http://bit.ly/dlsl2018

#DLUPC

Day 4 Lecture 3

Audio and Vision

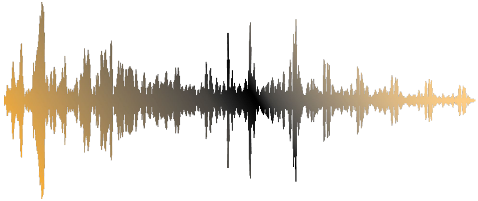
Xavier Giro-i-Nieto
xavier.giro@upc.edu

Associate Professor
Universitat Politecnica de Catalunya
Technical University of Catalonia

telecom BCN

UPC

Audio & Vision

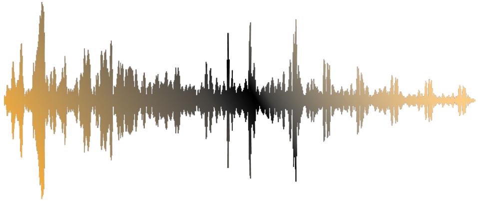


Audio

Vision

Speech

Audio & Vision



Audio

Vision

Video

Speech

Synchronization among modalities captured by **video** is exploited in a self-supervised manner.

Audio & Vision

- Feature Learning
- Cross-modal Retrieval
- Cross-modal Translation

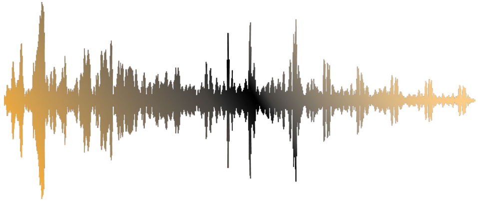
Audio & Vision

- **Feature Learning**
- Cross-modal Retrieval
- Cross-modal Translation

Visual Feature Learning

Vision

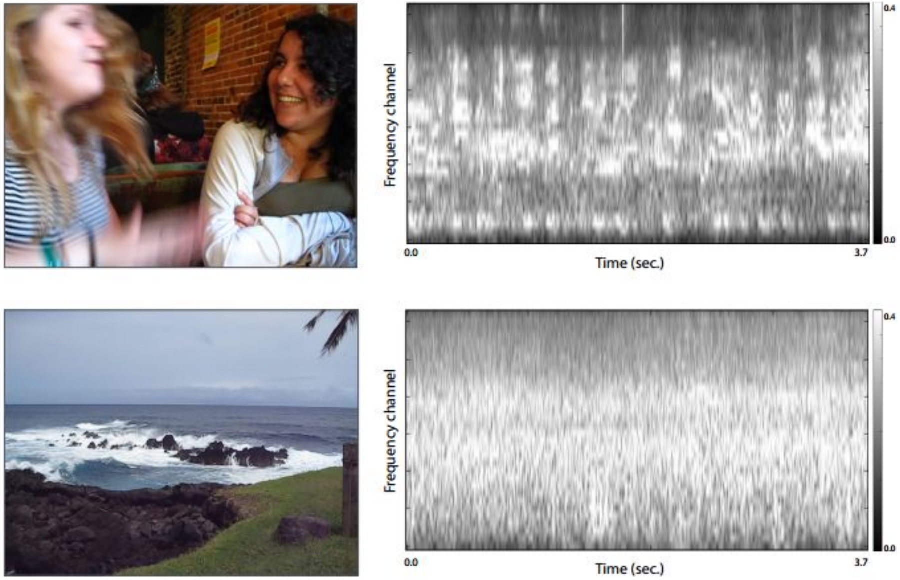
Video



Audio

Visual Feature Learning

Based on the assumption that ambient sound in video is related to the visual semantics.

