ECS 289G Paper Presentation

“Ambient Sound Provides Supervision for Visual Learning”

Written by: Andrew Owens, Jiajun Wu, Josh McDermott, William Freeman, Antonio Torralba

Presented by: Baotuan Nguyen
Date: 11/3/16

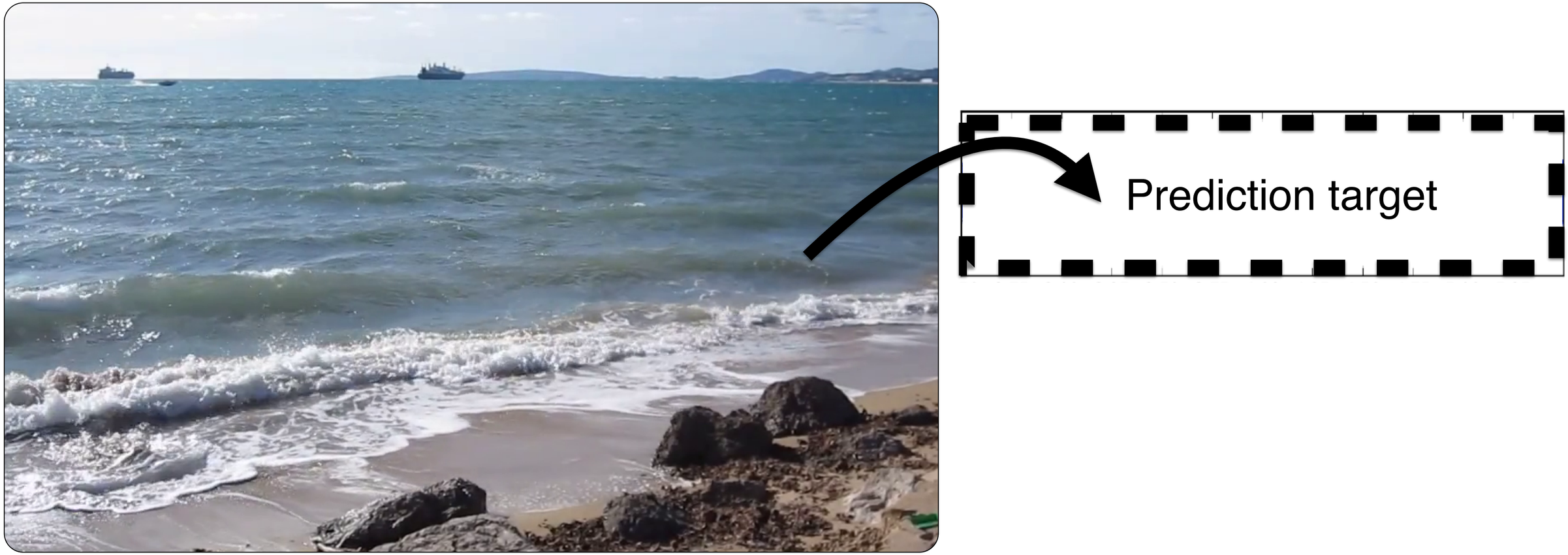
Original slides created by authors, adapted for this presentation

Goal: Self-supervised learning

Pretext Task:

Hold out one part of an image or video and predict it from the rest.

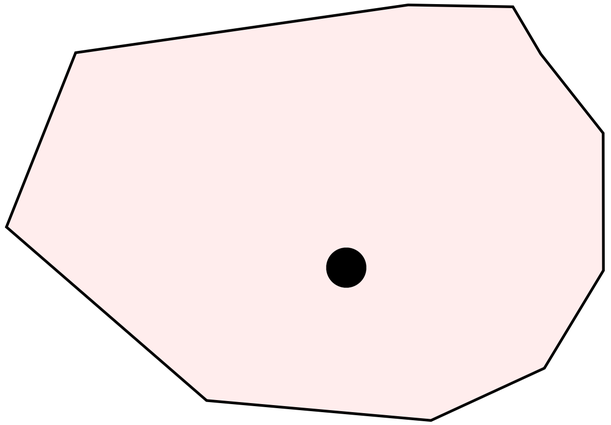
Self-supervised learning with sound



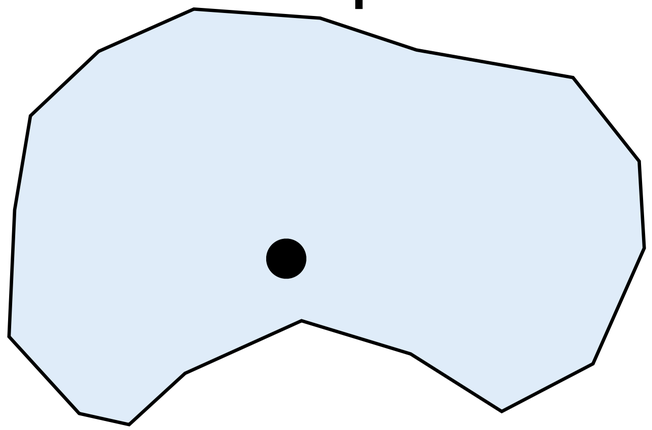
Analogously, hold out the sound from video and predict it from the images.

