Unsupervised Discriminative Learning of Sounds for Audio Event Classification

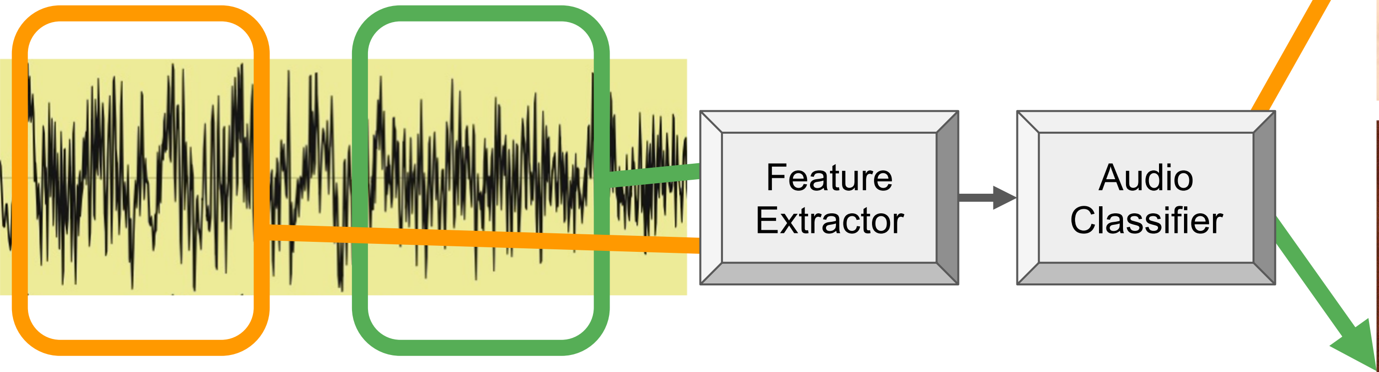
Sascha Hornauer **Ke Li** **Stella X. Yu**

UC Berkeley / ICSI

Shabnam Ghaffarzadegan **Liu Ren**

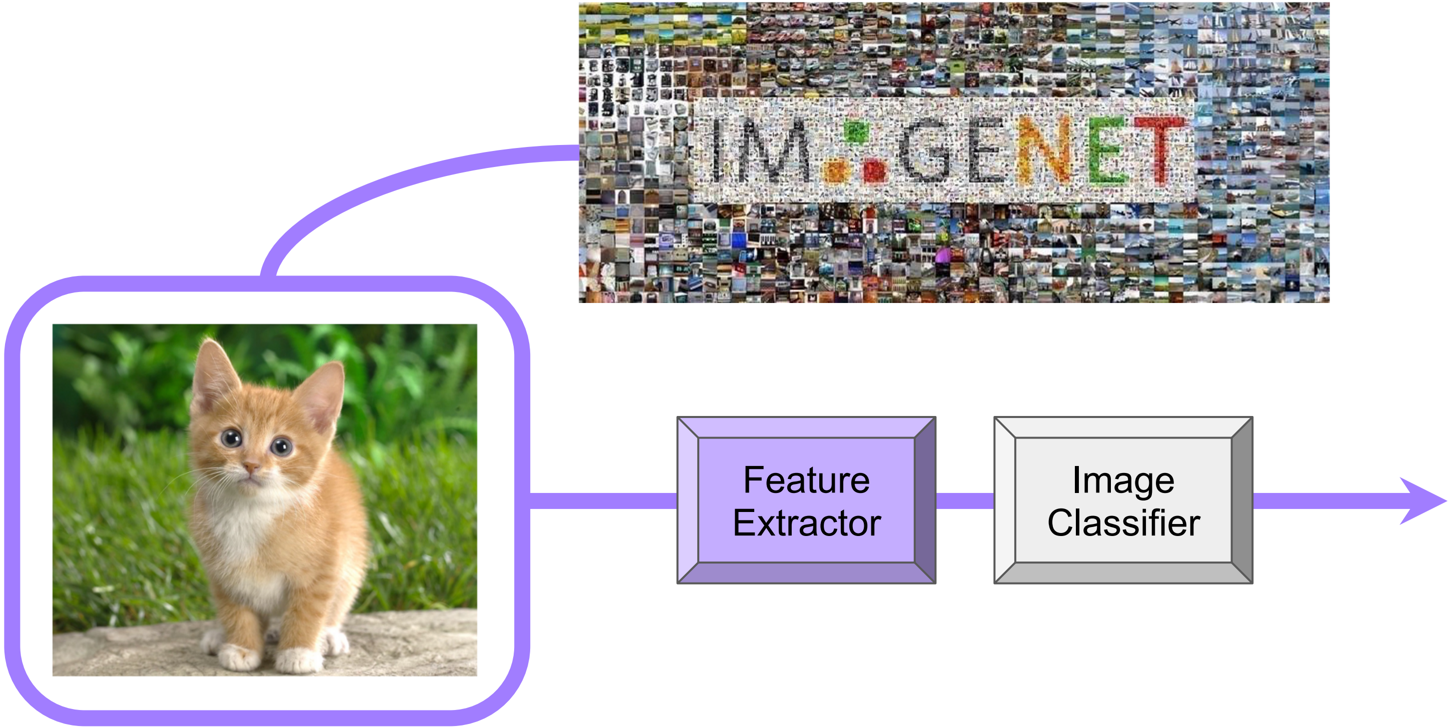
Robert Bosch LLC

Task: Audio Event Classification



Phone image by: https://commons.wikimedia.org/wiki/User:Holger.Ellgaard CC BY-SA 3.0