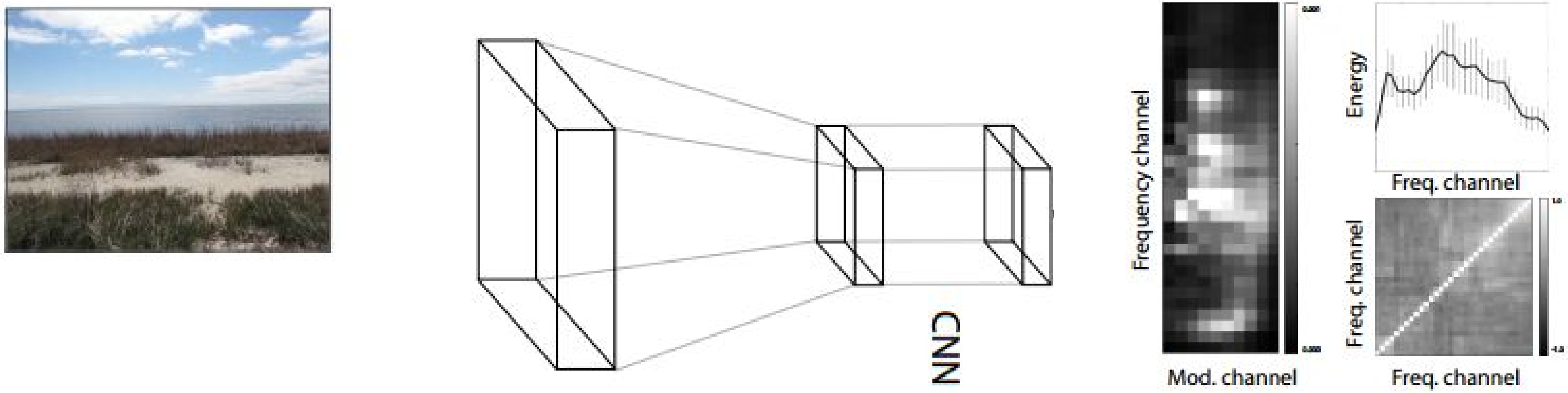


(a) Video frame (b) Cochleagram

Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

Visual Feature Learning

Use videos to train a CNN that predicts the audio statistics of a frame.

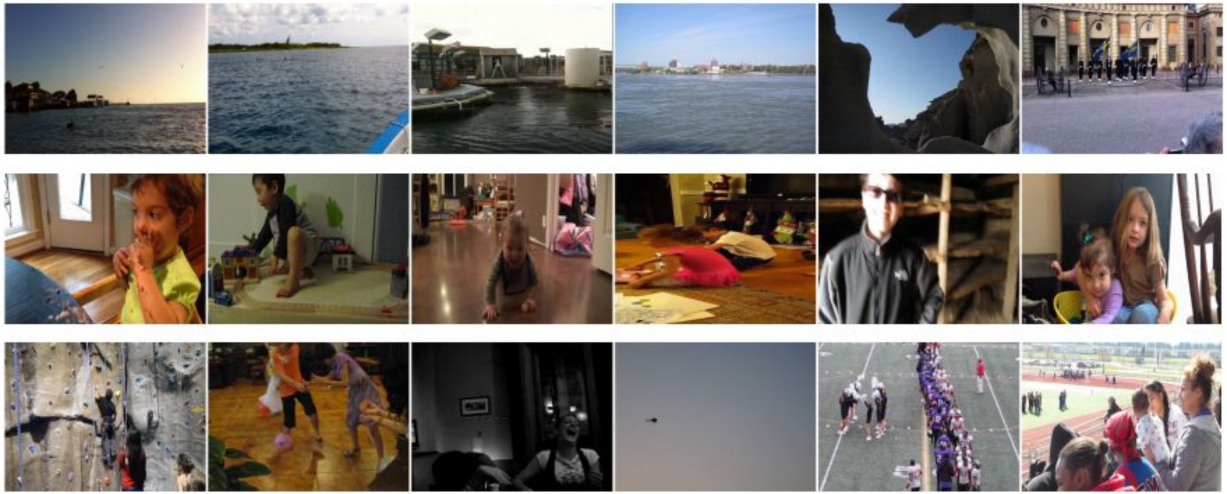


Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

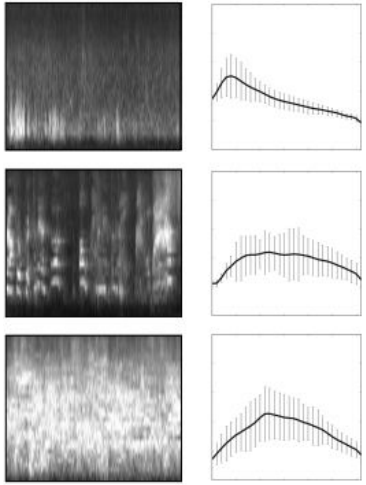
Visual Feature Learning

Task: Use the predicted audio stats to clusters images. Audio clusters built with K-means over training set