Audio is invariant to many visual transformations

Image space

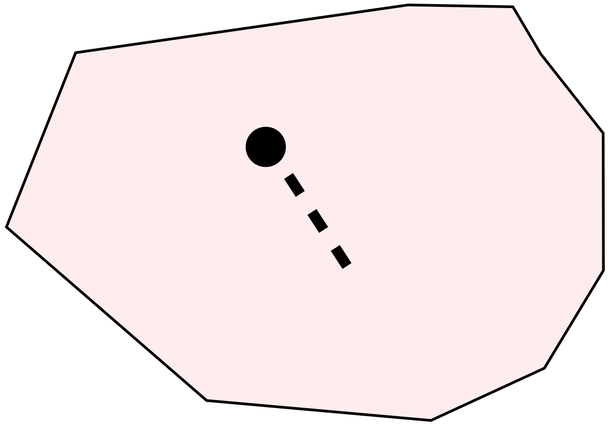


Audio space

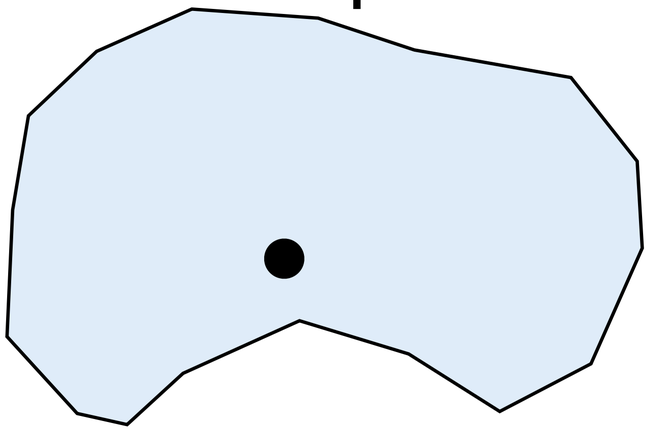


Audio is invariant to many visual transformations

Image space

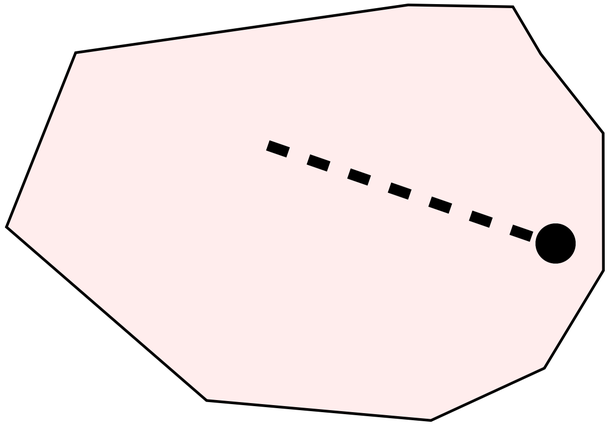


Audio space

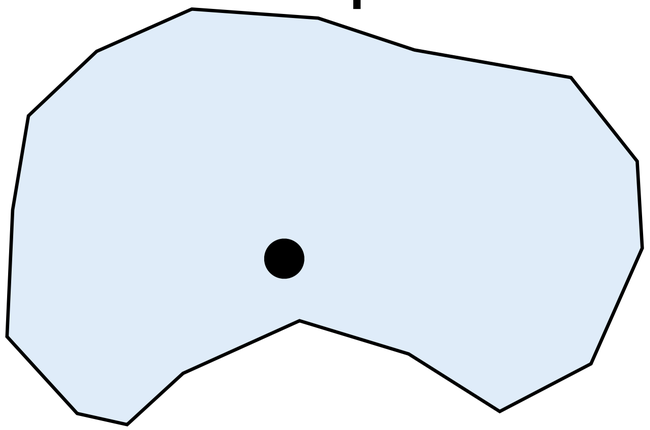


Audio is invariant to many visual transformations

Image space

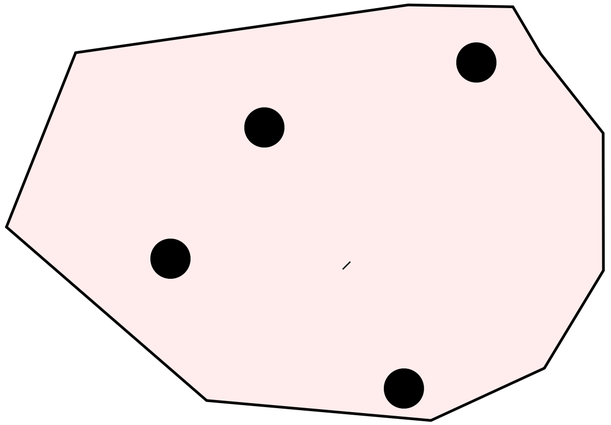


Audio space

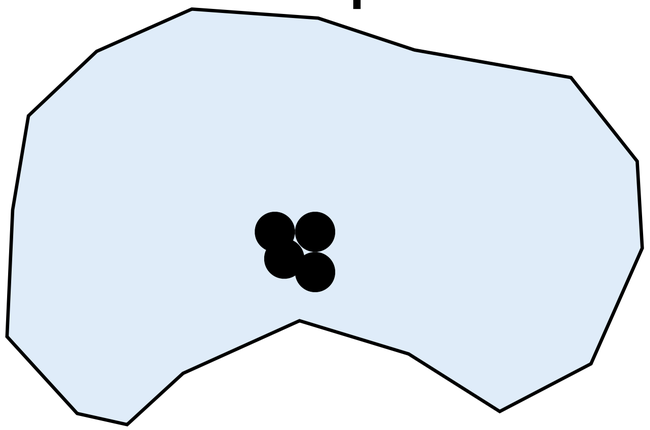


Audio is invariant to many visual transformations

Image space

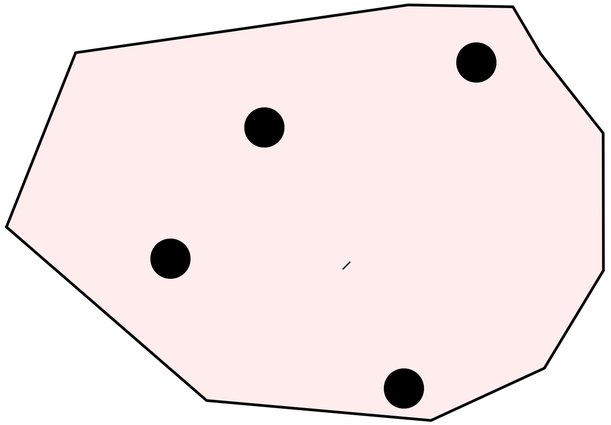


Audio space

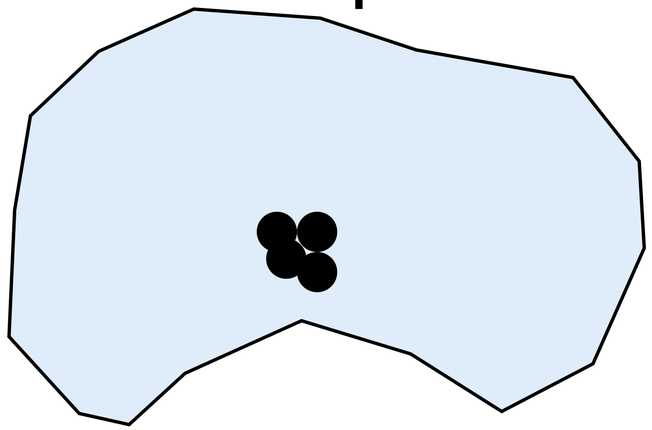


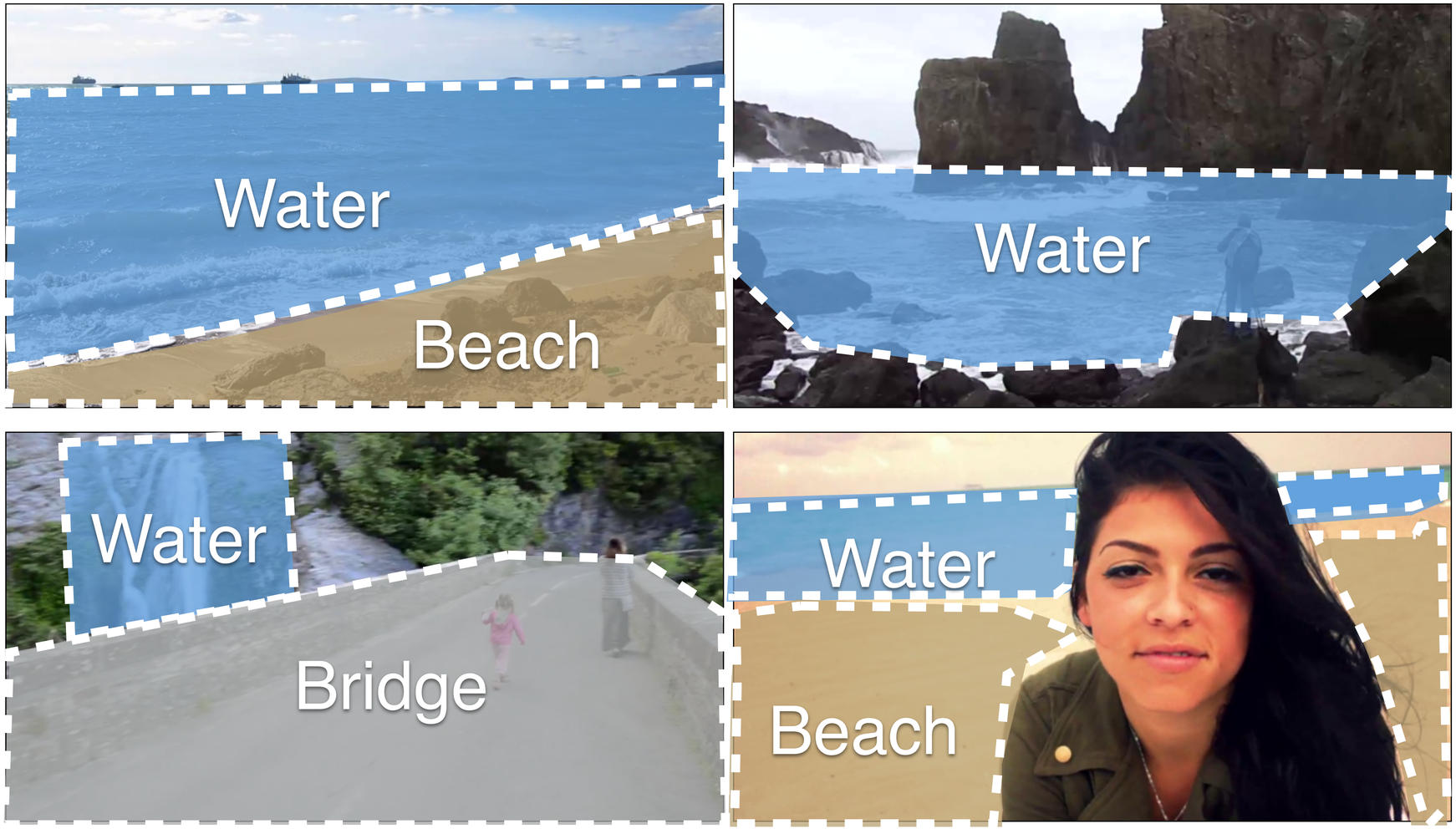
Audio is invariant to many visual transformations

