPAA 2021
F. Pishdadian, G. Wichern, and J. Le Roux, "**Finding strength in weakness: Learning to separate sounds with weak supervision,**" TASLP, 2020.

Source specific sound level estimation with multi-task learning



[energy consistency loss]

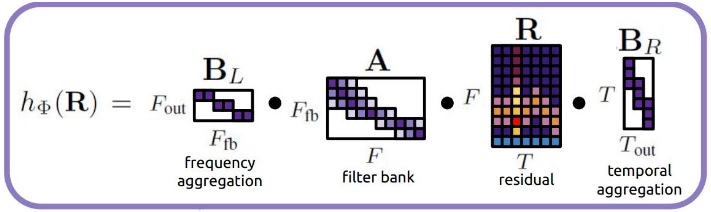
mixture **energy** = total **active source energy**

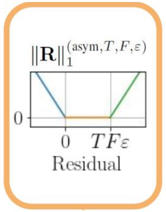
Inactive source energy = zero

active residual matrix · inactive residual matrix · frame-wise activity mask

$$\mathcal{L}_{\mathrm{E}}^{\Phi}(\mathbf{X},\mathbf{y},\hat{S}) = \frac{1}{TF_{\mathrm{fb}}}\left\| h_\Phi\left(\mathbf{M}_E \odot \mathbf{R}_{\mathrm{active}}\right)\right\|_1^{(\mathrm{asym},T,F_{\mathrm{fb}},\varepsilon)} + \frac{1}{TF_{\mathrm{fb}}}\left\| h_\Phi\left(\mathbf{M}_E \odot \mathbf{R}_{\mathrm{inactive}}\right)\right\|_1$$

resolution · projection function · asymmetric margin



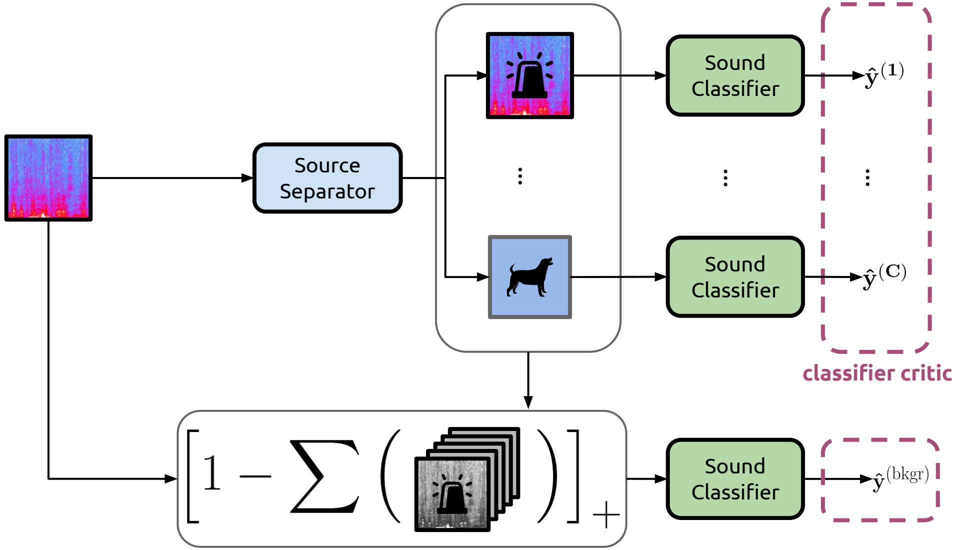


Apply energy consistency to multiple TF resolutions

$$\rightarrow \frac{\alpha}{|\mathcal{P}|}\sum_{\Phi\in\mathcal{P}} \mathcal{L}_{\mathrm{E}}^{\Phi}(\mathbf{X},\mathbf{y},\hat{S})$$

A. Cramer, M. Cartwright, F. Pishdadian, and J. P. Bello, "**Weakly supervised source-specific sound level estimation in noisy soundscapes**," WASPAA 2021
F. Pishdadian, G. Wichern, and J. Le Roux, "**Finding strength in weakness: Learning to separate sounds with weak supervision,**" TASLP, 2020.

Source specific sound level estimation with multi-task learning



classifier critic loss

separated **active source for class k**

→ **only class k present**

separated **inactive sources**

→ **no sources present**

background

→ **no sources present**

residual classification loss

$$\mathcal{L}_{\mathrm{C}}(\mathbf{X},\mathbf{y},\hat{\mathcal{S}}) = \sum_i H(\mathbf{y}\odot\mathbf{e}_i,\hat{\mathbf{y}}^{(i)}) + \beta\, H(\mathbf{0},\hat{\mathbf{y}}^{(\mathrm{bkgr})})$$

source classification loss

A. Cramer, M. Cartwright, F. Pishdadian, and J. P. Bello, "**Weakly supervised source-specific sound level estimation in noisy soundscapes**," WASPAA 2021
F. Pishdadian, G. Wichern, and J. Le Roux, "**Finding strength in weakness: Learning to separate sounds with weak supervision,**" TASLP, 2020.