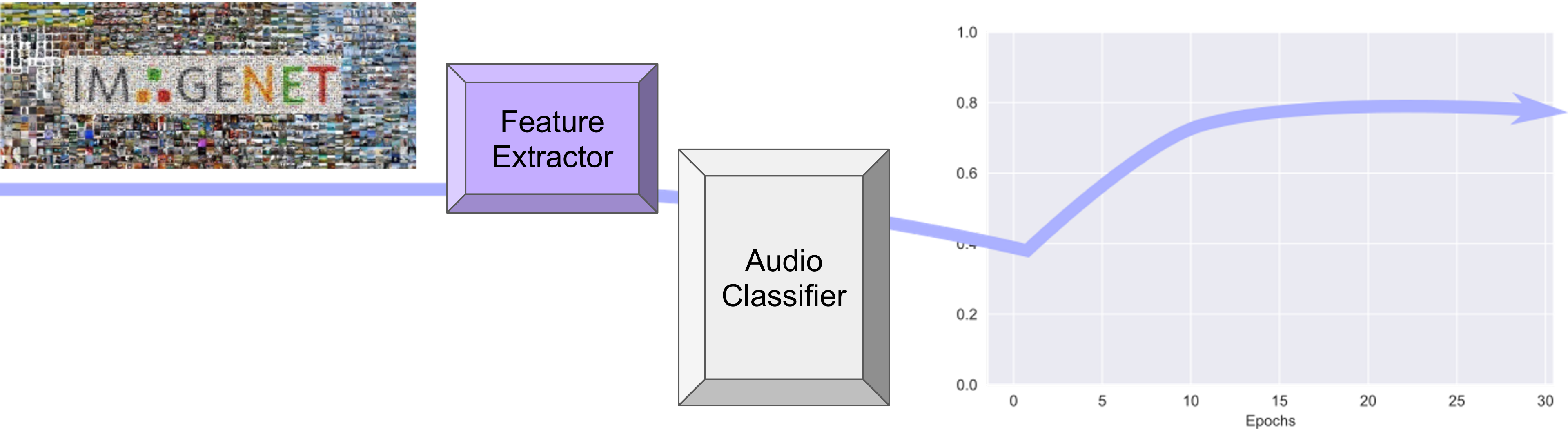
State-of-the-art Relies on ImageNet Pre-Training



Dog	0.10
Cat	**0.76**
Boat	0.05
Plane	0.09

Andrey Guzhov, Federico Raue, Jorn Hees, and Andreas Dengel, “Esresnet: Environmental sound classification based on visual domain models,” arXiv preprint arXiv:2004.07301, 2020.

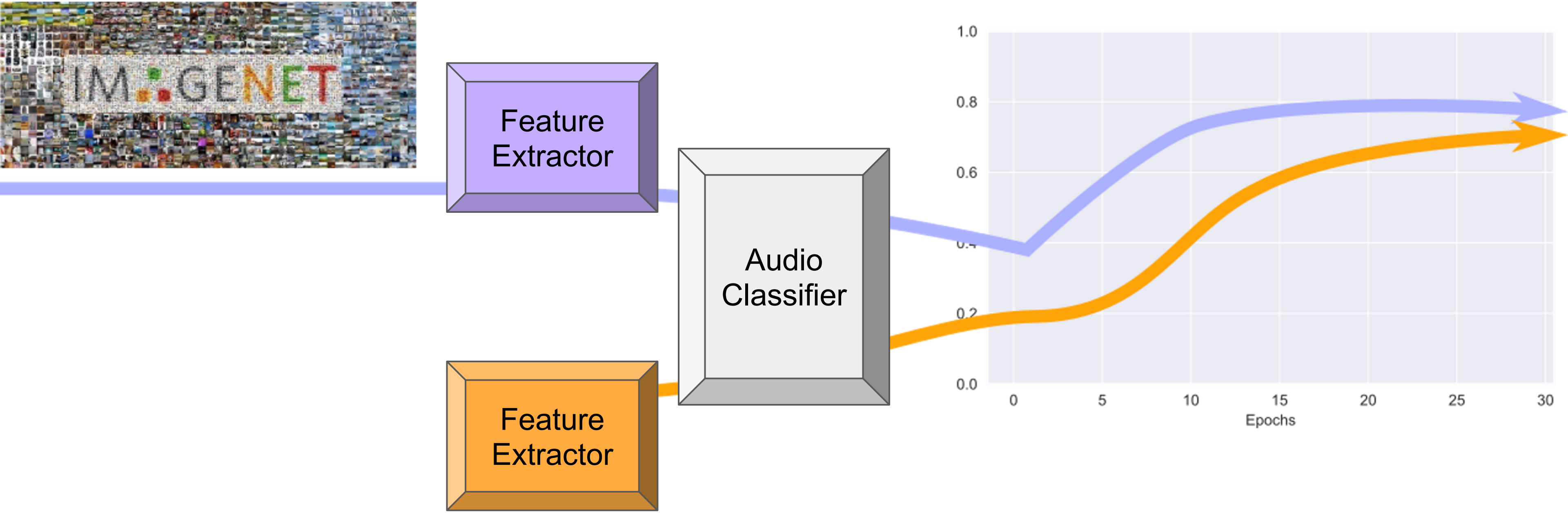
Subsequent Classifier Fine-Tuning on Audio for Quick Convergence