Image space

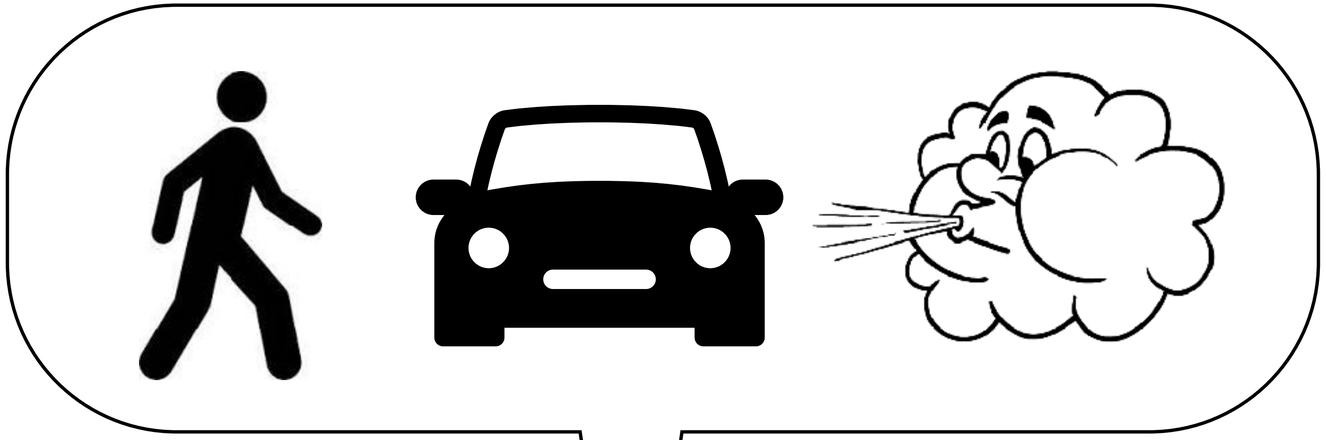


Audio space



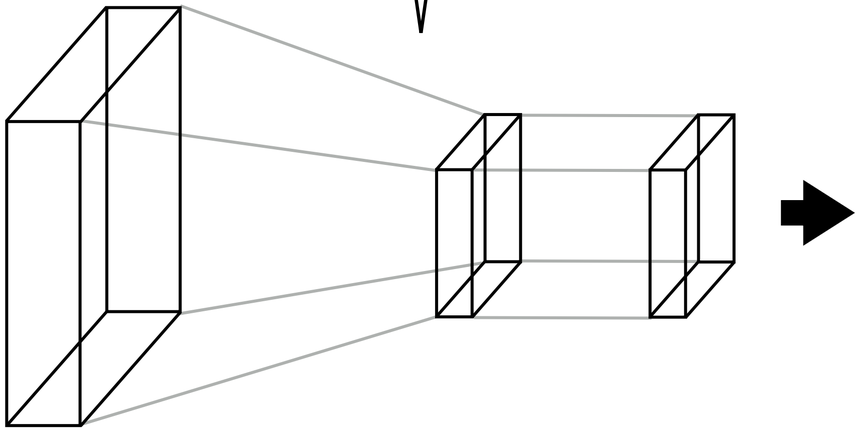


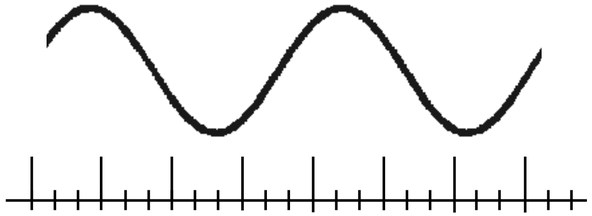
Experimental Design



Common sources

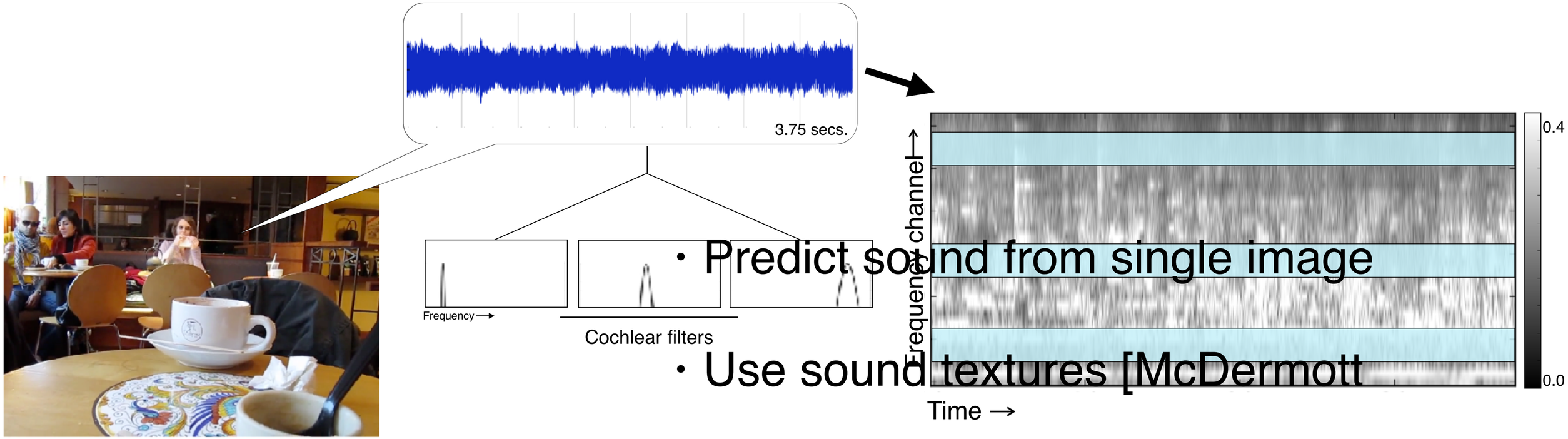
Image





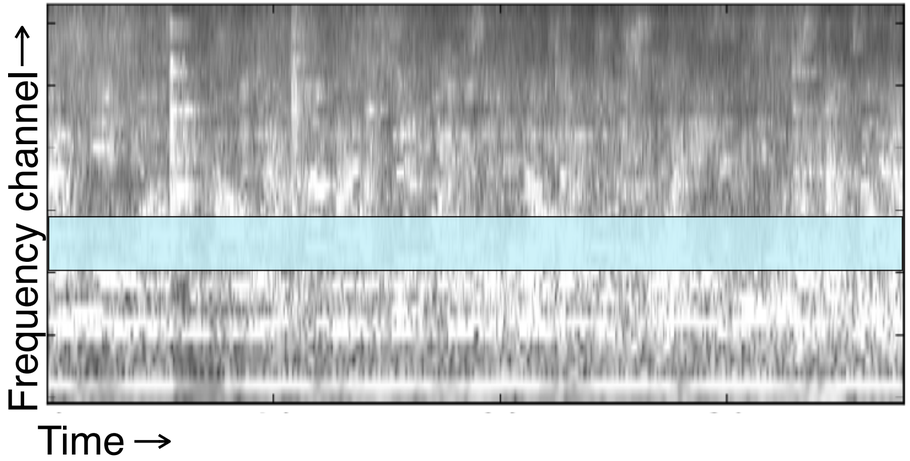
Audio

Representing ambient sound

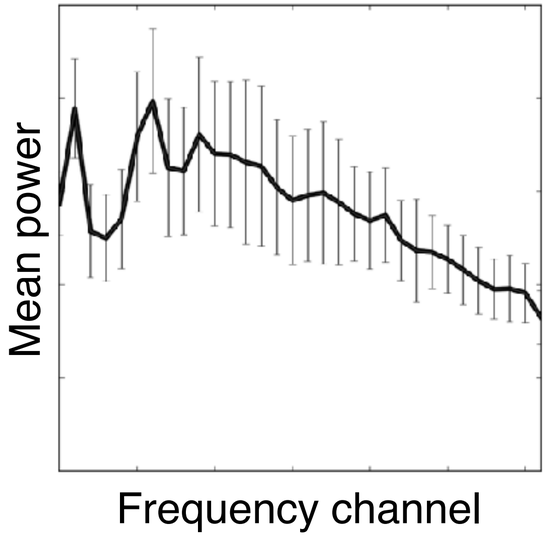


- Predict sound from single image
- Use sound textures [McDermott

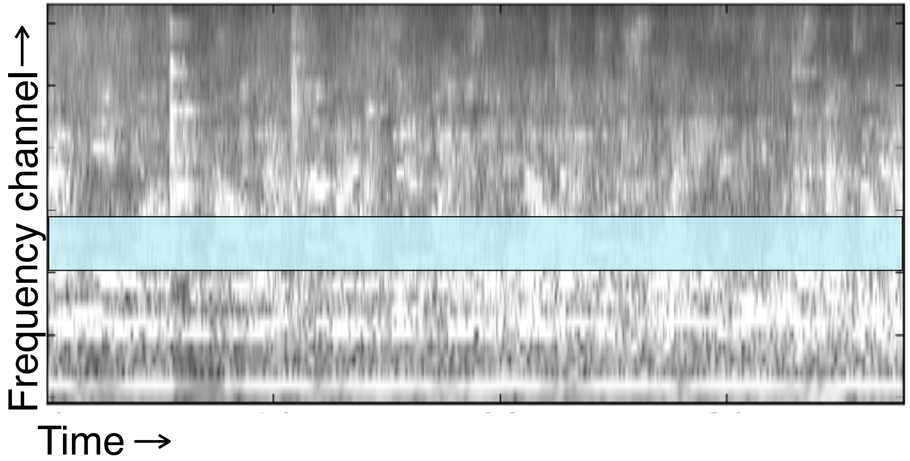
Representing ambient sound



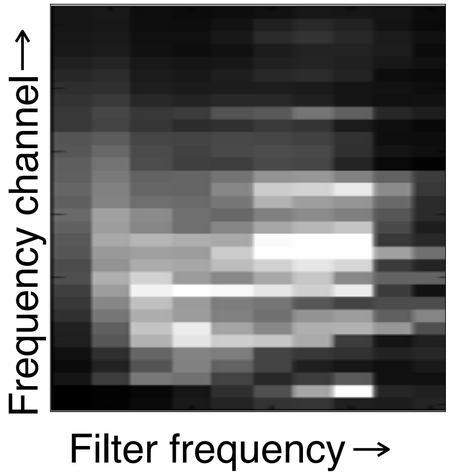