Cluster assignments at test time (one row=one cluster)



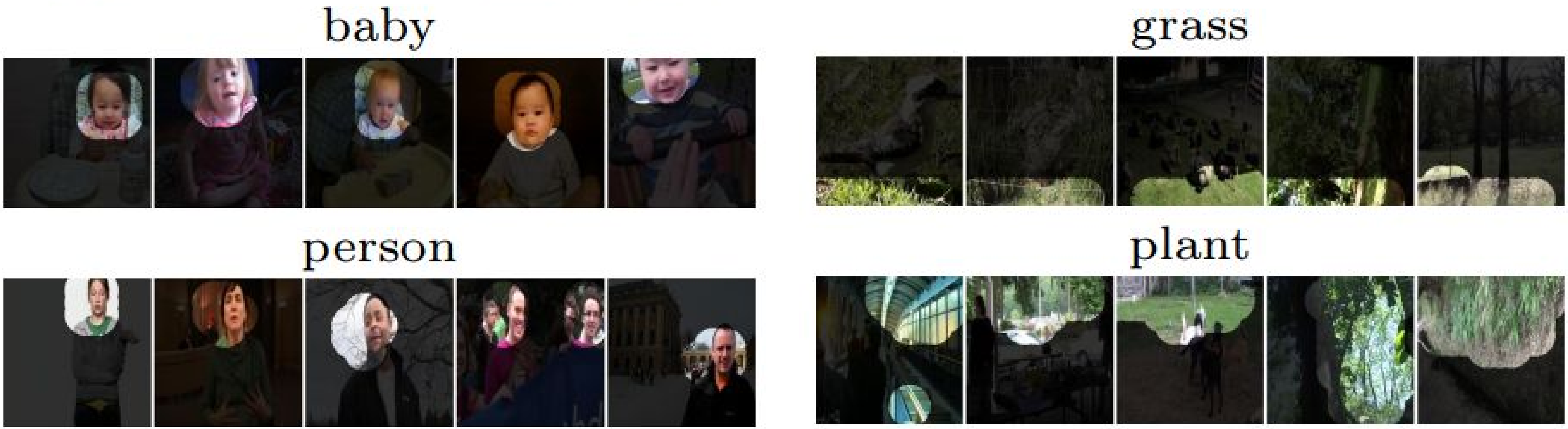
Average stats



Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

Visual Feature Learning

Although the CNN was not trained with class labels, local units with semantic meaning emerge.

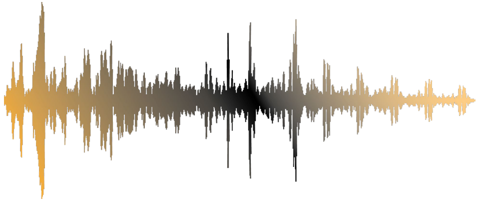


Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. ["Ambient sound provides supervision for visual learning."]() ECCV 2016

