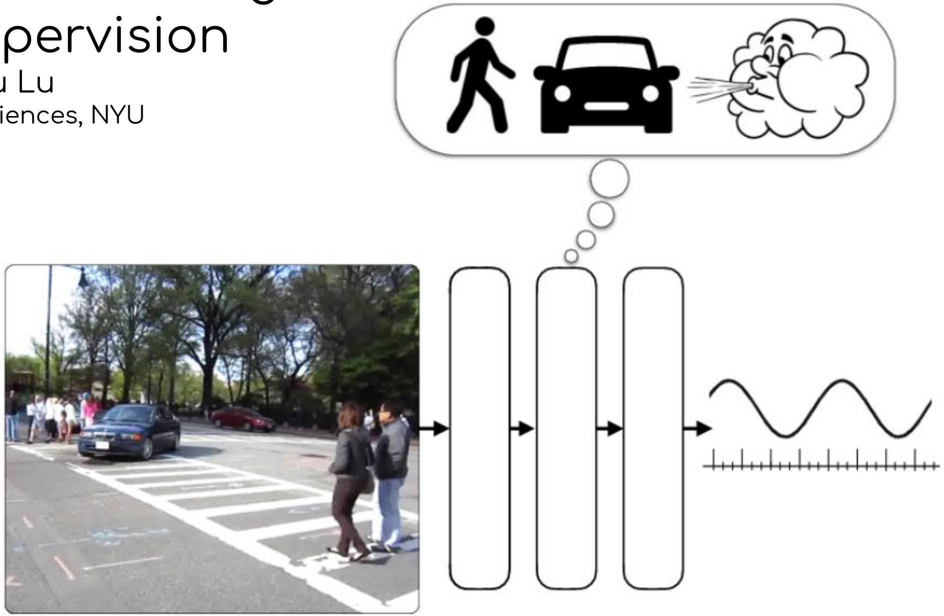
Improving Visual Recognition using Ambient Sound for Supervision

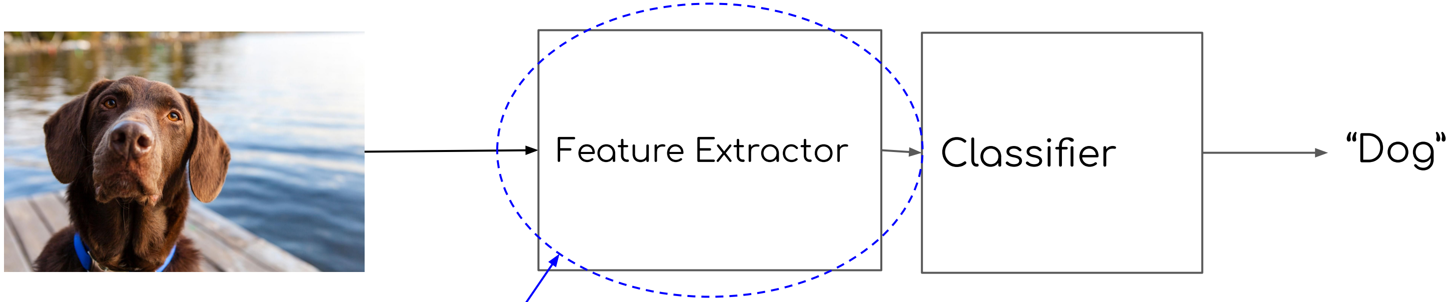
Rohan Mahadev, Hongyu Lu
Courant Institute of Mathematical Sciences, NYU

Sound conveys important information about objects in our surrounding. This fact can be exploited by using sound as a supervisory signal to train a model which improves image recognition performance.