$$\frac{1}{N}\sum_i x_i$$



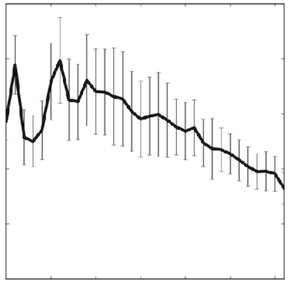
Representing ambient sound



$$\frac{1}{N}\|x * f_j\|^2$$

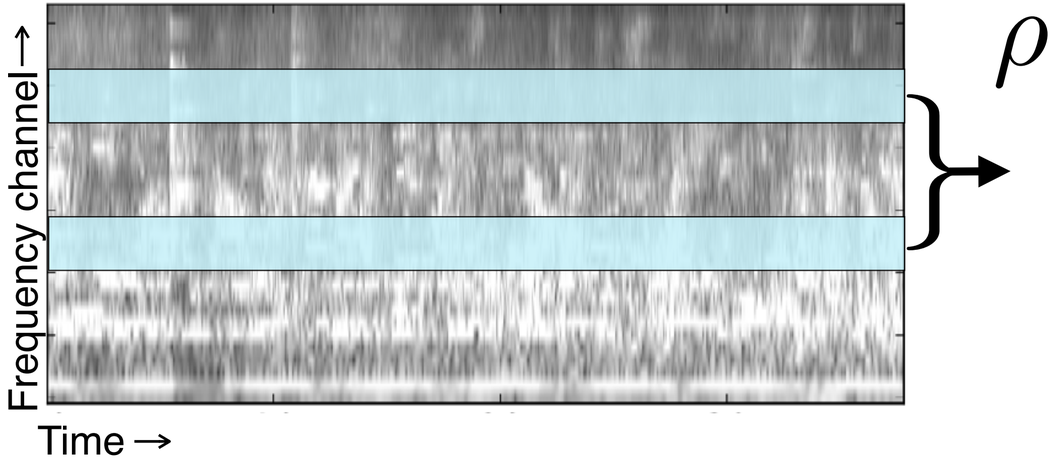


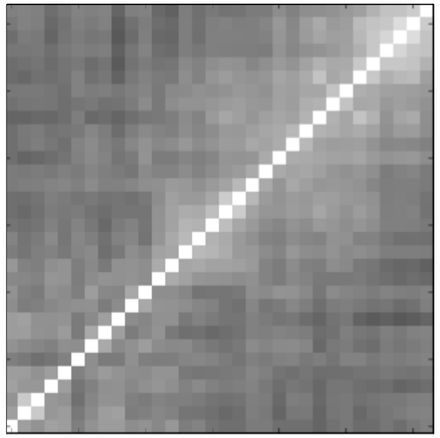
Sound texture
[McDermott & Simoncelli



Moments

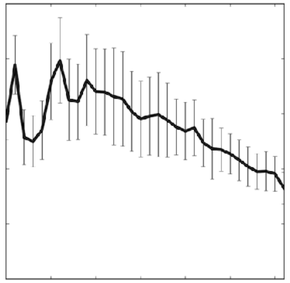
Representing ambient sound



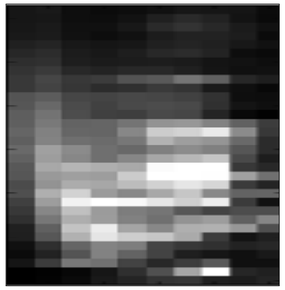


Correlation between frequency channels

Sound texture
[McDermott & Simoncelli

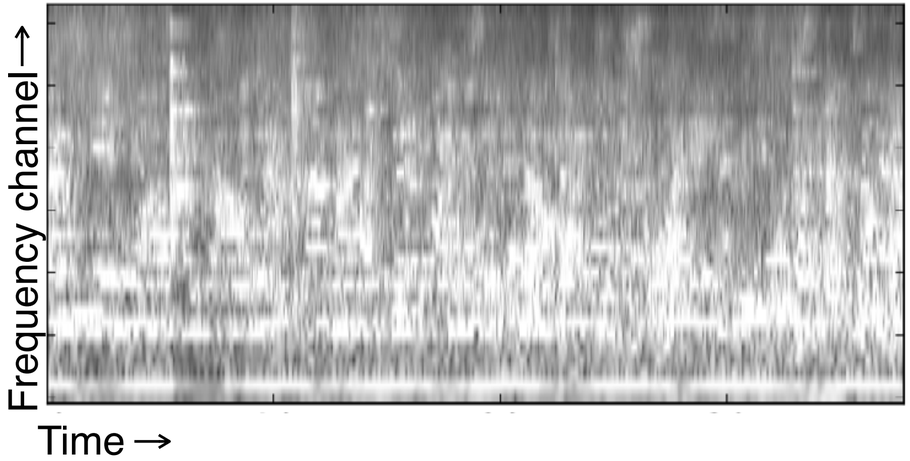


Moments

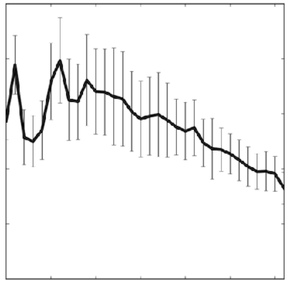