Audio Feature Learning

Vision

Video

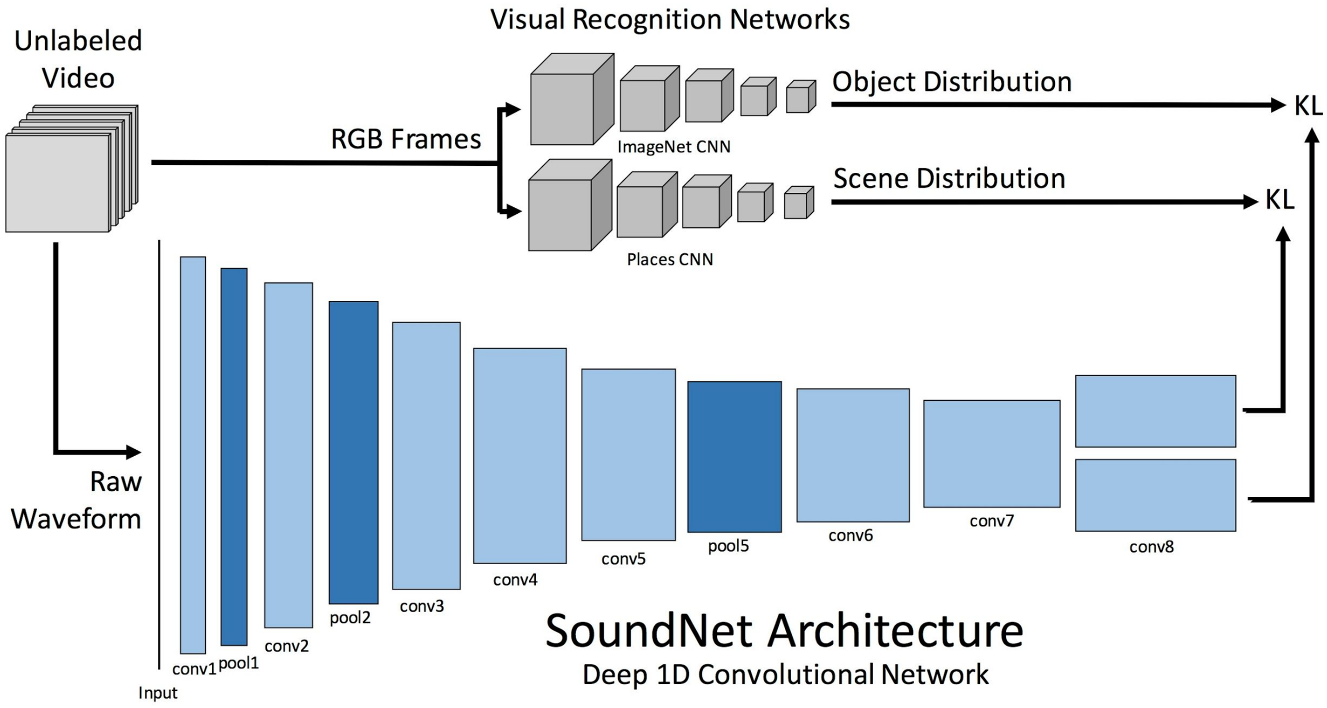


Audio

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Audio Feature Learning: SoundNet

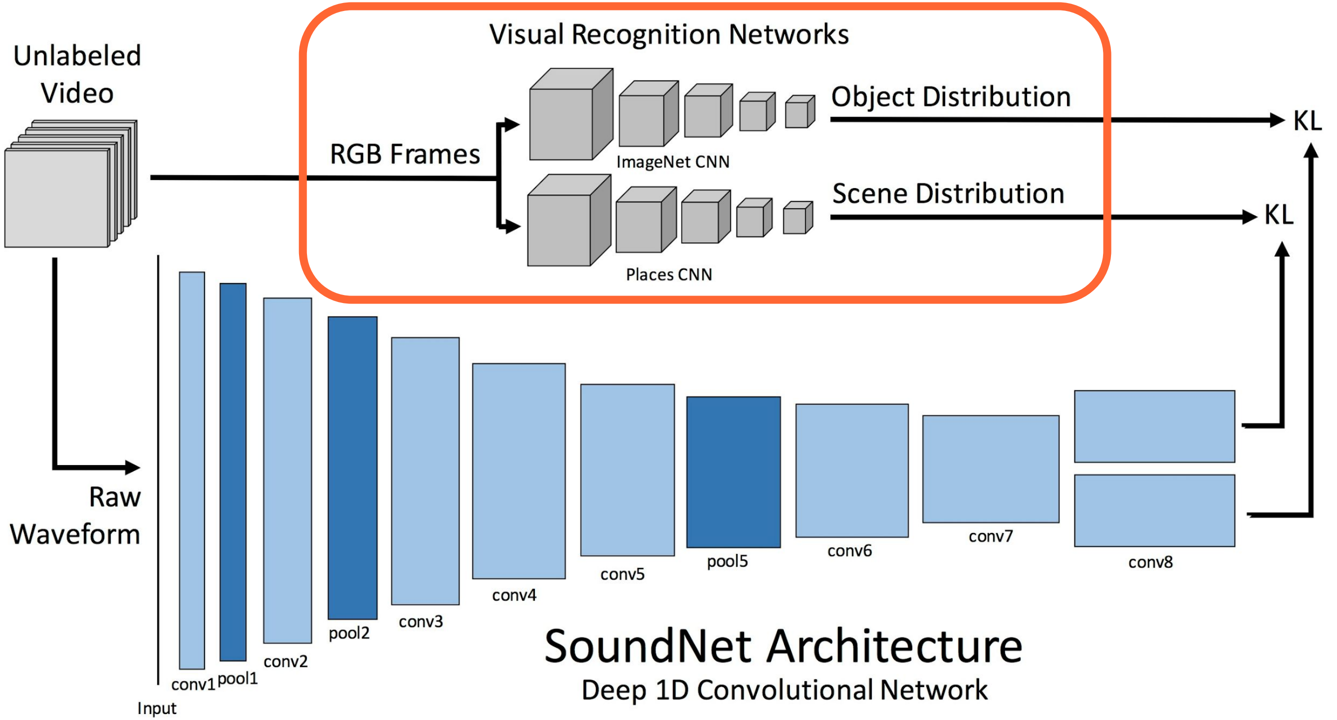
Pretrained visual ConvNets supervise the training of a model for sound representation