Credit: A. Owens

Motivation



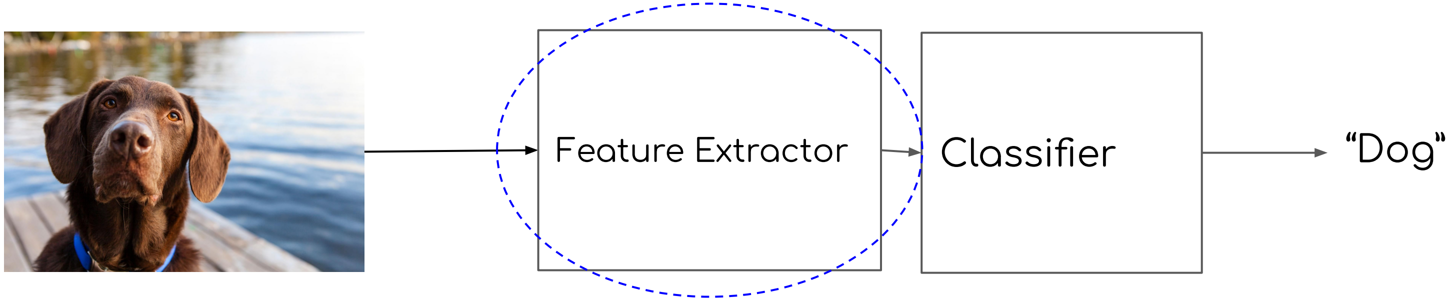
Repurposed for:

Classification
Detection
Semantic/Instance segmentation
Visual Question Answering

Recipe:

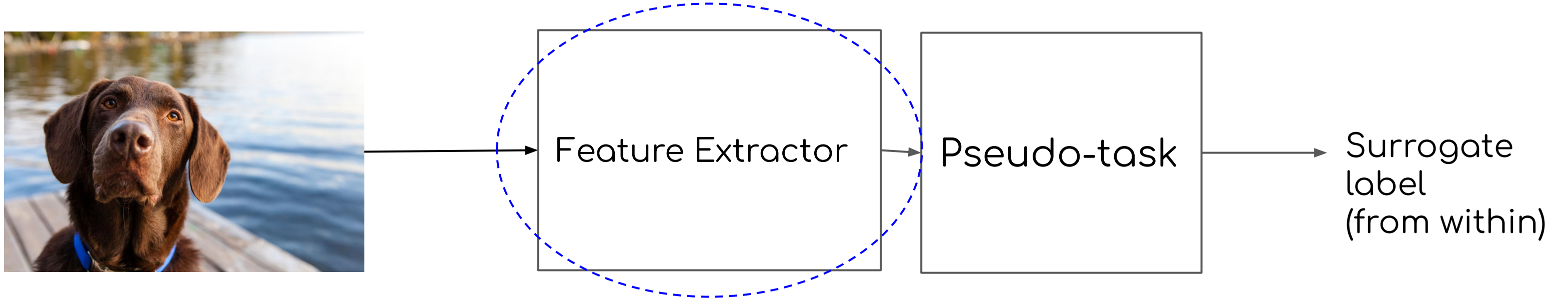
1. Pre-train on large supervised dataset
2. Collect a dataset of supervised images
3. Train a ConvNet

Motivation



<u>All predicated on human annotation</u>
Costly
Time consuming
Prone to error
Bias?
What about complex concepts? Medicine/Legal?