Training and evaluation

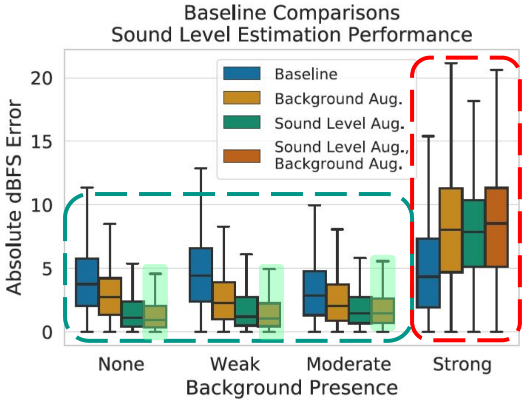
- Train and evaluate with 4-second synthetic soundscapes (16kHz) containing **urban sound events** (from UrbanSound8K) and different levels of **urban background noise** (SONYC-Backgrounds ✶)
    - Different background levels: **-50/-20/0 dB LUFS** (**weak/moderate/strong**)**, no background**

- Evaluate with respect to both **source separation performance** (SI-SDR improvement) and **sound level estimation** performance (absolute dBFS error), comparing with weakly supervised source separation baseline

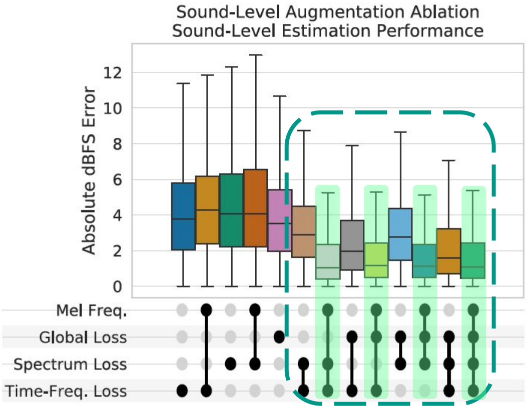
✶ SONYC-Backgrounds: https://doi.org/10.5281/zenodo.5129078

Salamon et al., "**A dataset and taxonomy for urban sound research,**" ACM Multimedia, 2014.

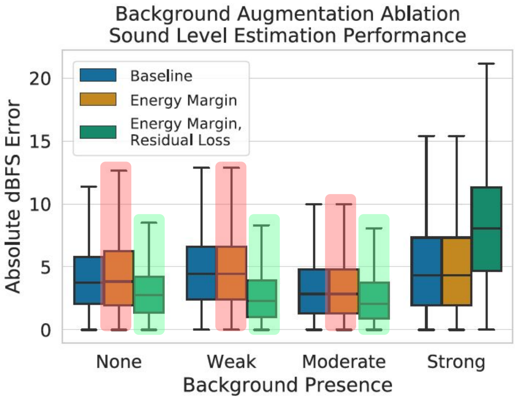
Experimental Results



SSSLE performance (and source separation performance) are improved in up to moderate background conditions!



Using multiple time-frequency resolutions is beneficial



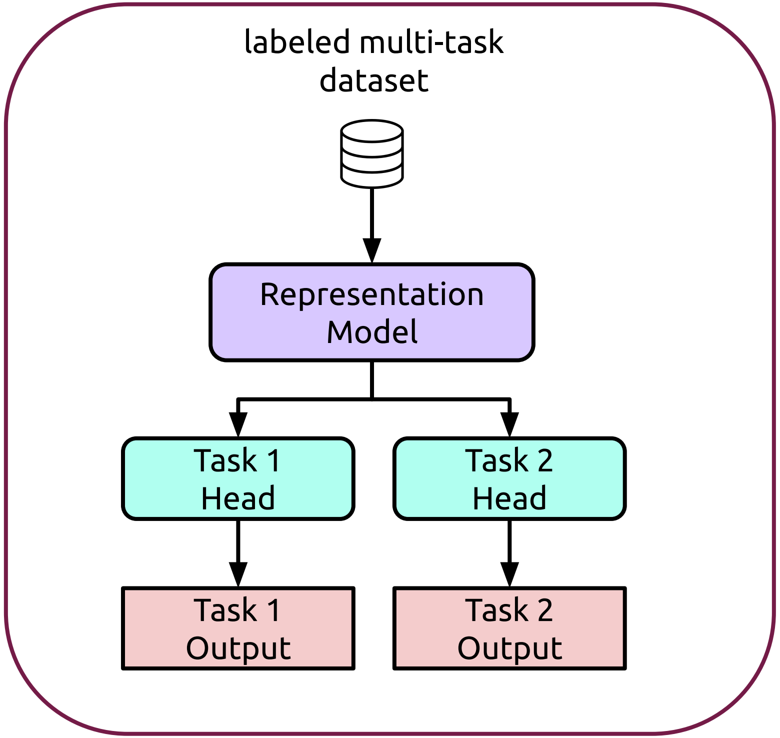
The background classification loss is crucial when using an asymmetric margin

SSSLE models can be trained with only clip-level source labels using multi-task learning!

A. Cramer, M. Cartwright, F. Pishdadian, and J. P. Bello, "**Weakly supervised source-specific sound level estimation in noisy soundscapes**," WASPAA 2021

Flavors of transfer learning

parallel transfer learning
(a.k.a. multi-task learning)



multi-task output
structure

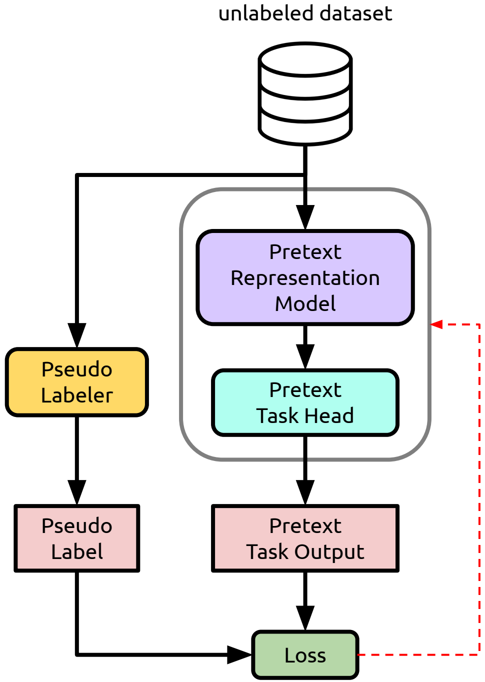
identifying and capitalizing on **multi-task output structure** improves **cross-task transfer**!