Feature Extractor

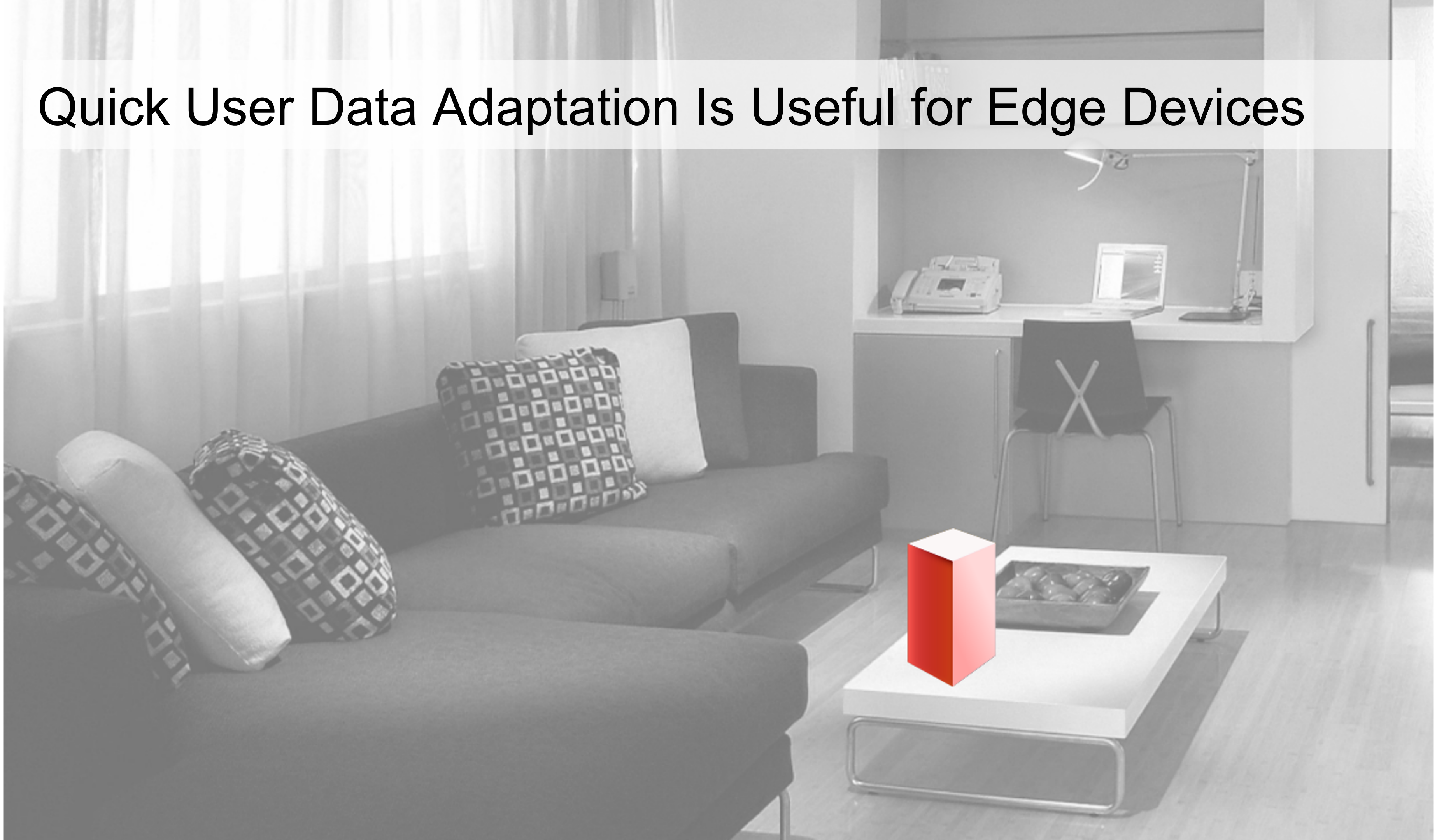
Audio Classifier

Pro: Quicker Accuracy Gain in Early Epochs Than No Pretraining

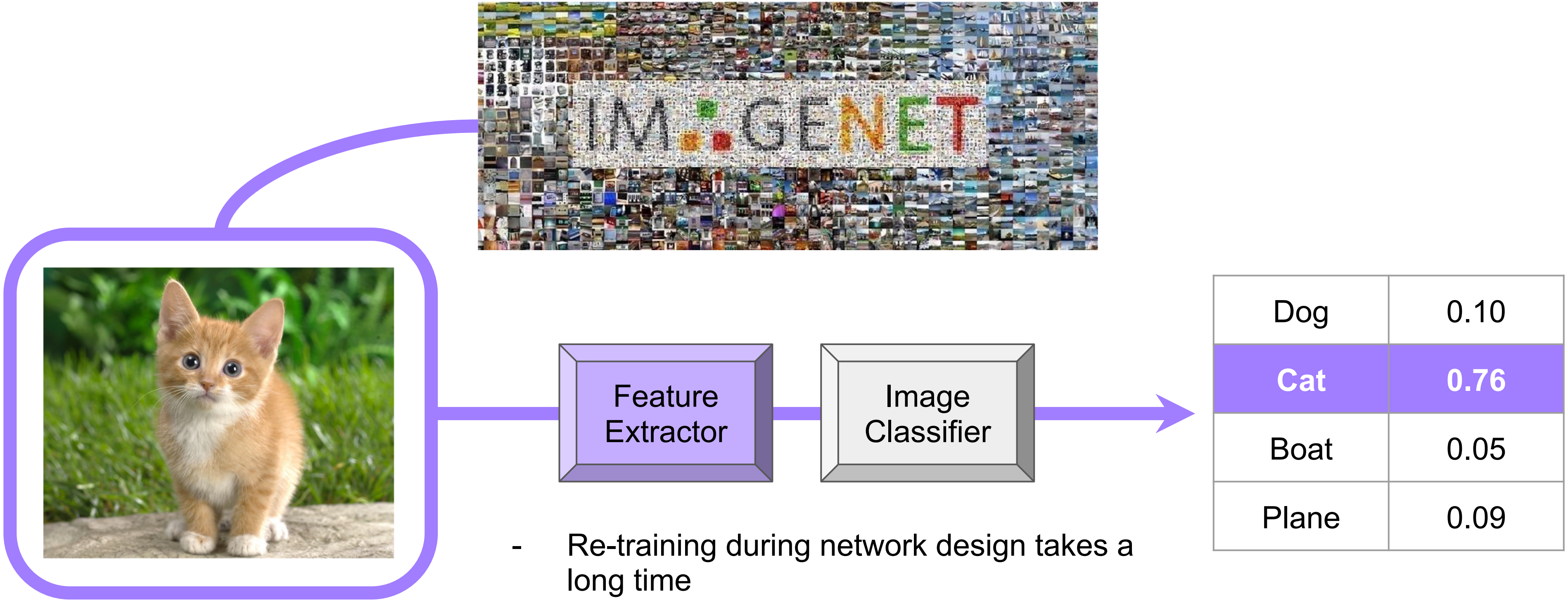


No Pretraining

Quick User Data Adaptation Is Useful for Edge Devices