Motivation

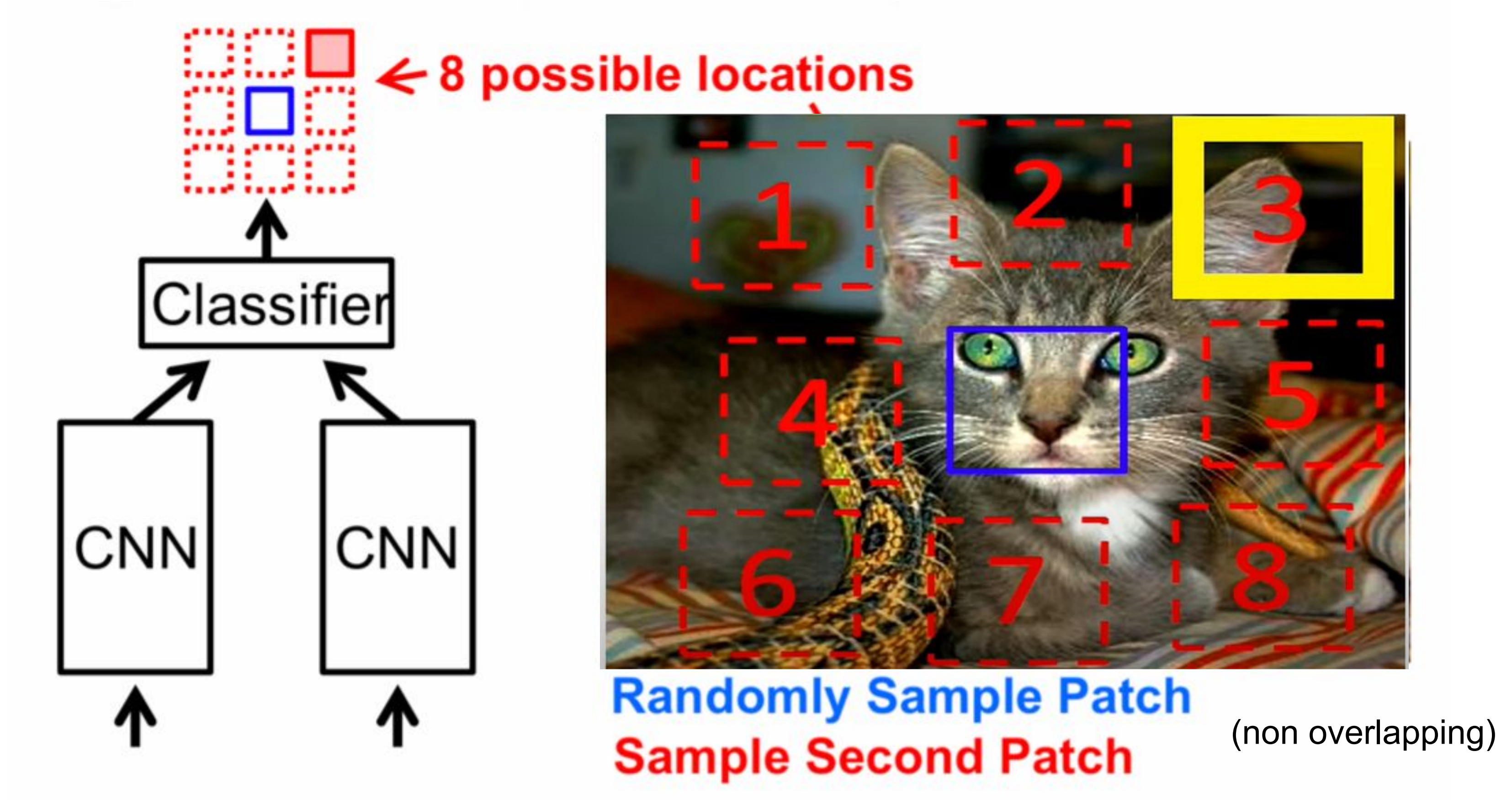


Instead of density modeling (unsupervised learning) where we want ρ_{model} similar to ρ_{data}

We find supervision signal y within the input data, which allows use of standard supervised learning losses and architectures