What if there still isn't enough data?

We can transfer knowledge from **other models** that have already learned useful representations!

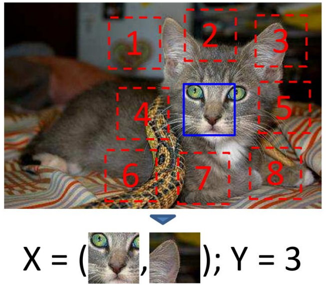
Sequential transfer using multi-modal self-supervision

Self-supervised transfer learning



- Train models with **unlabeled datasets** using “**pretext**” tasks with corresponding **pseudo-labels**

- Model **implicitly learns representation** encoding useful perceptual and/or semantic information

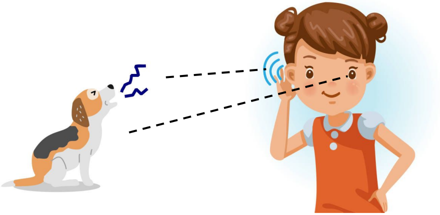


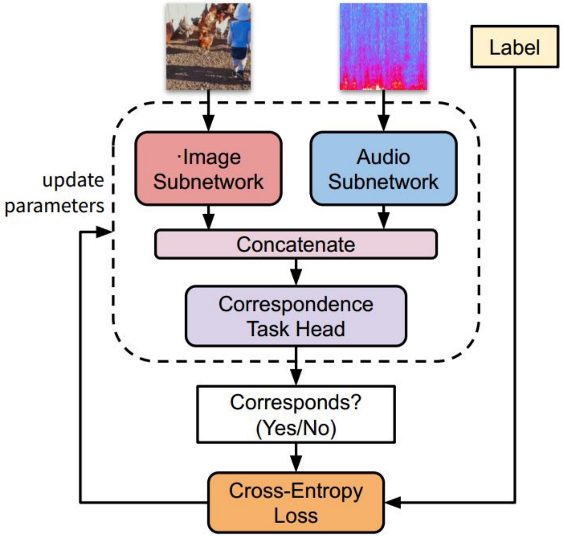
Audio-visual correspondence (AVC)

- **Auditory** and **visual** stimuli often correspond to a **common underlying source** → ***audio-visual correspondence***

- AVC defines a **simple self-supervision task** that learns embeddings from **unlabeled videos** that are useful for both **downstream audio and visual tasks**!

- Can we **better understand** how to **train effective audio embeddings using AVC**?



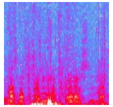


R. Arandjelovic and A. Zisserman, "**Look, listen and learn,**" ECCV 2017

How do design choices affect downstream audio classification performance?

Upstream Training

is it beneficial to use an audio-informed **input representation?** →



linear spectrogram (257 bins)
mel spectrogram (128 bins)
mel spectrogram (256 bins)

is it important that the **training data** for the **embedding match the downstream classification** task? →

“music” subset - instruments (+ tools)
“environmental” subset - urban sounds

how much training data is sufficient for training the embedding? →

Periodic model checkpoints

Downstream Evaluation

- Use embedding models as **feature extractors**
- Small multi-layer perception trained for audio classification

NYU TANDON SCHOOL OF ENGINEERING MARL

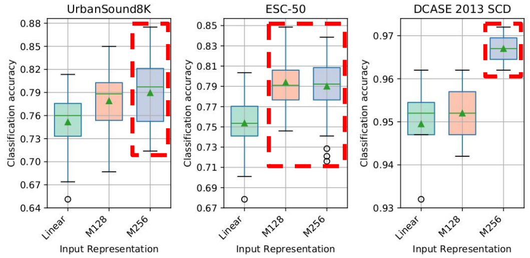
A. Cramer, H. H. Wu, J. Salamon, J. P. Bello, “**Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings**,” ICASSP 2019

Training and evaluation

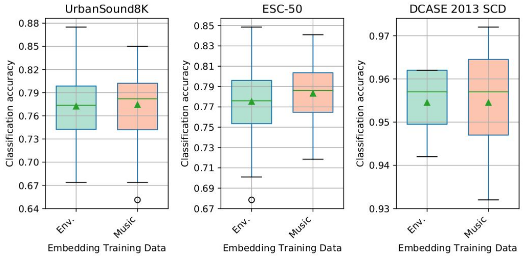
- Train and validate upstream AVC model using subsets of AudioSet
  - Generate pairs of 1 second audio clips and random overlapping video frame for positive, shuffle pairs for negatives
  - Data augmentation - video: crop, brightness, contrast, saturation; audio: loudness)
- Train and evaluate downstream models using multi-class audio classification datasets (UrbanSound8K, ESC-50, DCASE 2013 Scene Classification)
  - Obtain clip-level predictions by averaging framewise predictions on 1 second overlapping frames

A. Cramer, H. H. Wu, J. Salamon, J. P. Bello, "**Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings**," ICASSP 2019

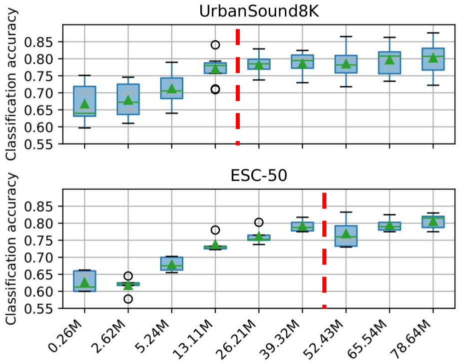
Experimental Results



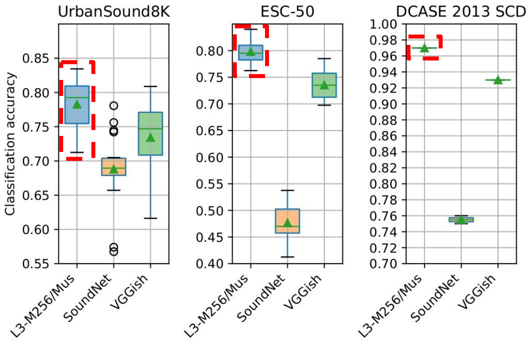
Mel spectrograms with half the number of bins outperform linear spectrograms!