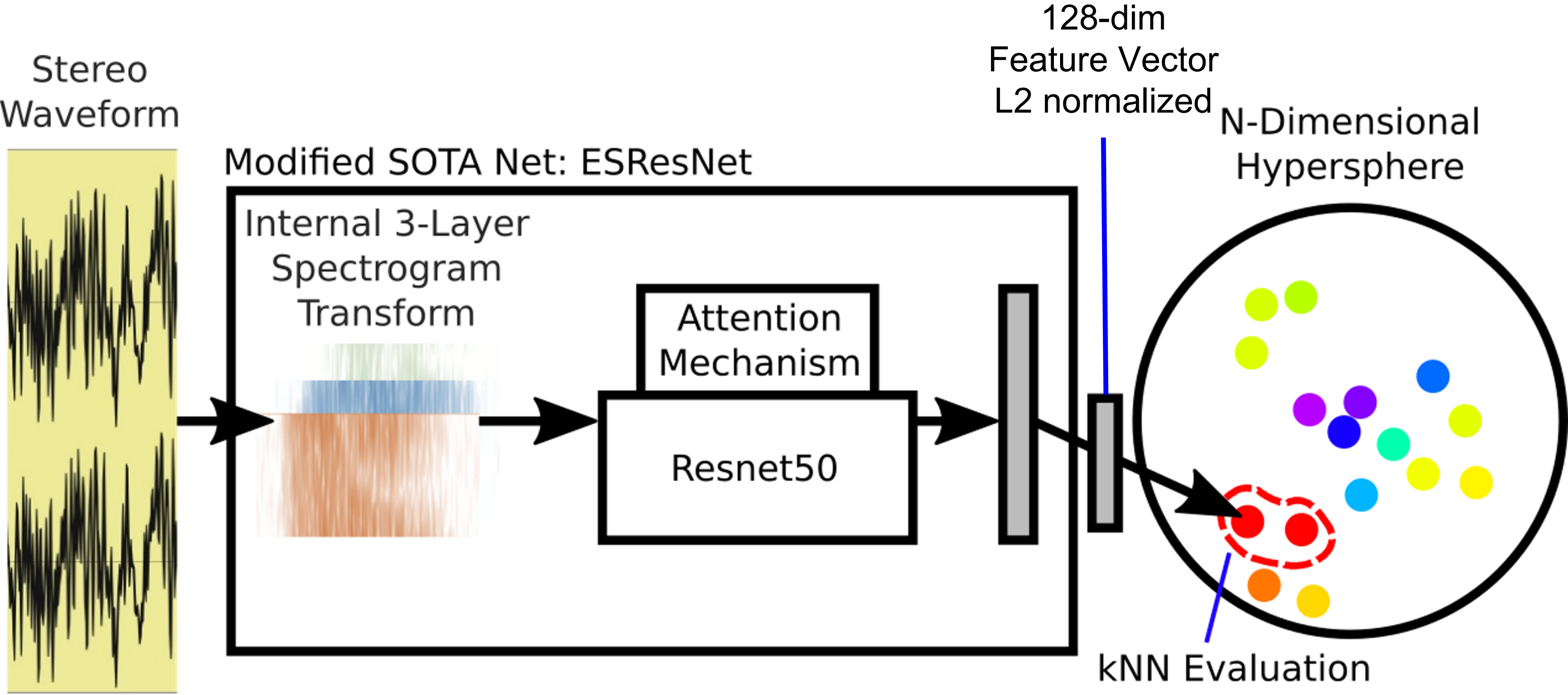
Con: ImageNet PreTraining Needs Large, Image-Parsing Nets



Dog	0.10
Cat	**0.76**
Boat	0.05
Plane	0.09

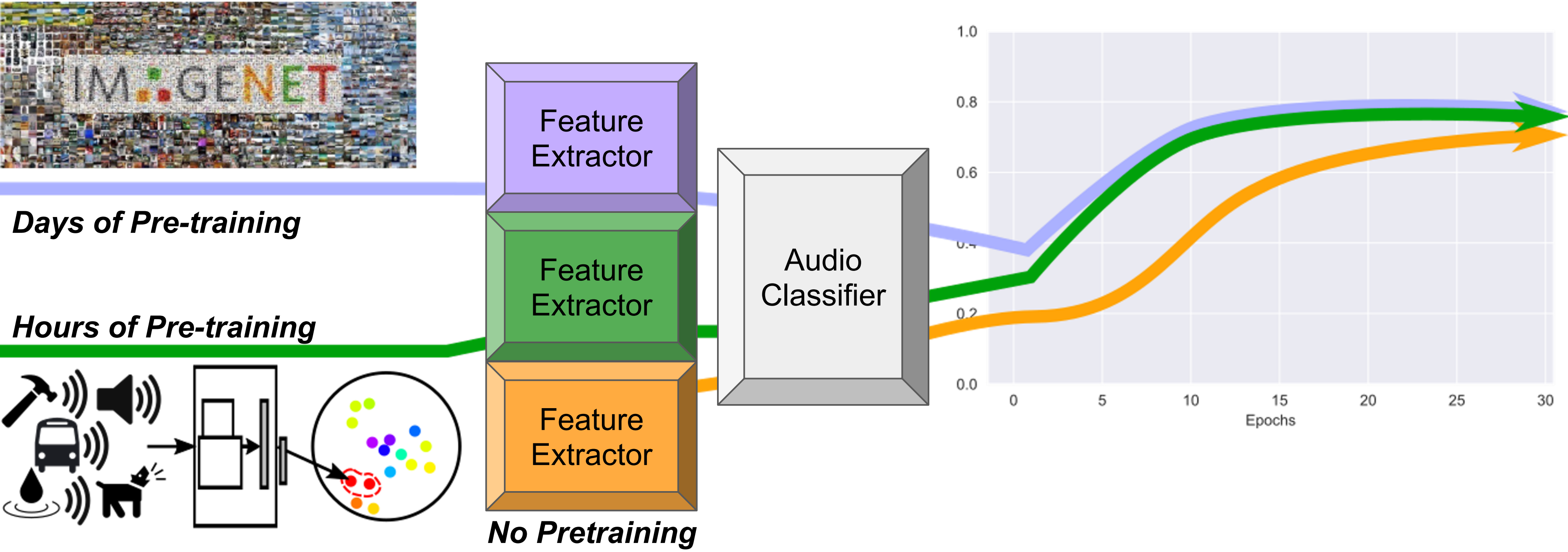
- Re-training during network design takes a long time
- Image data requires layers with many parameters