Filter responses

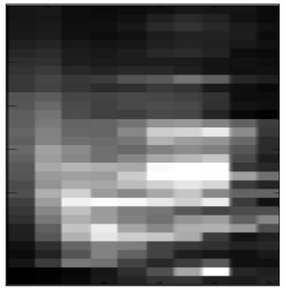
Representing ambient sound



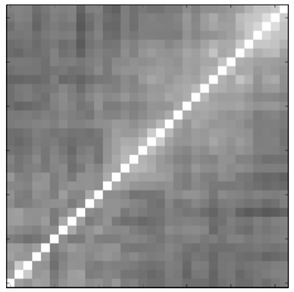
Sound texture
[McDermott & Simoncelli



Moments



Filter responses

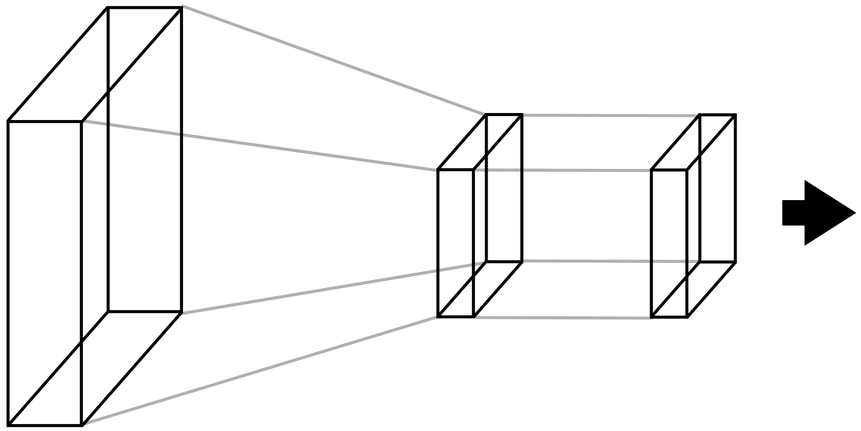


Correlations

Training Data

- Flickr video dataset.
- 180K videos, 10 random frames from each.
- Trained from scratch

Video frame

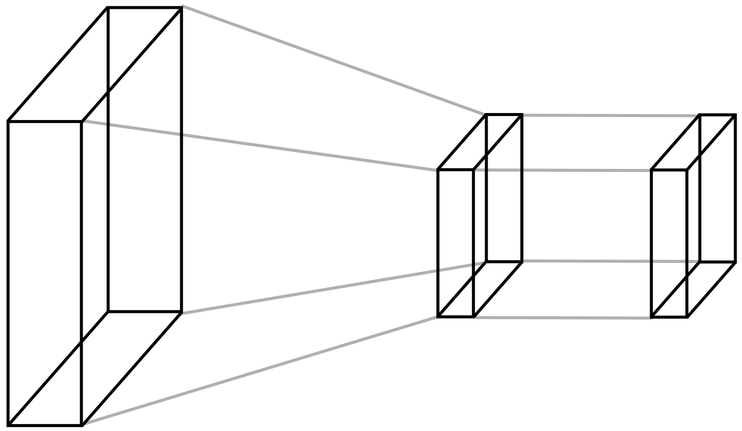


ConvNet

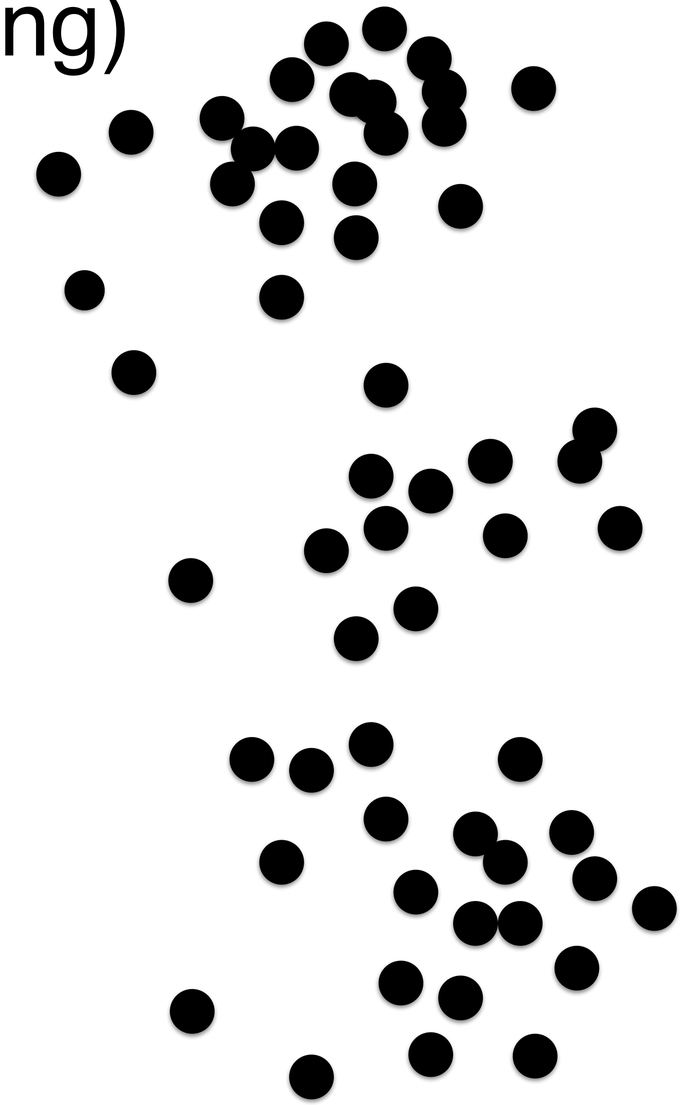
Sound texture
[McDermott & Simoncelli

Predicting sound (clustering)