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning: SoundNet

Videos for training are unlabeled. Relies on Convnets trained on labeled images.



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning: SoundNet

Hidden layers of Soundnet are used to train a standard SVM classifier that outperforms state of the art.

Method	Accuracy
RG [29]	69%
LTT [21]	72%
RNH [30]	77%
Ensemble [34]	78%
SoundNet	**88%**

Table 3: **Acoustic Scene Classification on DCASE:** We evaluate classification accuracy on the DCASE dataset. By leveraging large amounts of unlabeled video, SoundNet generally outperforms hand-crafted features by 10%.

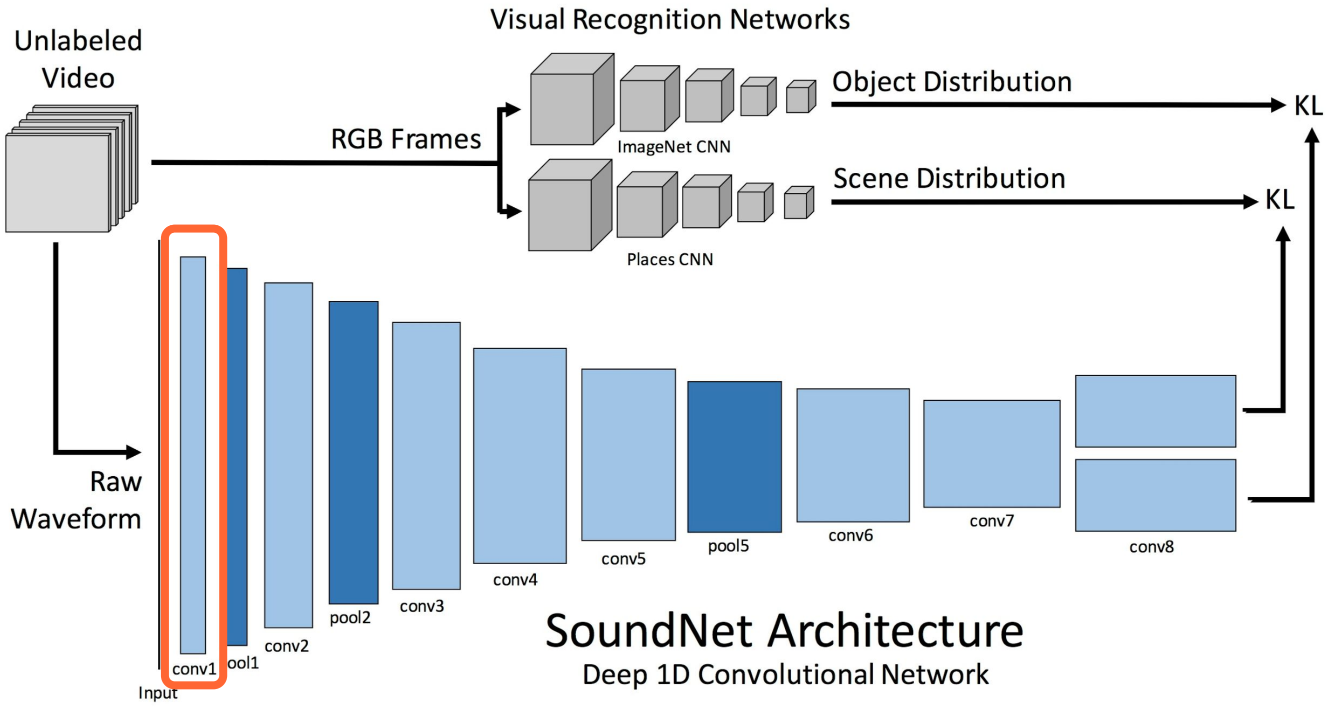
	Accuracy on	
Method	ESC-50	ESC-10
SVM-MFCC [28]	39.6%	67.5%
Convolutional Autoencoder	39.9%	74.3%
Random Forest [28]	44.3%	72.7%
Piczak ConvNet [27]	64.5%	81.0%
SoundNet	**74.2%**	**92.2%**
Human Performance [28]	81.3%	95.7%