Our Idea: Pretrain on Sound Data Directly without Supervision



Stereo Waveform
Modified SOTA Net: ESResNet
Internal 3-Layer Spectrogram Transform
Attention Mechanism
Resnet50
128-dim Feature Vector L2 normalized
N-Dimensional Hypersphere
kNN Evaluation

Using Audio Data Alone, Faster Pre-Training Performs On Par



Days of Pre-training

Hours of Pre-training

Feature Extractor

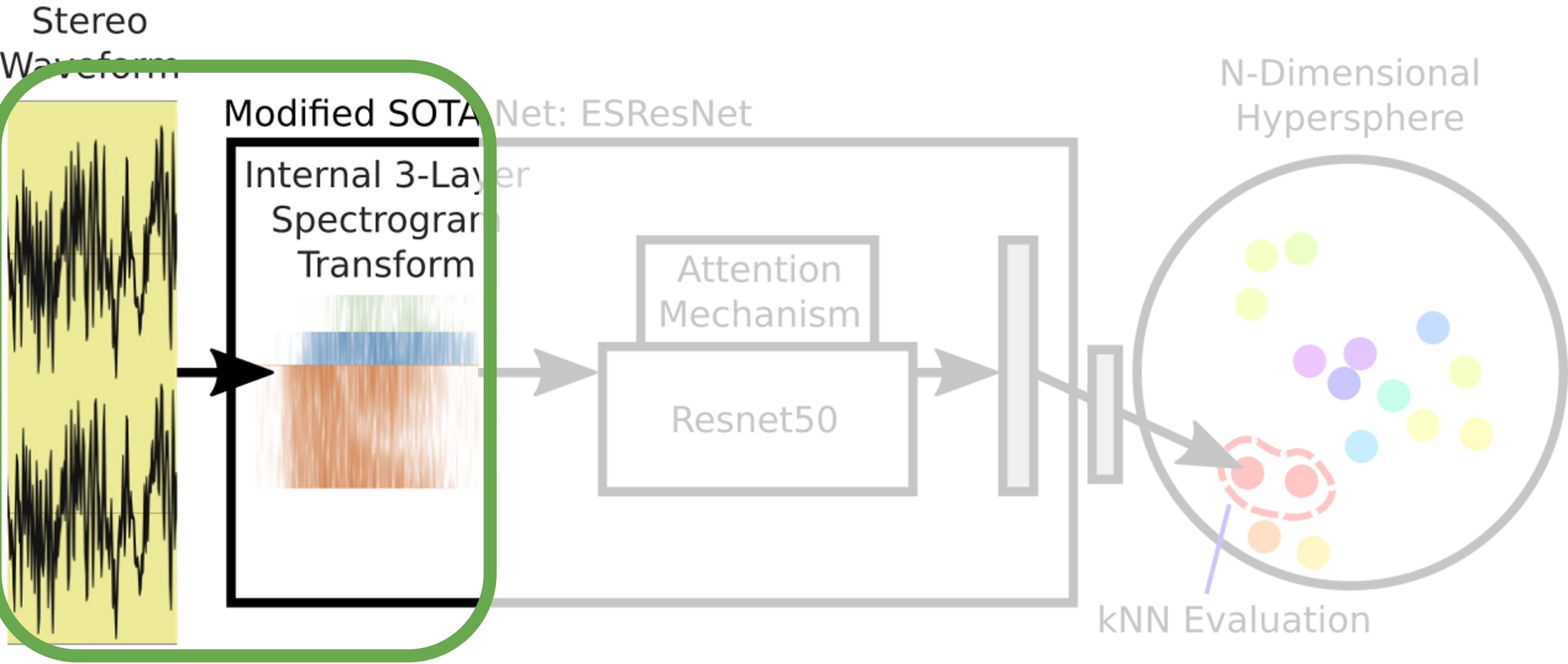
Feature Extractor

Feature Extractor

No Pretraining

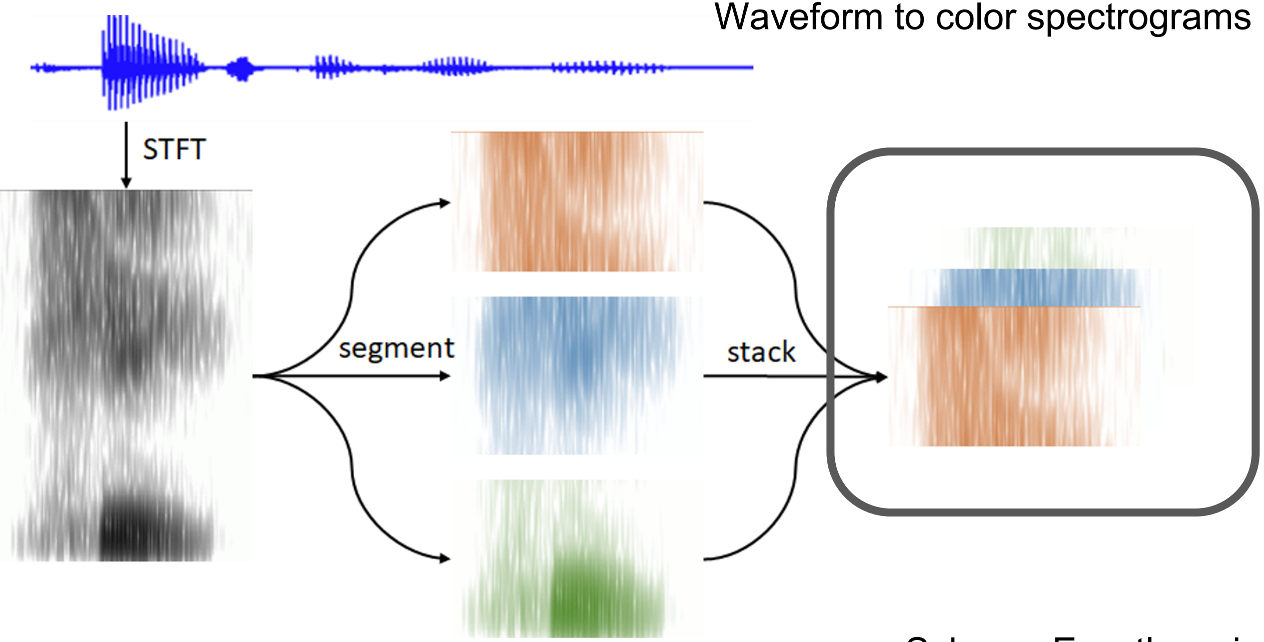
Audio Classifier

Details of the Sound Encoding


Stereo Waveform
Modified SOTA Net: ESResNet
Internal 3-Layer Spectrogram Transform
Attention Mechanism
Resnet50
N-Dimensional Hypersphere
kNN Evaluation