Video frame



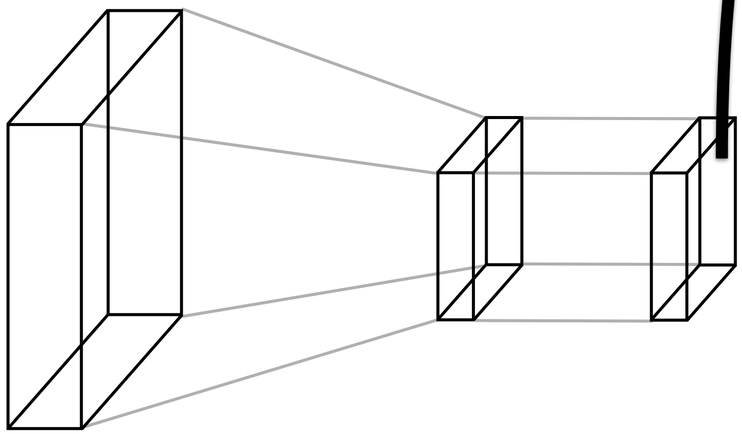
ConvNet



Sound feature vectors

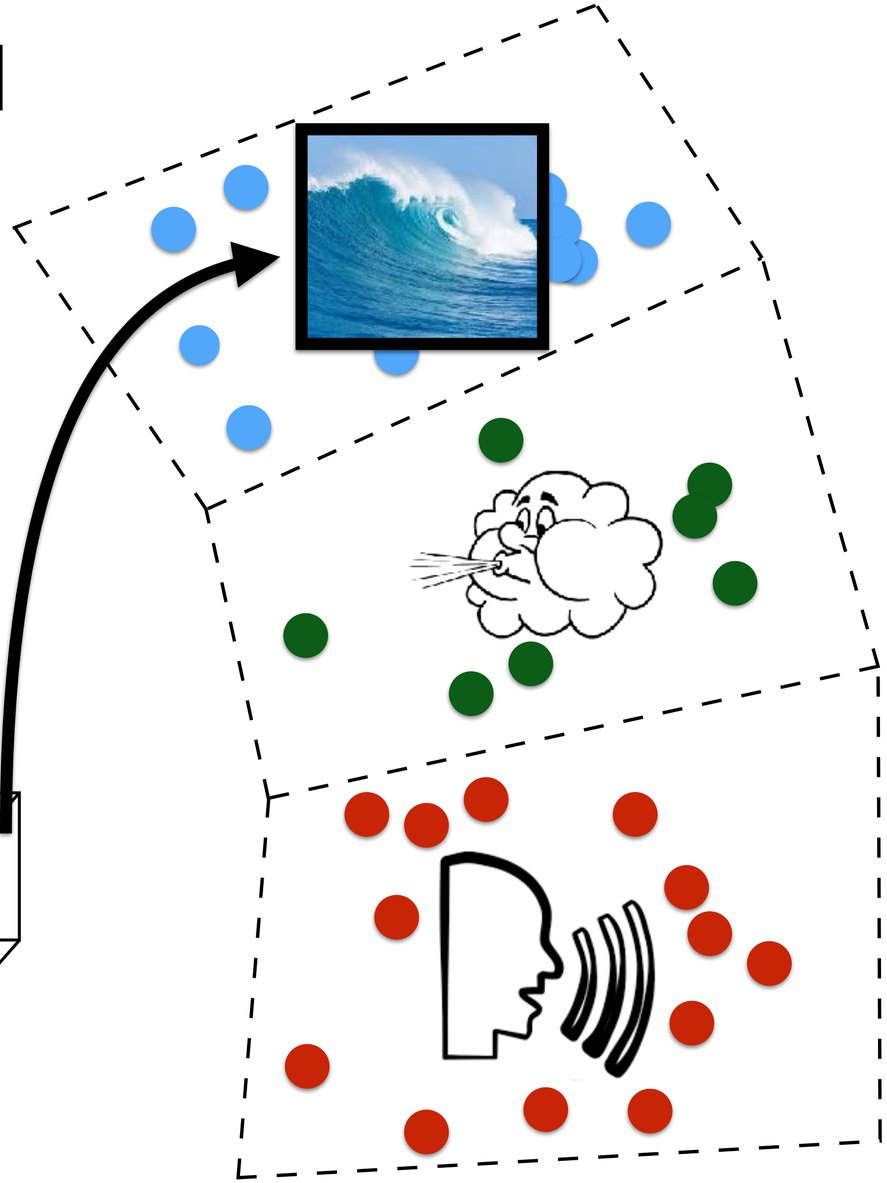
Predicting sound

Video frame



ConvNet

Audio cluster

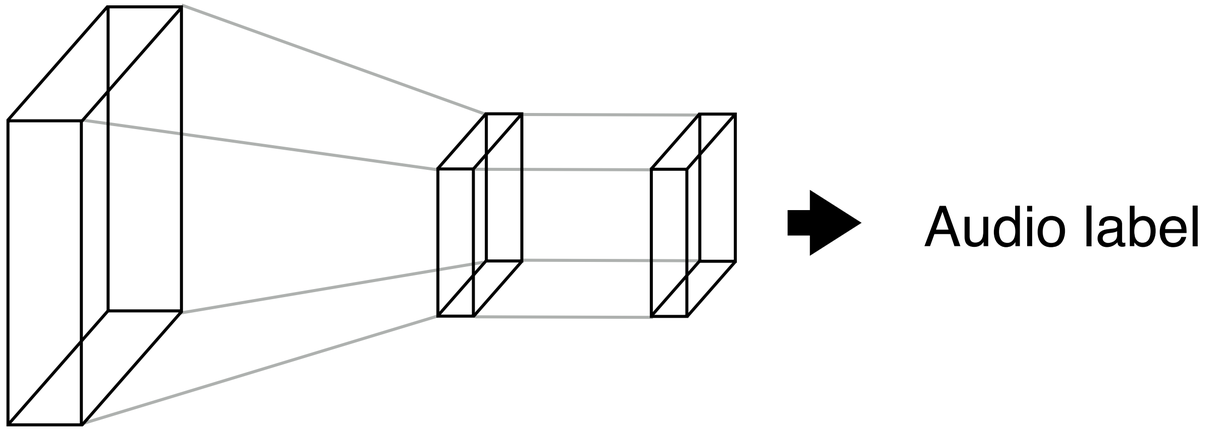


K-means or LSH

Top audio clips for one cluster (of 30).

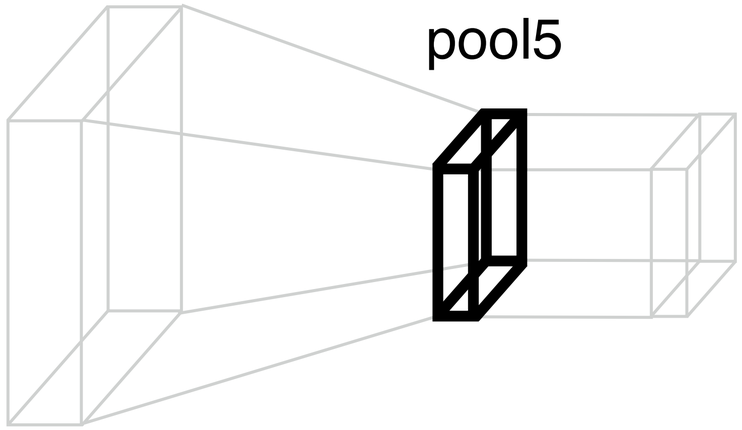

Top audio clips for one cluster (of 30).

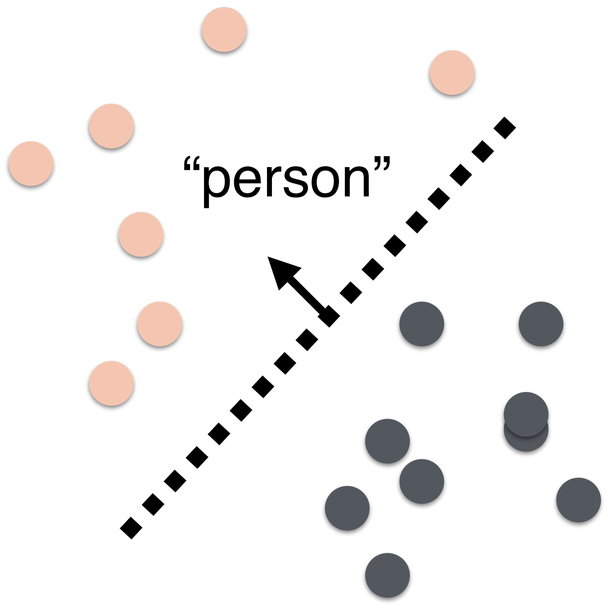
What did the network learn?



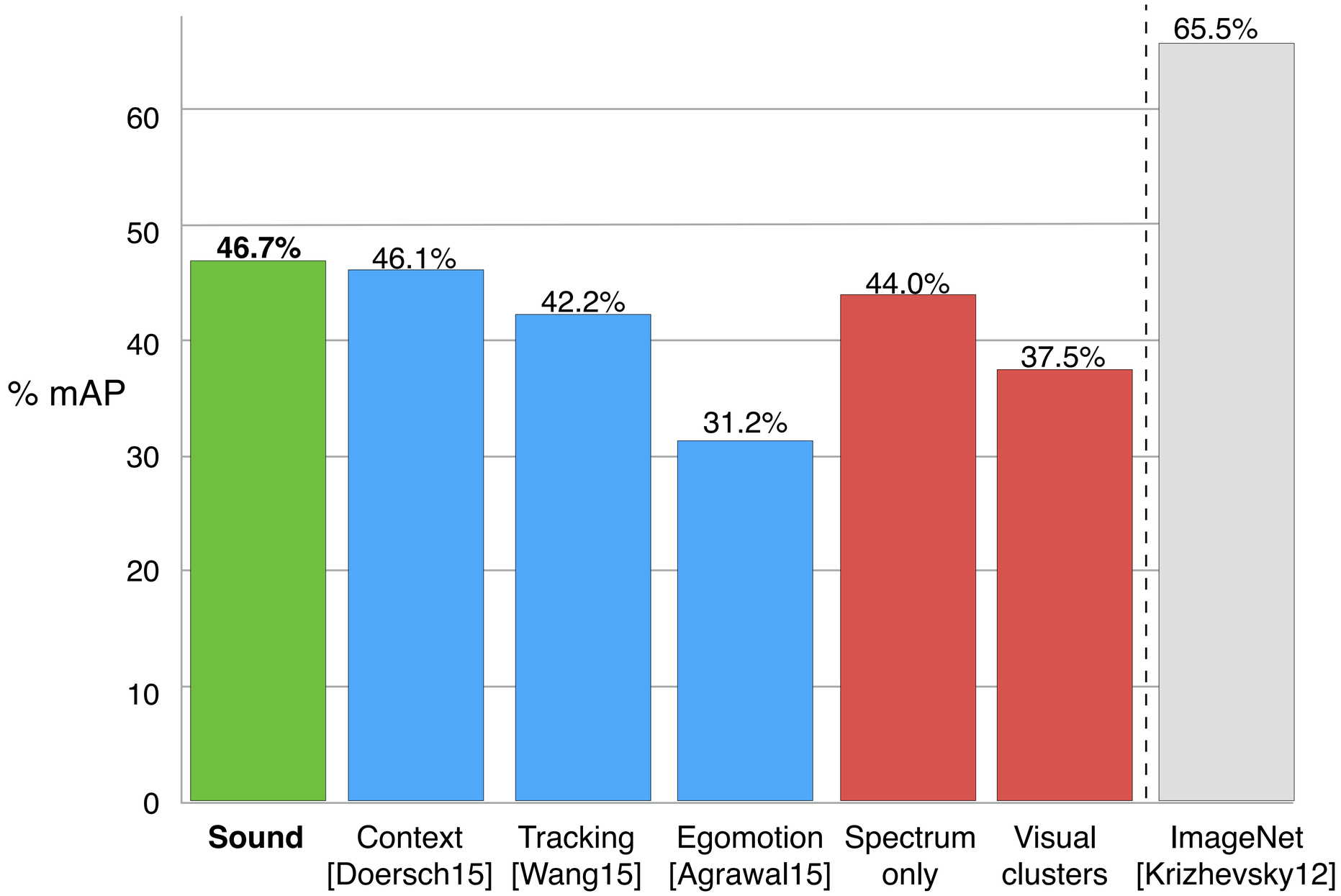
What did the network learn?