Table 4: **Acoustic Scene Classification on ESC-50 and ESC-10:** We evaluate classification accuracy on the ESC datasets. Results suggest that deep convolutional sound networks trained with visual supervision on unlabeled data outperforms baselines.

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning: SoundNet

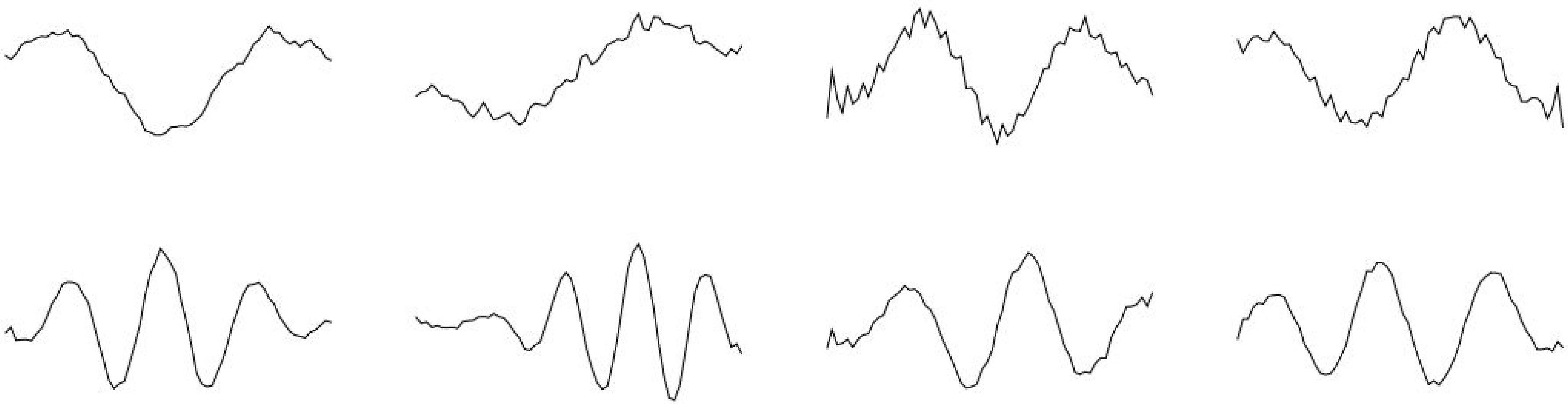
Visualization of the 1D filters over raw audio in conv1.