Frequency Bins are Distributed Along Channel Dimension

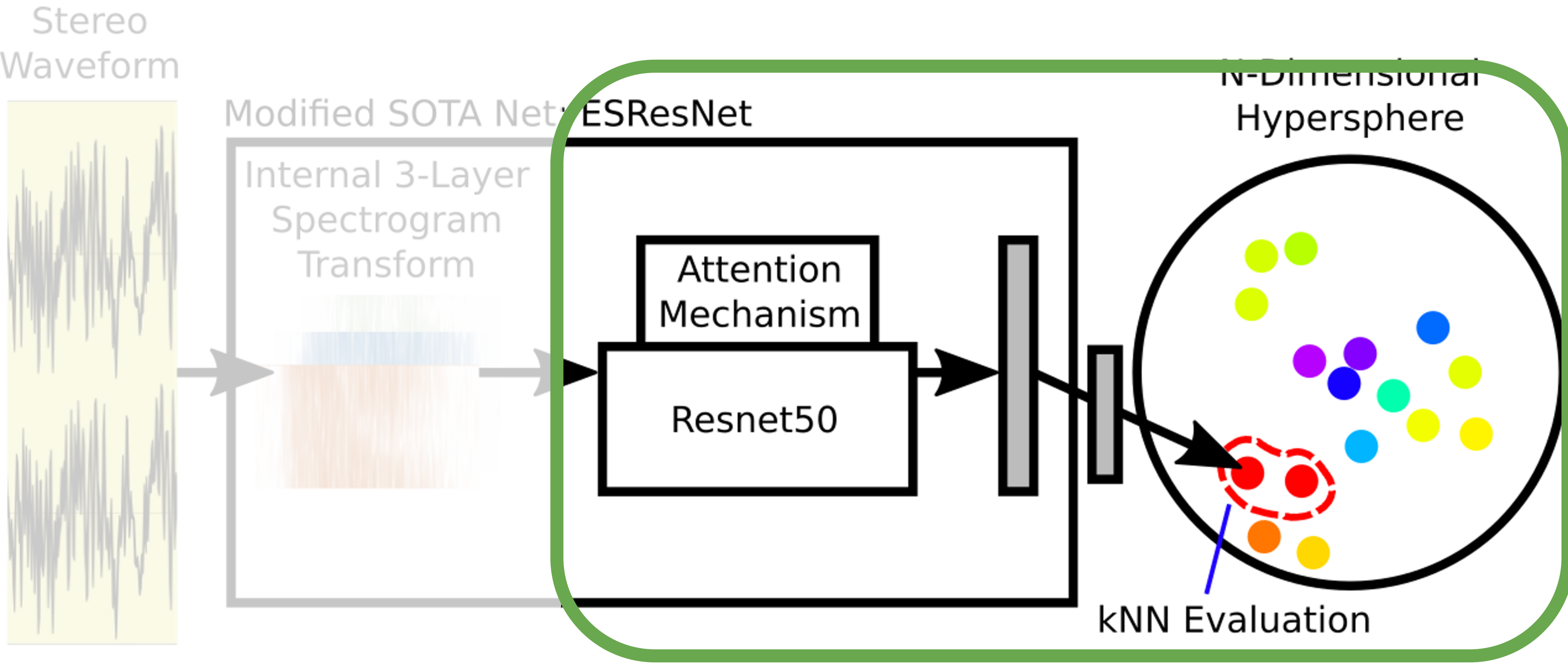
Waveform to color spectrograms



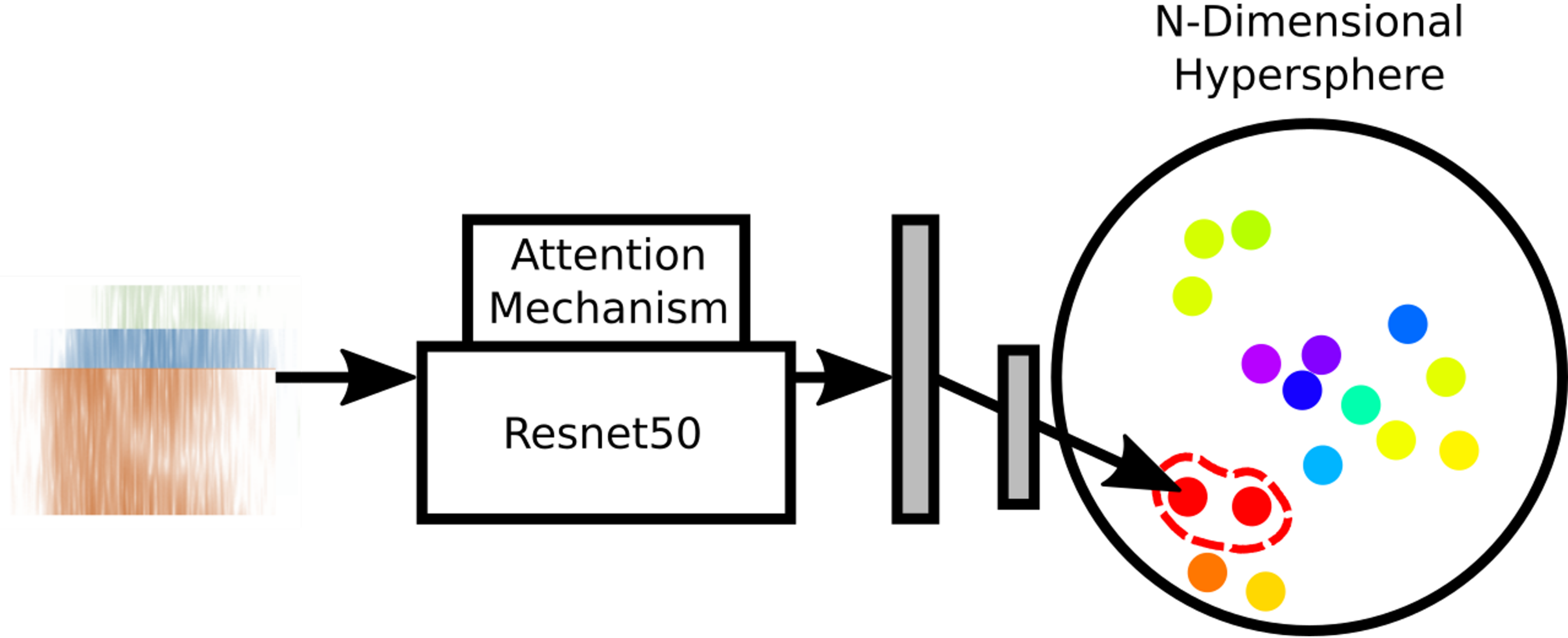
Each resulting instance has three Channels, containing different parts of the frequency spectrum.