Training with either subsets produces similar results — semantic content is less relevant than a strong AVC signal



25-50% of AVC examples needed to match best performance

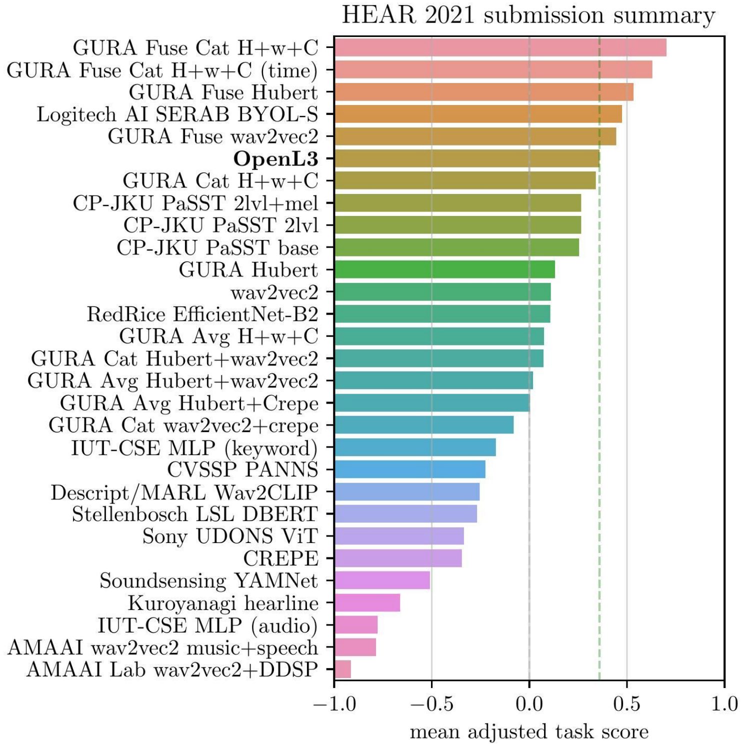


L3-Net outperforms other competitive embeddings!

`$ pip install` **`openl3`**

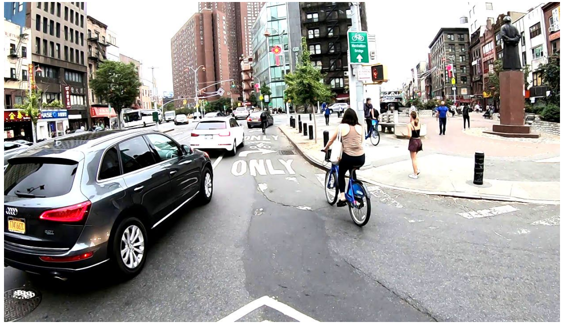
A. Cramer, H. H. Wu, J. Salamon, J. P. Bello, "**Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings**," ICASSP 2019

OpenL3 remains competitive!



J. Turian, et al. "**HEAR: Holistic evaluation of audio representations.**" NeurIPS 2021 Competitions and Demonstrations Track. PMLR, 2022

Urban sound exhibits natural urban rhythm



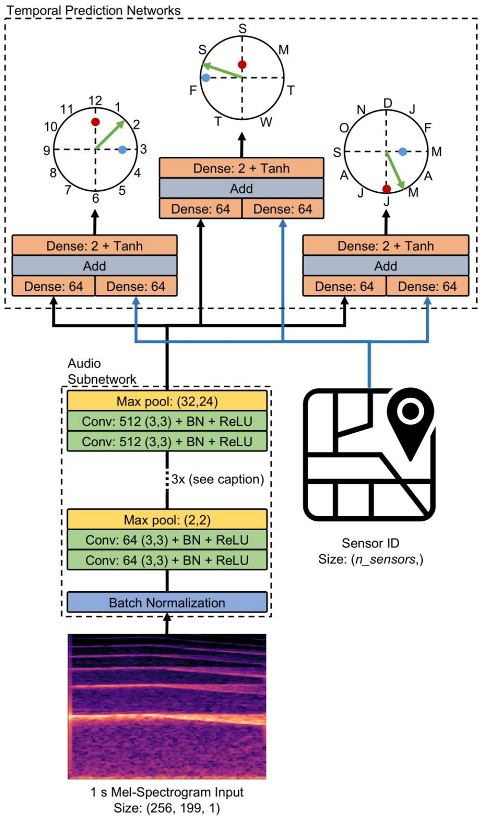
Soundscapes often sound different depending on the **time of day**, **day of the week**, or **month of the year**

Temporal cycle prediction

- Leverage **self-supervised learning** with an L3-like architecture using ***temporal cycle prediction*** as the **pretext task**

- Model predicts the phase in **daily**, **weekly**, and **yearly cycles** using only sensor network audio

- Condition on sensor ID to help prevent model from overfitting to sensor characteristics



NYU TANDON SCHOOL OF ENGINEERING MARL

M. Cartwright, **A. Cramer**, J. Salamon, and J.P. Bello. "**TriCycle: Audio representation learning from sensor network data using self-supervision**," WASPAA, 2019