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning: SoundNet

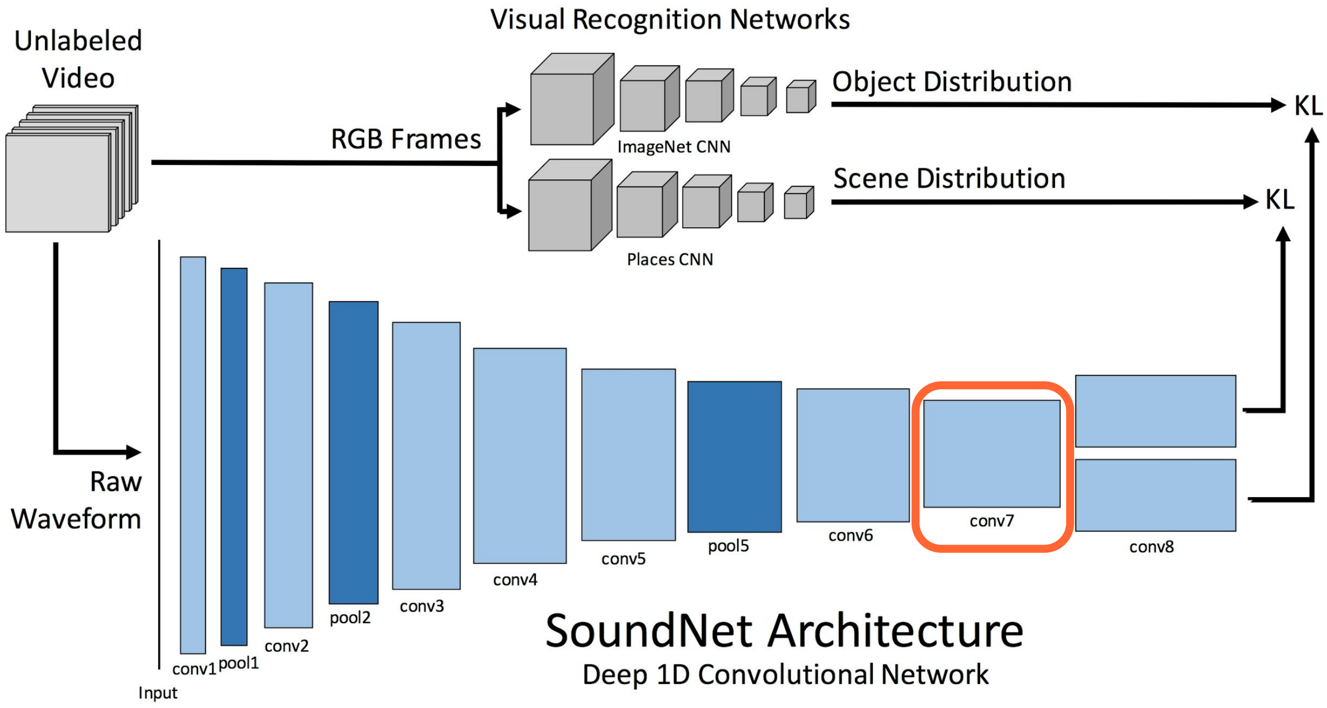
Visualization of the 1D filters over raw audio in conv1.



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning: SoundNet

Visualize samples that mostly activate a neuron in a late layer (conv7)



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning: SoundNet

Visualization of the video frames associated to the sounds that activate some of the last hidden units (conv7):