Schema Exactly as in ESResnet for Direct Comparison

Details of the Unsupervised Embedding for **Pretraining**

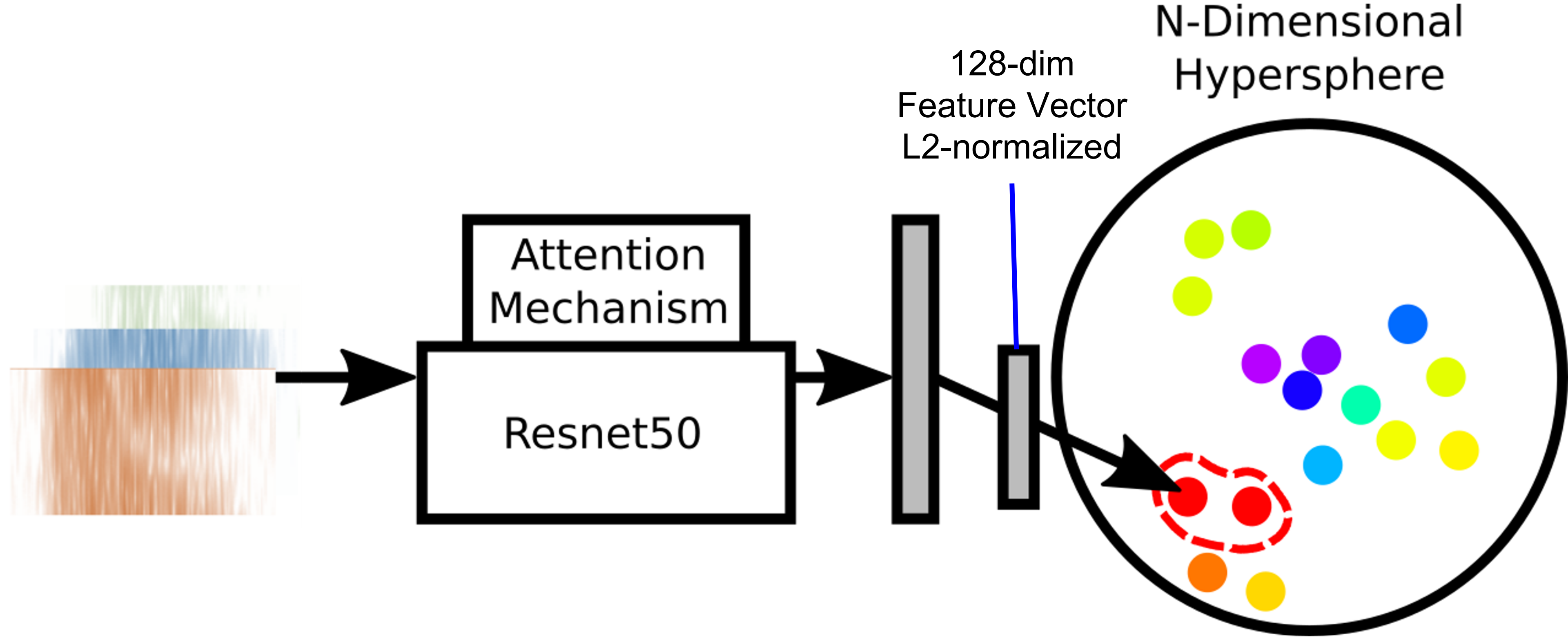

Stereo Waveform
Modified SOTA Net: ESResNet
Internal 3-Layer Spectrogram Transform
Attention Mechanism
Resnet50
N-Dimensional Hypersphere
kNN Evaluation

Train Instance Discrimination on Spectrograms



[1] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin, "Unsupervised feature learning via non-parametric instance discrimination," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018