Training and evaluation

- **Train with unlabeled clips** from a year of **urban soundscape recordings** obtained from the SONYC acoustic sensor network
  - Sample evenly in time, focusing on potentially meaningful events by randomly selecting recordings for each hour in the top 15th percentile of SPL difference:

$$\sqrt{\sum_{n=0}^{79}(d_{m,n} - d_{m,n-1})^2}$$

- Evaluate with a labeled **urban sound tagging dataset** from temporally-disjoint SONYC recordings (SONYC-UST v1)

M. Cartwright, **A. Cramer**, J. Salamon, and J.P. Bello. "**TriCycle: Audio representation learning from sensor network data using self-supervision**," WASPAA, 2019
M. Cartwright, A. E. M. Mendez, G. Dove, **A. Cramer**, V. Lostanlen, H. Wu, J. Salamon, O. Nov, and J. P. Bello, "SONYC urban sound tagging (SONYC-UST): a multilabel dataset from an urban acoustic sensor network," Zenodo, (2019). https://zenodo.org/record/2590742

Experimental Results

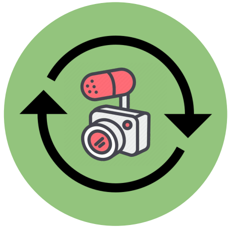
Able to produce embeddings **comparable with L3-Net**
on downstream urban sound tagging

Name	Init.	TriCycle Train	Variation	MAD Day	MAD Week	MAD Year	UST F1@0.5	UST P@0.5	UST R@0.5	UST AUPRC	Sensor Acc.
l3	L^3-Net	No	—	—	—	—	0.638	**0.767**	0.547	**0.751**	0.792
rand	Rand.	No	—	—	—	—	0.531	0.697	0.429	0.632	0.721
rand-tc	Rand.	Yes	—	0.480	0.508	0.562	0.622	0.734	0.540	0.712	0.781
l3-tc-llr	L^3-Net	Yes	Low LR	0.370	0.531	0.540	0.638	0.764	0.548	0.739	0.824
l3-tc-hlr	L^3-Net	Yes	High LR	**0.338**	0.443	0.545	0.638	0.749	0.556	0.737	**0.851**
rand-tc-rs	Rand.	Yes	Rand. Sampling	0.416	0.508	0.542	0.610	0.739	0.520	0.702	0.801
rand-tc-pcen	Rand.	Yes	PCEN	0.351	**0.423**	**0.444**	**0.650**	**0.767**	**0.564**	0.744	0.831

NYU TANDON SCHOOL OF ENGINEERING MARL

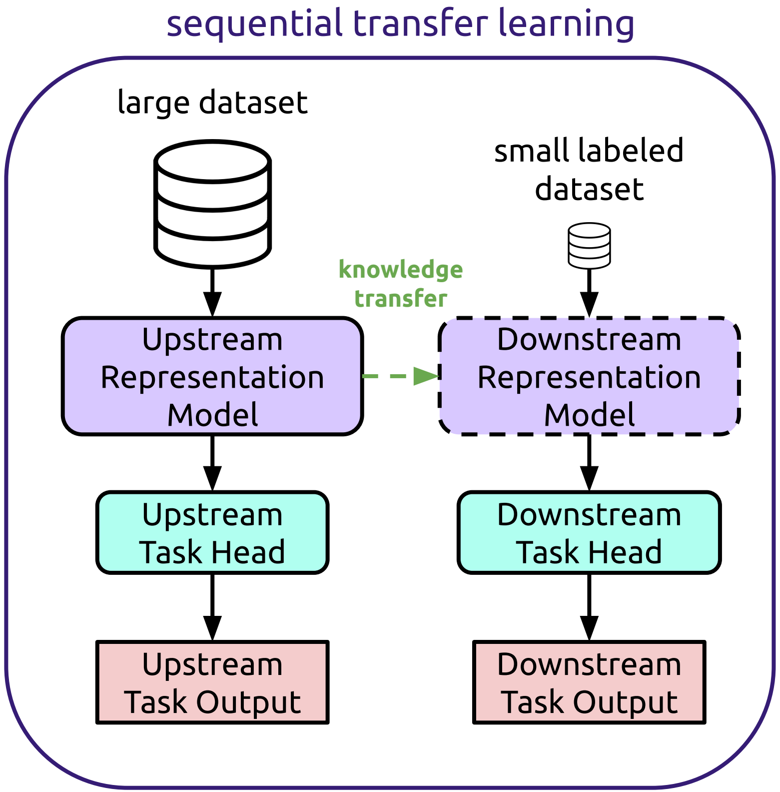
M. Cartwright, **A. Cramer**, J. Salamon, and J.P. Bello. "**TriCycle: Audio representation learning from sensor network data using self-supervision**," WASPAA, 2019
M. Cartwright, A. E. M. Mendez, G. Dove, **A. Cramer**, V. Lostanlen, H. Wu, J. Salamon, O. Nov, and J. P. Bello, "SONYC urban sound tagging (SONYC-UST): a multilabel dataset from an urban acoustic sensor network," Zenodo, (2019). https://zenodo.org/record/2590742

Flavors of transfer learning

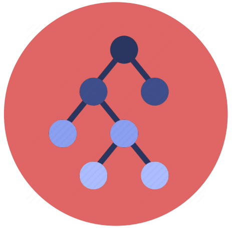


multi-modal
self-supervision

natural **multi-modal correspondence** provide structure **without human annotations**!

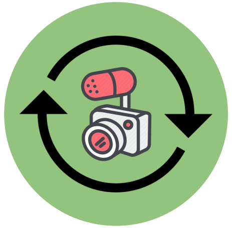


What's next?