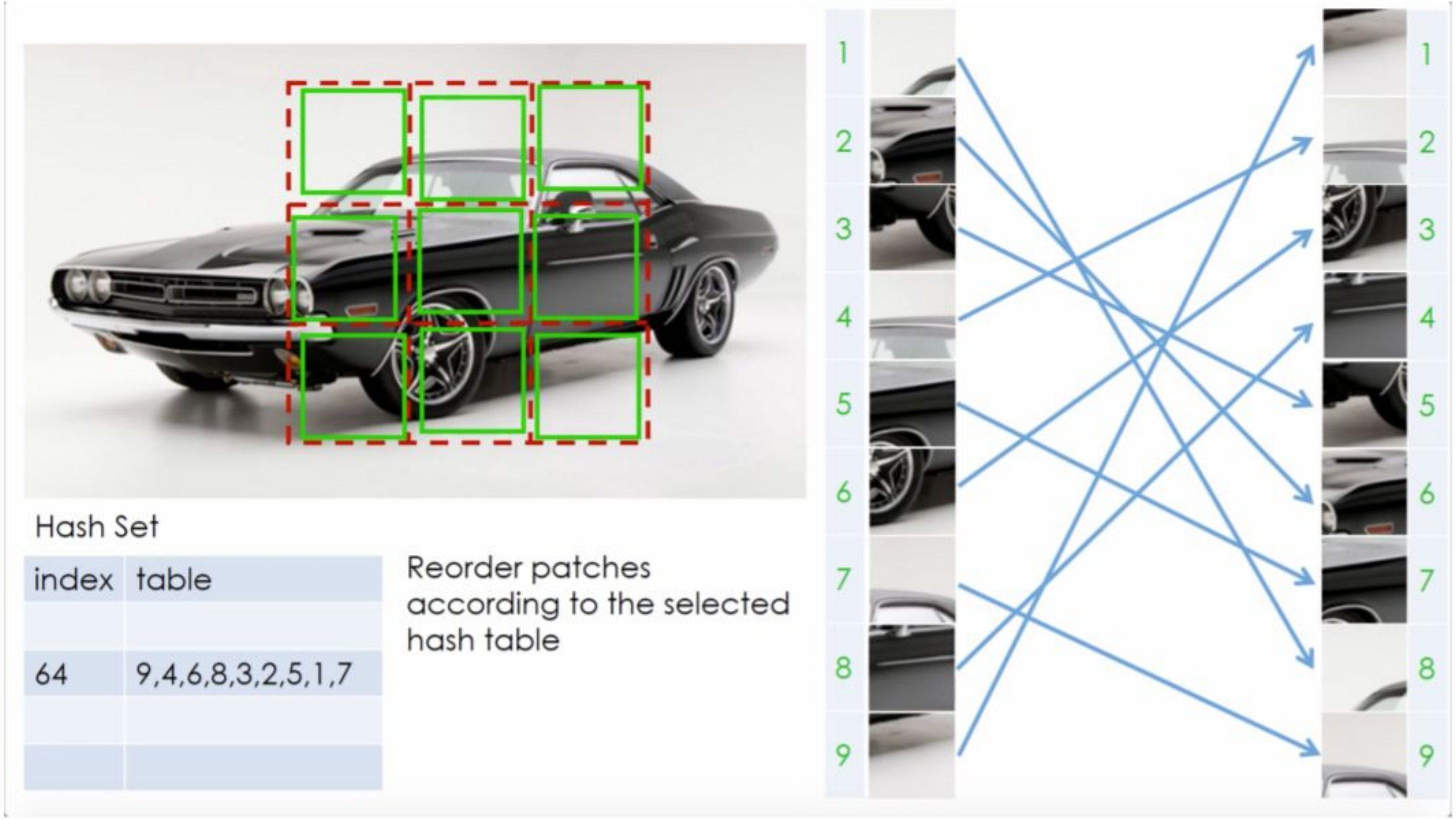
$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y(x_i))$$

Self supervision in CV



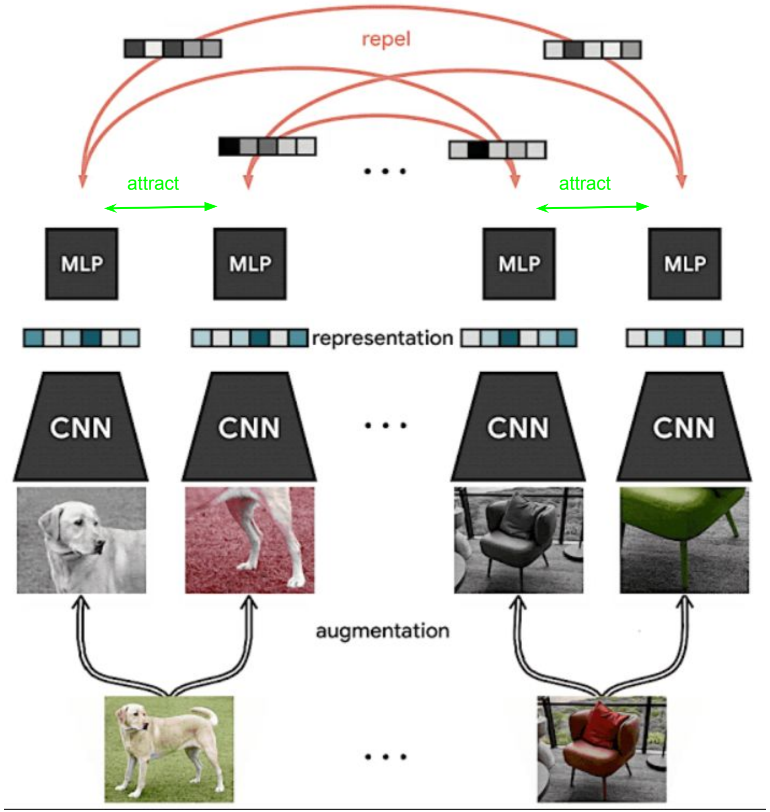
Unsupervised visual representation learning by context prediction, Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Self supervision in CV