PASCAL VOC 2007

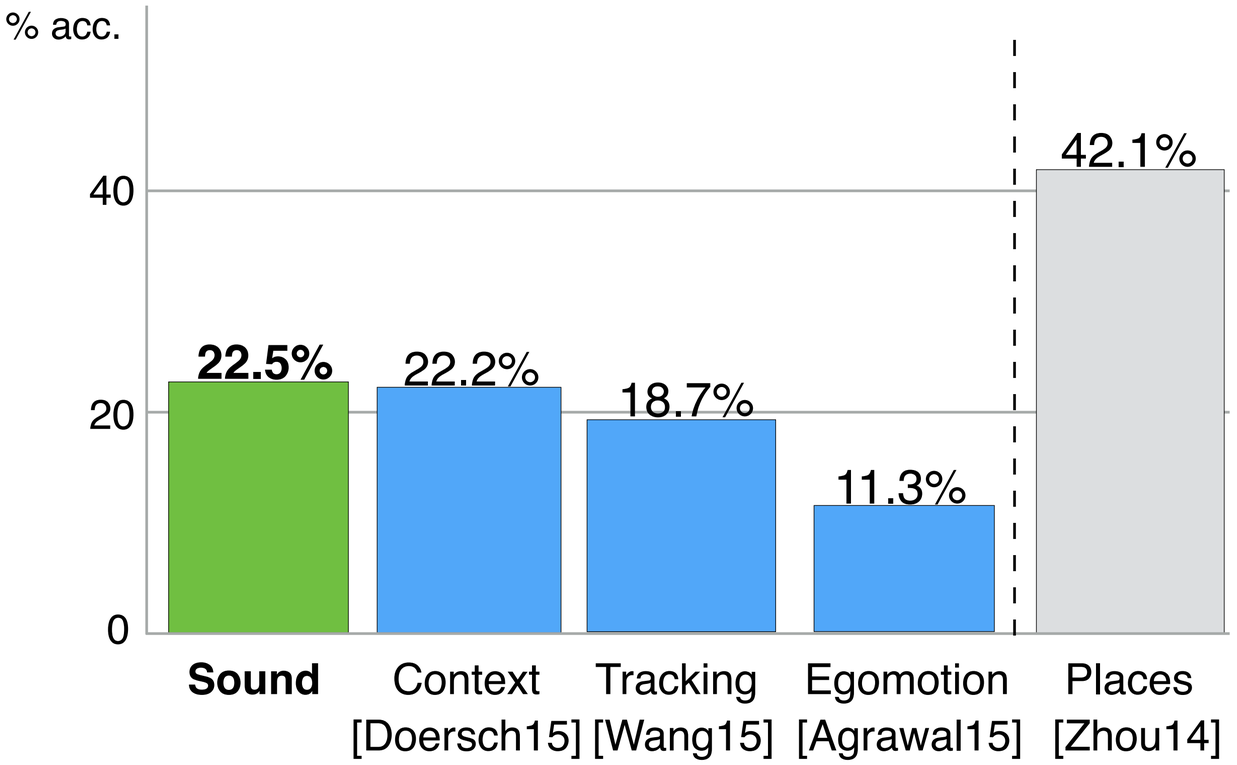




PASCAL VOC Classification



SUN397 Scene Recognition



Classification with linear SVM

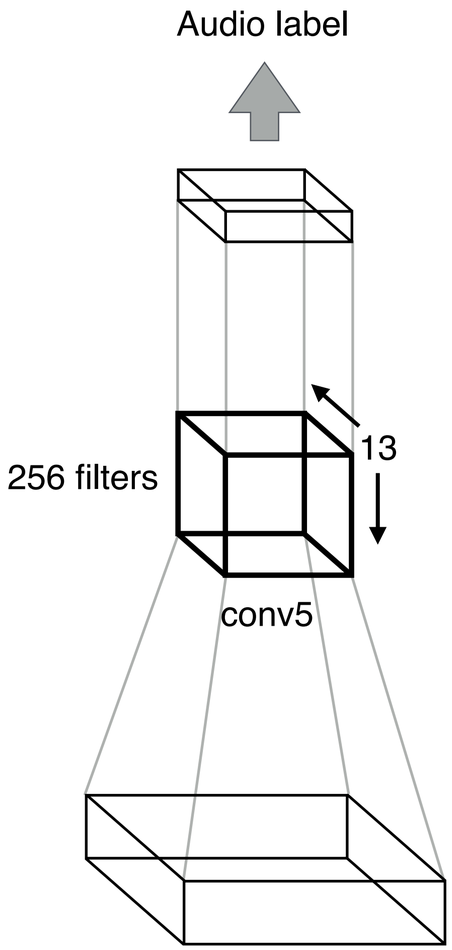
Method	VOC Cls. (%mAP)				SUN397 (%acc.)			
	max5	pool5	fc6	fc7	max5	pool5	fc6	fc7
Sound (cluster)	36.7	45.8	44.8	44.3	**17.3**	**22.9**	20.7	14.9
Sound (binary)	**39.4**	**46.7**	**47.1**	**47.4**	17.1	22.5	**21.3**	**21.4**
Sound (spectrum only)	35.8	44.0	44.4	44.4	14.6	19.5	18.6	17.7
Texton-CNN	28.9	37.5	35.3	32.5	10.7	15.2	11.4	7.6
K-means (Krähenbühl et al.)	27.5	34.8	33.9	32.1	11.6	14.9	12.8	12.4
Tracking (Wang and Gupta)	33.5	42.2	42.4	40.2	14.1	18.7	16.2	15.1
Patch pos. (Doersch et al.)	26.8	46.1	-	-	9.8	22.2	-	-
Egomotion (Agrawal et al.)	22.7	31.1	-	-	9.1	11.3	-	-
ImageNet (Krizhevsky et al.)	**63.6**	**65.6**	**69.6**	**73.6**	29.8	34.0	37.8	37.8
Places (Zhou et al.)	59.0	63.2	65.3	66.2	**39.4**	**42.1**	**46.1**	**48.8**

Fine-tuning Fast R-CNN

Method	(%mAP)
Random init. (Krähenbühl et al.)	41.3
Sound (cluster)	44.1
Sound (binary)	43.3
Tracking (Wang and Gupta)	44.0
Egomotion (Agrawal et al.)	41.8
Patch pos. (Doersch et al.)	46.6
Calib. + Patch (Krähenbühl et al.)	**51.1**
ImageNet (Krizhevsky et al.)	**57.1**
Places (Zhou et al.)	52.8