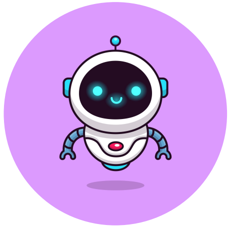


sound source hierarchies

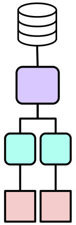
multi-task output structure



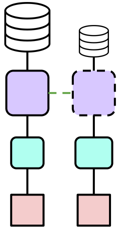
multi-modal self-supervision



embodied agents



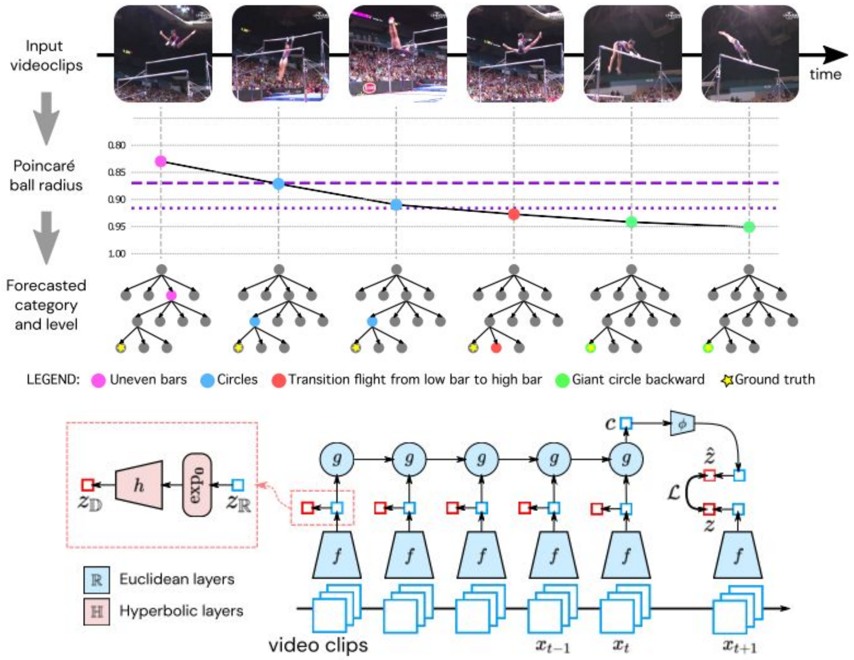
Transfer Learning



Beyond predefined hierarchies

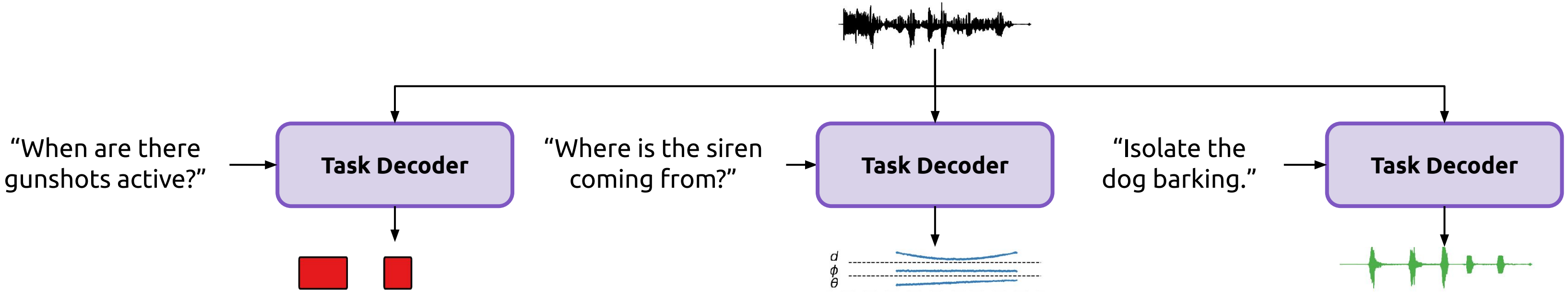
- Unlike **Euclidean** embeddings, **Hyperbolic embeddings** can directly **encode hierarchical structure** without distortion

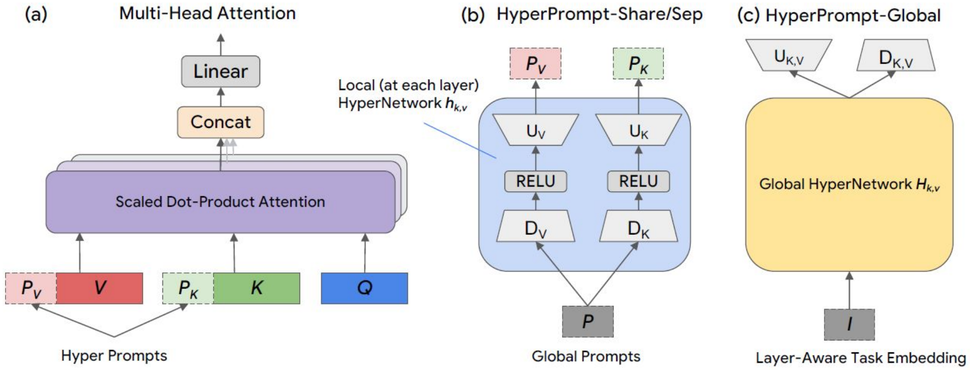
- **Self-supervised contrastive predictive coding** with **hyperbolic embeddings** can **implicitly learn hierarchies** and **uncertainty**