Mehdi Noroozi (University of Bern), Hamed Pirsiavash (University of Maryland Baltimore County), Paolo Favaro

Self supervision in CV



$$\ell_{i,j}^{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, Geoffrey Hinton

Self supervision in CV

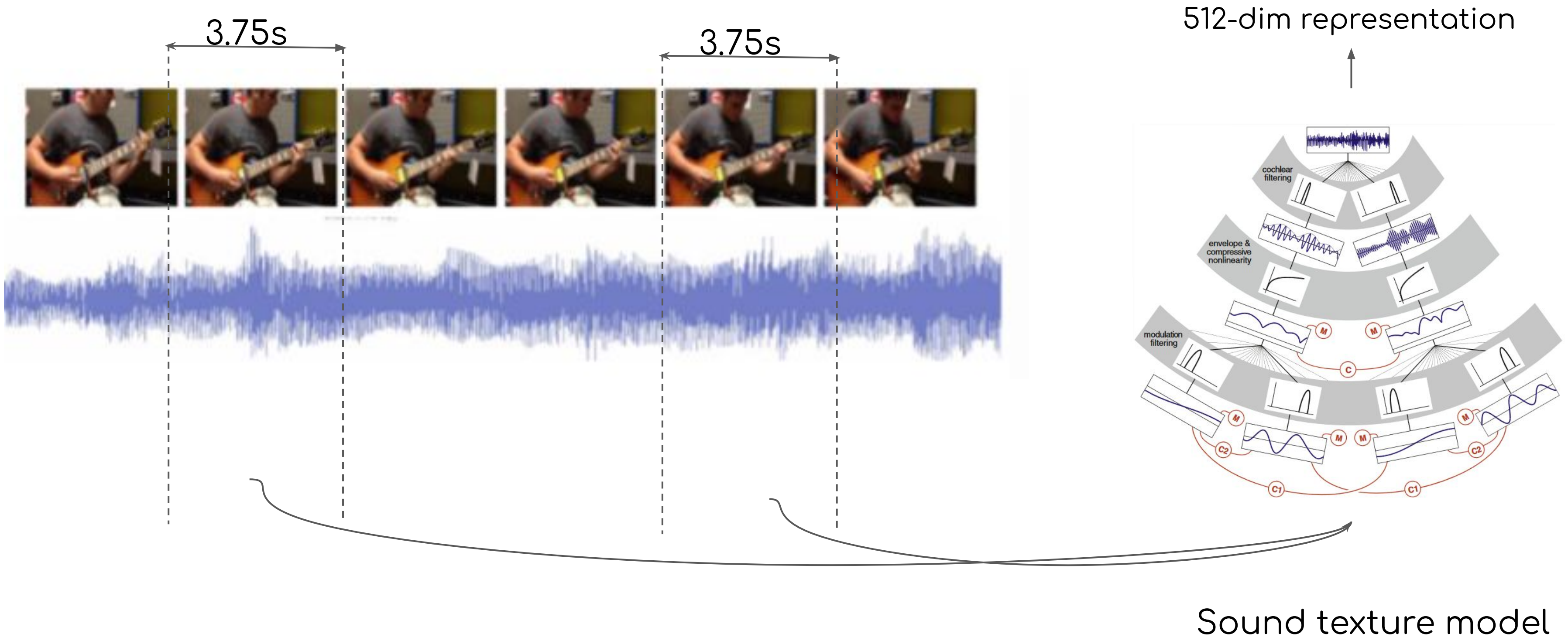
Colorization
Image inpainting
Split-brain (cross channel prediction)
Counting visual primitives
Video shuffling
Deep Clustering
....,
....,
Image GPT,
SimCLR (v2)
MoCo

All use images/video frames. What about audio?