Train Instance Discrimination on Spectrograms



Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018

Training on Sound Classification Task Yields Fast Improvement

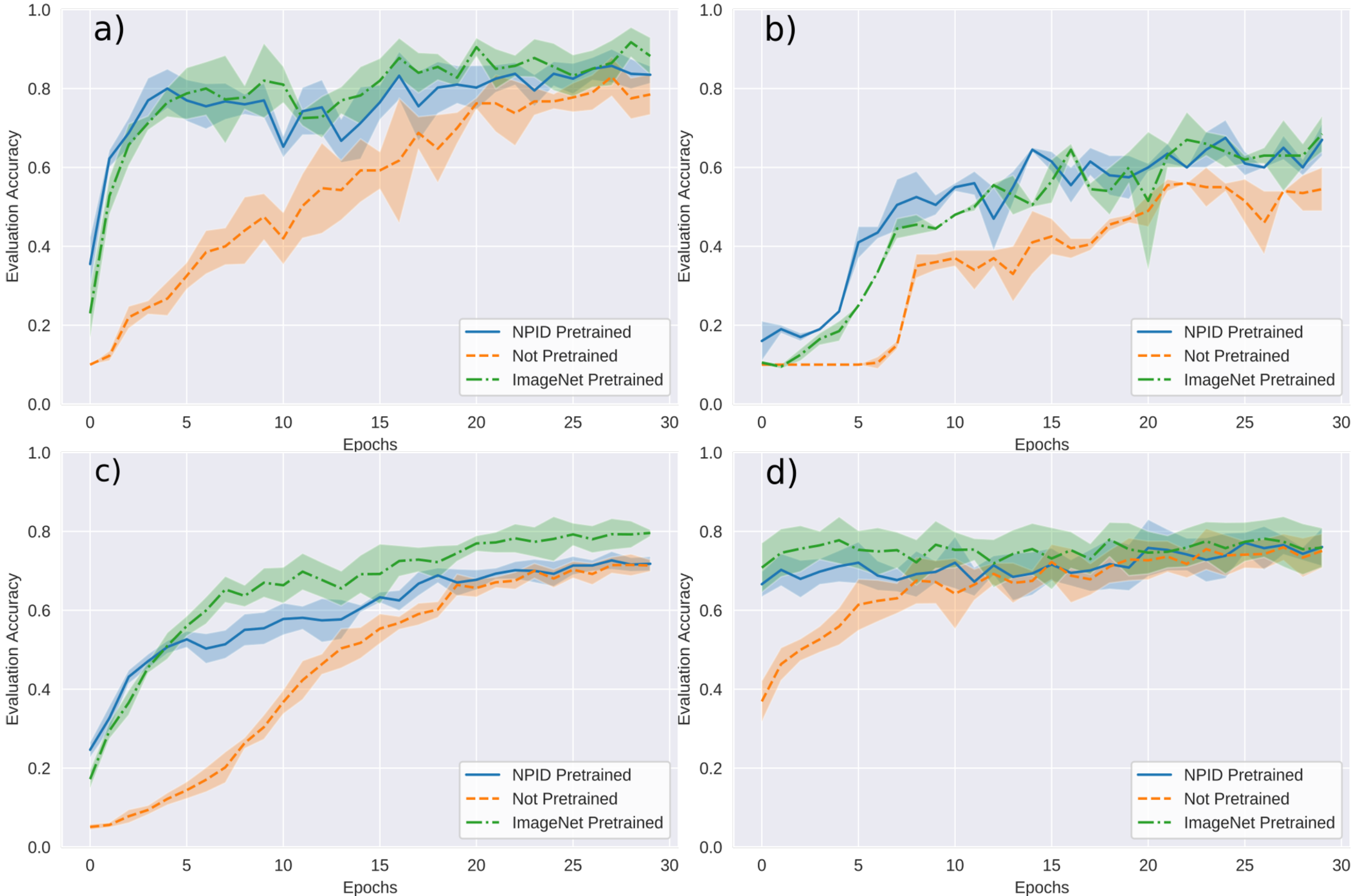
Datasets:

a) ESC10
b) DCASE2013
c) ESC50
d) US8K

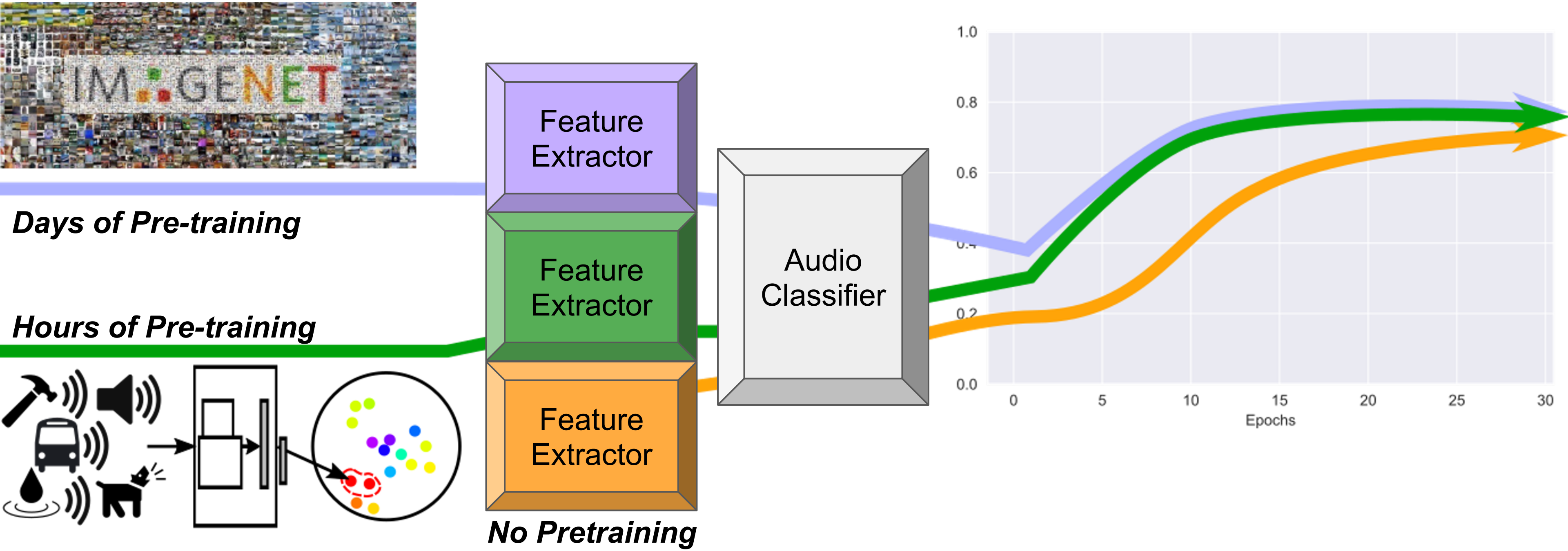
Pre-Training with NPID unsupervised on all Datasets.

Downstream training follows official train/val folds.

Results are averages over all Folds



In Summary: Quick Performance Gain for Training Arbitrary Networks on the Edge



Days of Pre-training

Hours of Pre-training

No Pretraining

Feature Extractor

Feature Extractor

Feature Extractor

Audio Classifier