D. Suris, R. Liu, and C. Vondrick, “**Learning the Predictability of the Future**,” CVPR *2021*

Beyond predefined tasks

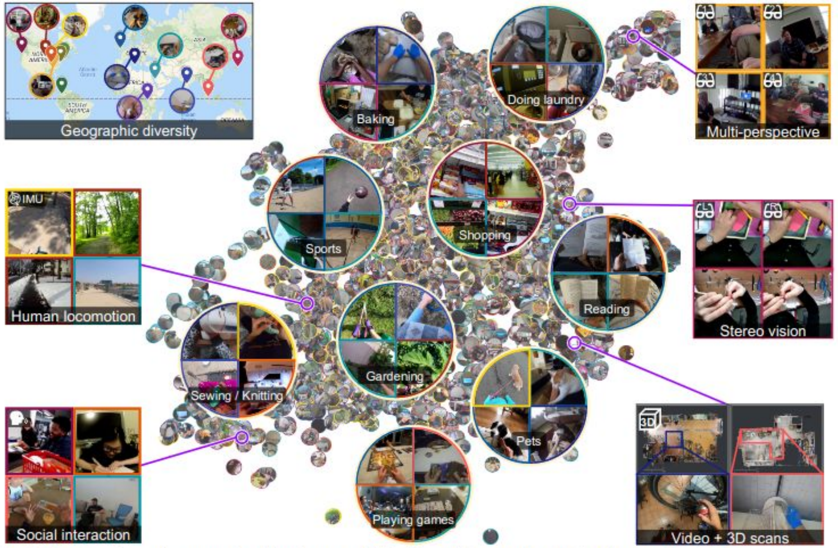




Y. He, et al., "**HyperPrompt: Prompt-based Task-Conditioning of Transformers**," ICML *2022*

Beyond audio-visual correspondence

- Egocentric videos may better align to human perception than videos filmed with handheld devices
- Telemetry data like accelerometry, recording timestamps, and location can be used for further self-supervision
- Embodied agents that interact with the environment may even more effectively align to human perception



K. Grauman et al., “**Ego4D: Around the World in 3,000 Hours of Egocentric Video**,” 2021.
S. Zhang et al., “**EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices**,” *ECCV*, 2022.

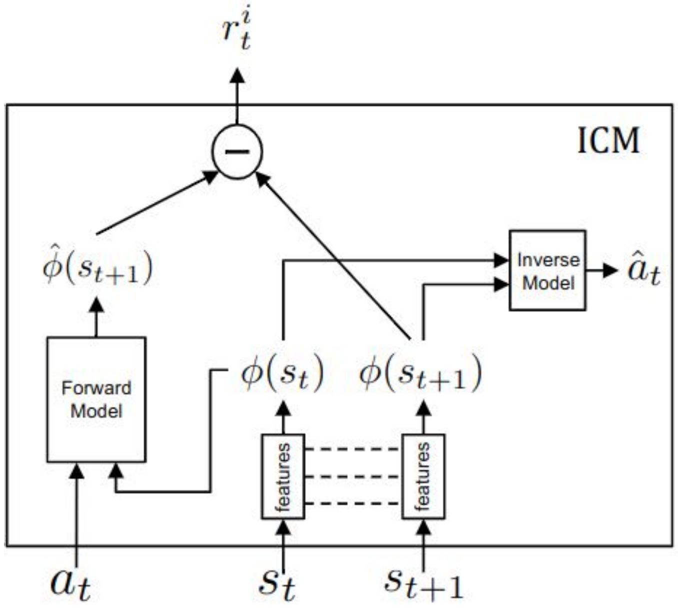
Embodied navigation

Photorealistic renders from point clouds
+ spatialized audio from realistic RIRs

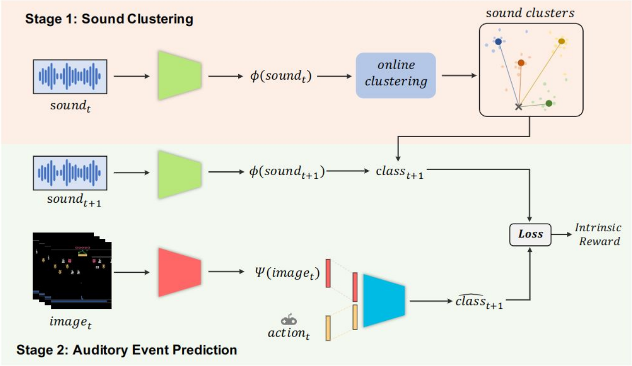
Agent must navigate to audible goal using audio and visual sensory inputs

C. Chen *et al.*, "**SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning**," *ArXiV*, 2022
C. Chen *et al.*, "**SoundSpaces: Audio-Visual Navigation in 3D Environments**," 2020.
C. Gan *et al.*, "**ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation**," *NeurIPS*, 2021

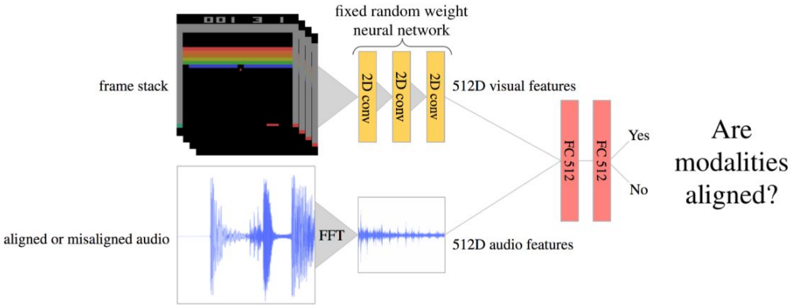
Self-supervision through intrinsic valuation



Intrinsic Curiosity Module: predict action from state embeddings before and after action