The thud of a bouncing ball, the onset of speech as lips open — when visual and audio events occur together, it suggests that there might be a common, underlying event that produced both signals.

We want (x,y) pairs of (image, sound representation)

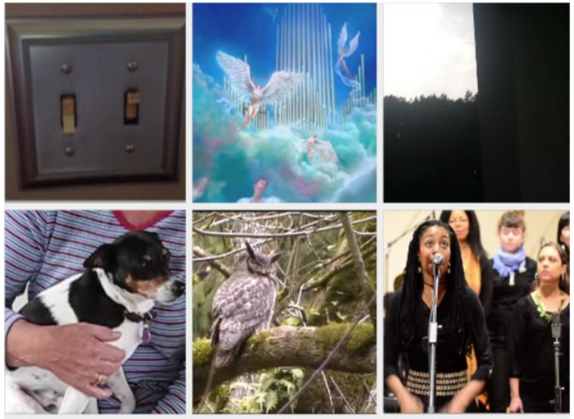
Relationship b/w sound and images

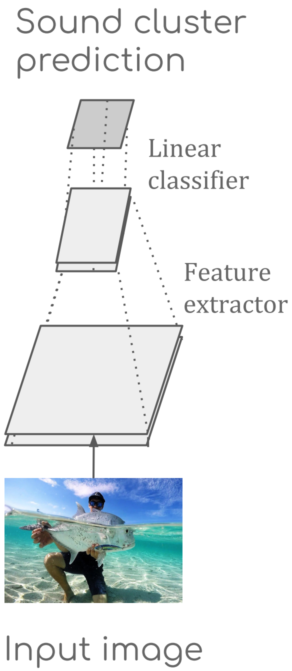


Credit: J. McDermott, E. Simoncelli

Relationship b/w sound and images

Cluster sound using K-means and
use cluster as label for image





We see that frames fall into categories such as “outdoor scenes”, “indoor scenes”, “people laughing”, and “music”.

And we get pairs of {(img_1, cluster=1)…(img_n, cluster=k)}

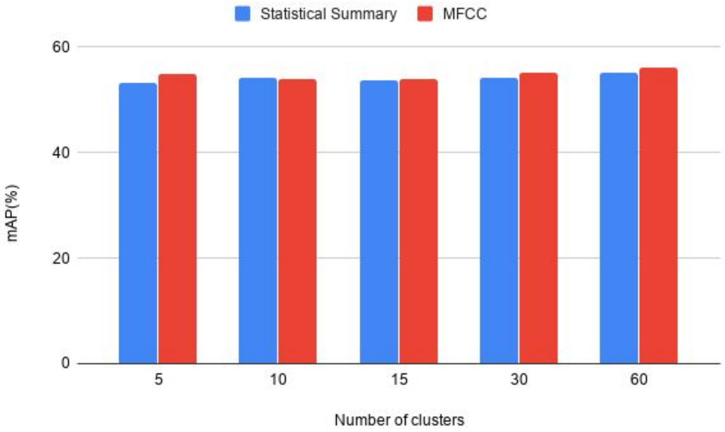
Dataset

Need for ambient sound

AudioSet consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos.