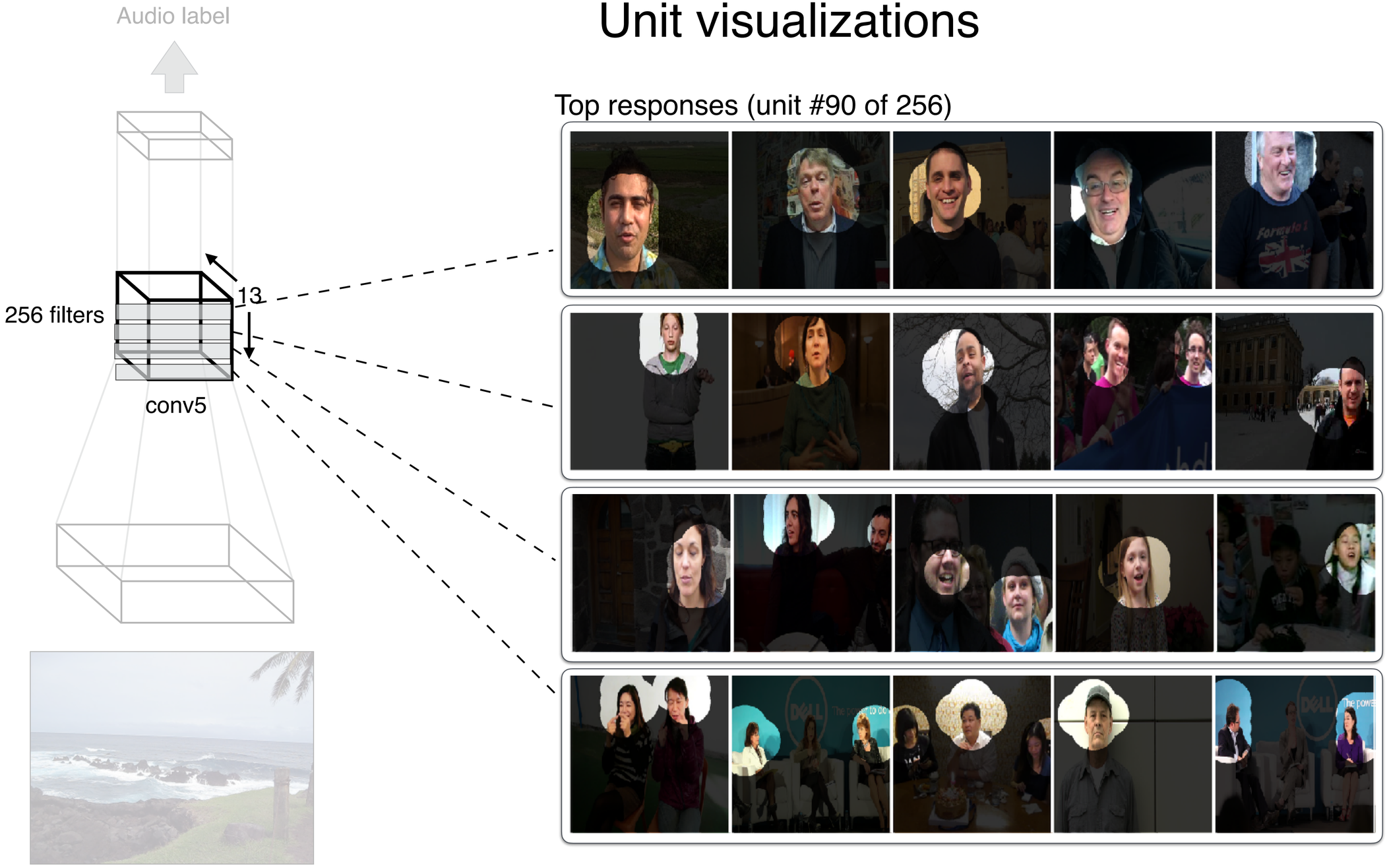
For 30-way classification ask: 15.8% classification accuracy (vs. 6.6% for choosing the most common label)

Did the model recognize sound sources?



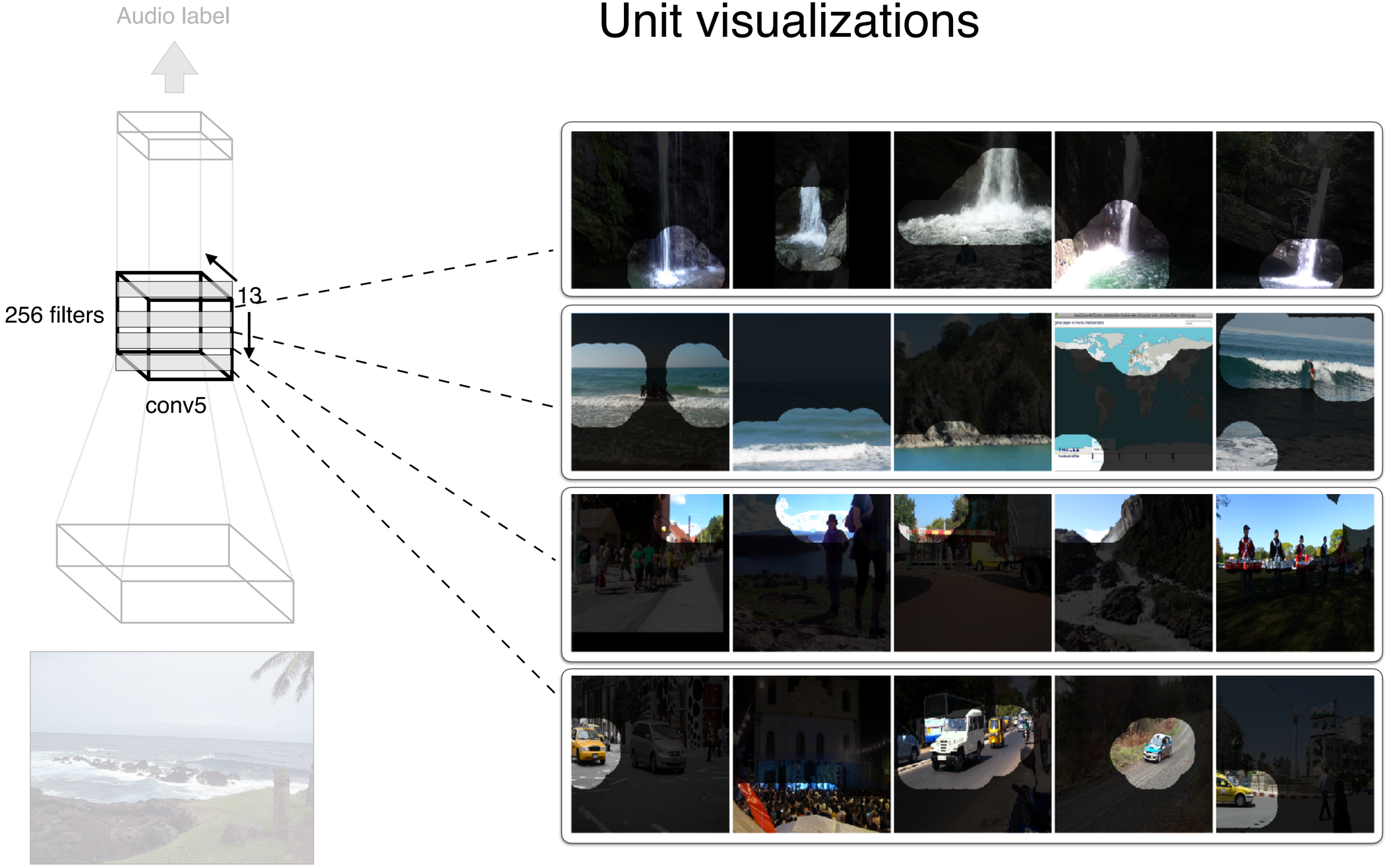
Object-selective units emerge when recognizing scenes [Zhou et al. 2015]

Unit visualizations



Unit visualizations

Unit visualizations



Strengths

- Model learns a representation that conveys semantics.
- Orthogonal to image-based approaches.
- Invariance to visual transformation.

Weaknesses

- Requires characteristics sounds, but not all objects emit sounds.
- Sound can be misleading.

Other Possible Extensions?

-Leveraging data from social media.
(i.e. using comments on photos/videos as a supervisory signal)