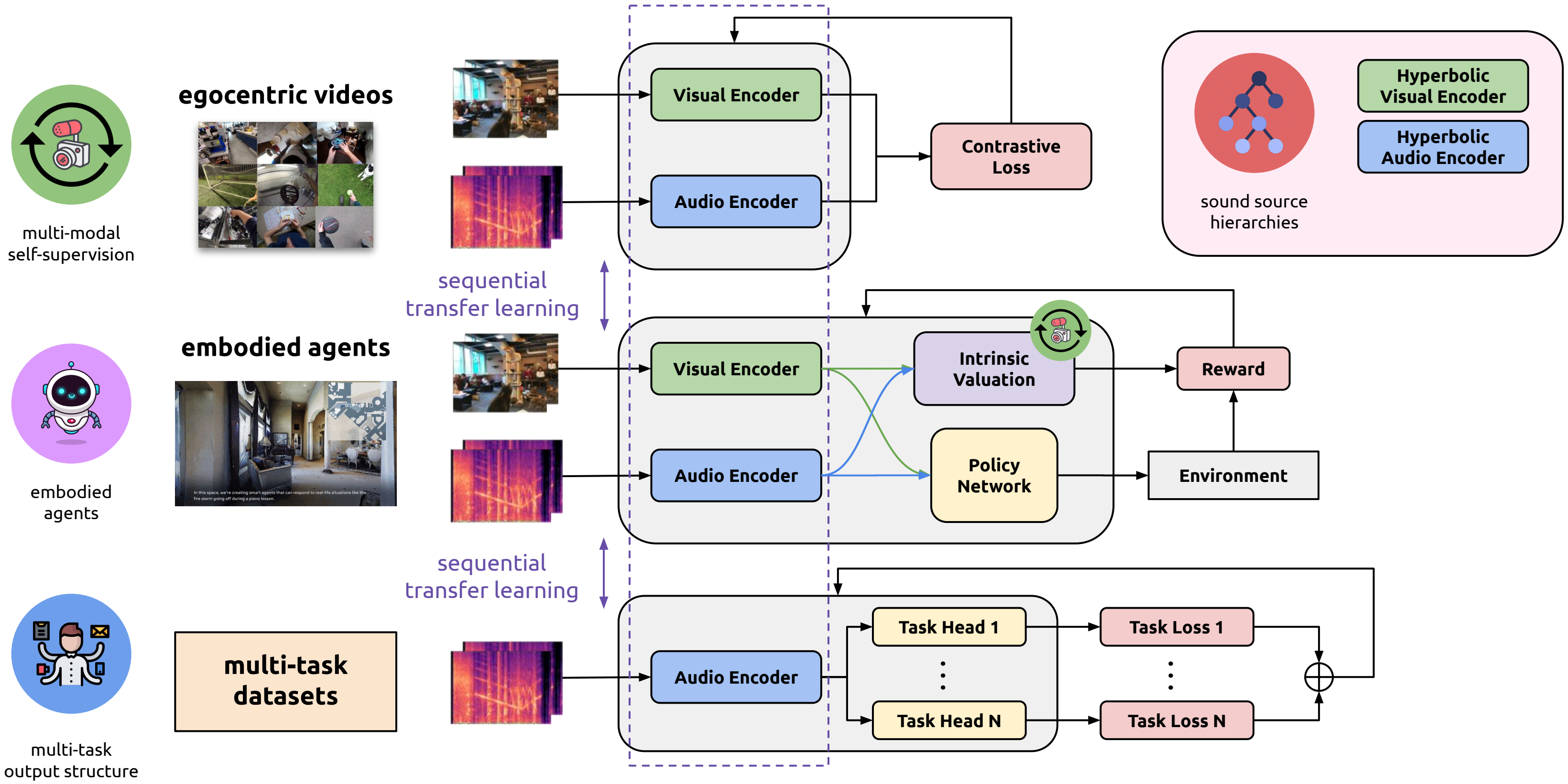
Predict auditory events from visual stream



Audio-visual alignment

D. Pathak *et al.*, “**Curiosity-driven Exploration by Self-supervised Prediction**,” *ICML*, 2017.
C. Gan *et al.*, “**Noisy Agents: Self-supervised Exploration by Predicting Auditory Events**,” 2020.
V. Dean, S. Tulsiani, and A. Gupta, “**See, Hear, Explore: Curiosity via Audio-Visual Association**,” *NeurIPS*, 2020.

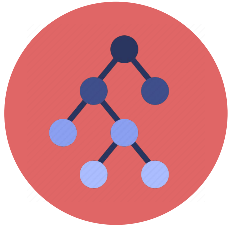
Putting it all together



In summary:

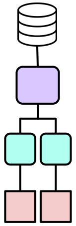
Incorporating natural structure is a promising path towards alleviating data-scarcity and improving robustness in machine listening models!

parallel transfer learning
(a.k.a. multi-task learning)



sound source hierarchies

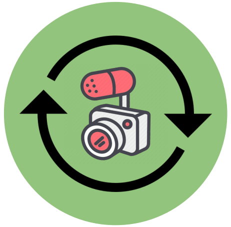
hierarchical structure improves model robustness!



multi-task output structure

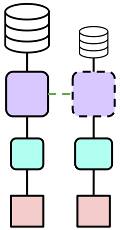
identifying and capitalizing on multi-task structure improves cross-task transfer!

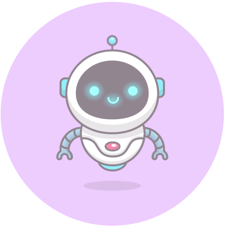
sequential transfer learning



multi-modal self-supervision

natural multi-modal correspondence provide structure without human annotations!





embodied agents

embodied navigation may provide structure better aligning with everyday experiences

Thank you!


Juan Pablo Bello


Yao Wang


Ivan Selesnick


Michael Mandel


Dan Ellis


Ho-Hsiang Wu


Justin Salamon


Charlie Mydlarz


Mark Cartwright


Ana Elisa Méndez


Yu Wang


Vincent Lostanlen


Andrew Farnsworth


Benjamin Van Doren


Fatemeh Pishdadian


Magdalena Fuentes


Mahin Salman


Bea Steers


Gordon Wichern


Jonathan Le Roux


Rafael Valle


Bryan Catanzaro


Sivan Ding


Olive

\+ many others!