Do not use the annotations in any way. Only needed the dataset as it was a good collection of videos with only ambient sounds

	Head	All
Imagenet(pretrained) [15]	78.9	79.9
Random [14]	29	33.2
Pathak *et al.* [25]	34.6	56.5
Donahue *et al.* [6]	52.3	60.1
Owens *et al.* [22]	52.3	61.3
Pathak *et al.* [24]	-	61
Wang *et al.* [29]	55.6	63.1
Doersch *et al.* [5]	55.1	65.3
Bojanowski *et al.* [2]	56.7	65.3
Zhang *et al.* [34]	61.5	65.9
Zhang *et al.* [35]	**63**	67.1
Noroozi and Favaro [20]	-	67.6
Noroozi *et al.* [21]	-	**67.7**
Our model	52.8	55.1



Pascal VOC classification

	aer	bk	brd	bt	btl	bus	car	cat	chr	cow	din	dog	hrs	mbk	prs	pot	shp	sfa	trn	tv
Imagenet(pretrained)[15]	79	71	73	75	25	60	80	75	51	45	60	70	80	72	91	42	62	56	82	62
Owens *et al.* [22]	68	47	38	54	15	45	66	45	42	23	37	28	71	58	**85**	**25**	26	32	67	42
Colorization [34]	70	50	45	58	15	45	71	50	39	20	38	41	72	57	81	17	**42**	41	66	38
Tracking [29]	67	35	41	54	11	35	62	35	39	21	30	26	70	53	78	22	32	37	61	34
Object motion [24]	65	39	39	50	13	33	61	36	39	24	35	28	69	49	82	14	19	34	56	31
Patch position [5]	70	44	43	60	12	44	66	**52**	44	24	45	31	73	48	78	14	28	39	62	43
Egomotion [1]	60	24	21	35	10	19	57	24	27	11	22	18	61	40	69	13	12	24	48	28
Texton-CNN [17]	65	35	28	46	11	31	63	30	41	17	28	23	64	51	74	9	19	33	54	30
k-means [14]	61	31	27	**69**	9	27	58	34	36	12	25	21	64	38	70	18	14	25	51	25
Ours - Sound Texture	**76**	58	45	57	**20**	**60**	**76**	48	44	35	46	42	**75**	**69**	**90**	**33**	34	**43**	**76**	43
Ours - MFCC	74	**60**	**48**	57	20	54	**76**	49	**45**	**37**	**51**	**43**	74	**69**	87	31	40	42	75	**45**

Update - July 2020

Method	Architecture	Accuracy
Colorization	R101	39.6
Jigsaw	R50w2x	44.6
Exemplar	R50w3x	46
DeepCluster	VGG	48.4
Relative Position	R50w2x	51.4
Rotation	Rv50w4x	55.4
BigBiGAN	Rv50w4x	61.3
SimCLRv1	R50w4x	64.5
MoCo	R50w4x	68.6
SimCLRv2	R50w4x	69.28
ImageGPT	iGPT-XL	72
SimCLRv2	R152w3x	76.6
This method	R50	49.71
Pretrained SOTA	RX-101 32x48d	88.5

Imagenet top 1 accuracy - using linear probe

Unsupervised pretraining dataset differs*

Improvements

Bigger models

Better sound representation, WaveNet/Contrastive Predictive Coding (CPC)

Audio augmentation to produce a contrastive learning environment? Image augmentation paired with audio clusters?...(several ideas to steer it towards SimCLR)

Built on top of a long list of (audio related) works

- Andrew Owens, Alexei A. Efros. *Audio-Visual Scene Analysis with Self-Supervised Multisensory Features*
- *Andrew Owens, Jiajun Wu, Josh McDermott, William T. Freeman, Antonio Torralba. Ambient Sound Provides Supervision for Visual Learning*
- Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, In So Kweon. *Learning to Localize Sound Source in Visual Scenes*
- Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, Michael Rubinstein. *Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation*
- Aviv Gabbay, Asaph, Shamir, Shmuel Peleg. *Visual Speech Enhancement using Noise-Invariant Training*
- Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, Antonio Torralba. *The Sound of Pixels*
- Relja Arandjelovic, Andrew Zisserman. *Objects that Sound*
- Ruohan Gao, Rogerio Feris, Kristen Grauman. *Learning to Separate Object Sounds by Watching Unlabeled Video*
- Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman. *The Conversation: Deep Audio-Visual Speech Enhancement*