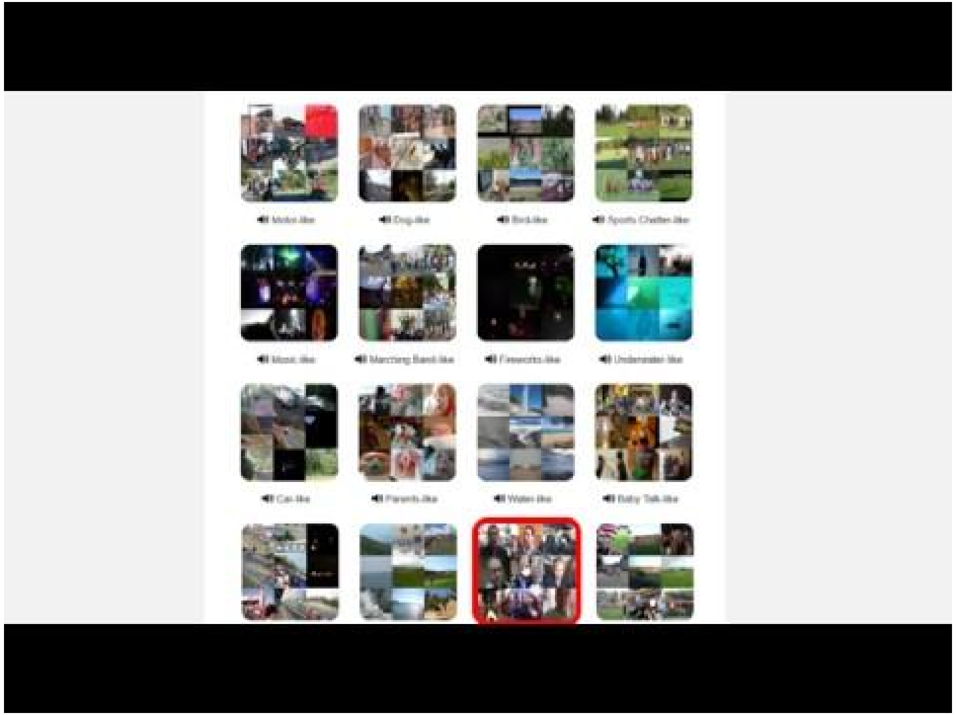
Baby Talk

Bubbles

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning: SoundNet

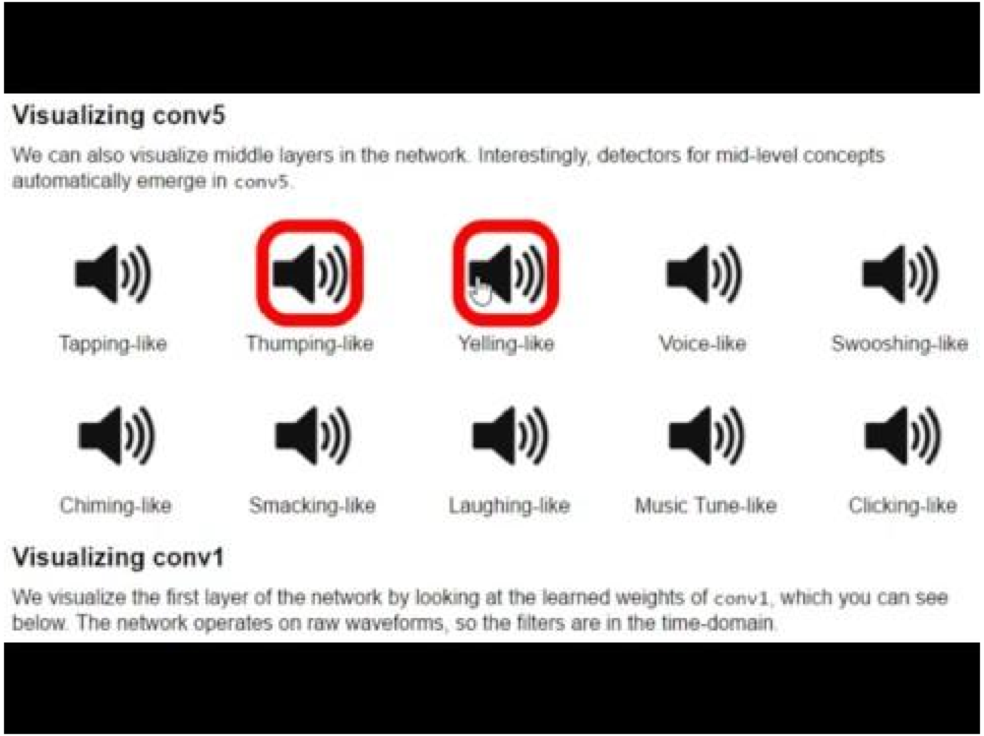
Hearing sounds that most activate a neuron in the sound network (conv7)



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning: SoundNet

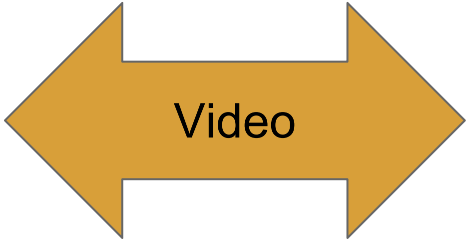
[Hearing sounds that most activate a neuron in the sound network](#) (conv5)

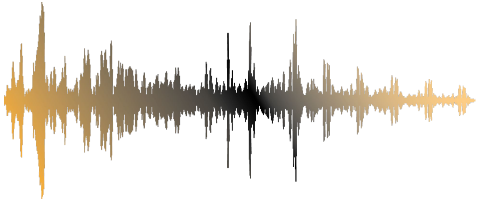


Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio & Visual Feature Learning

Vision

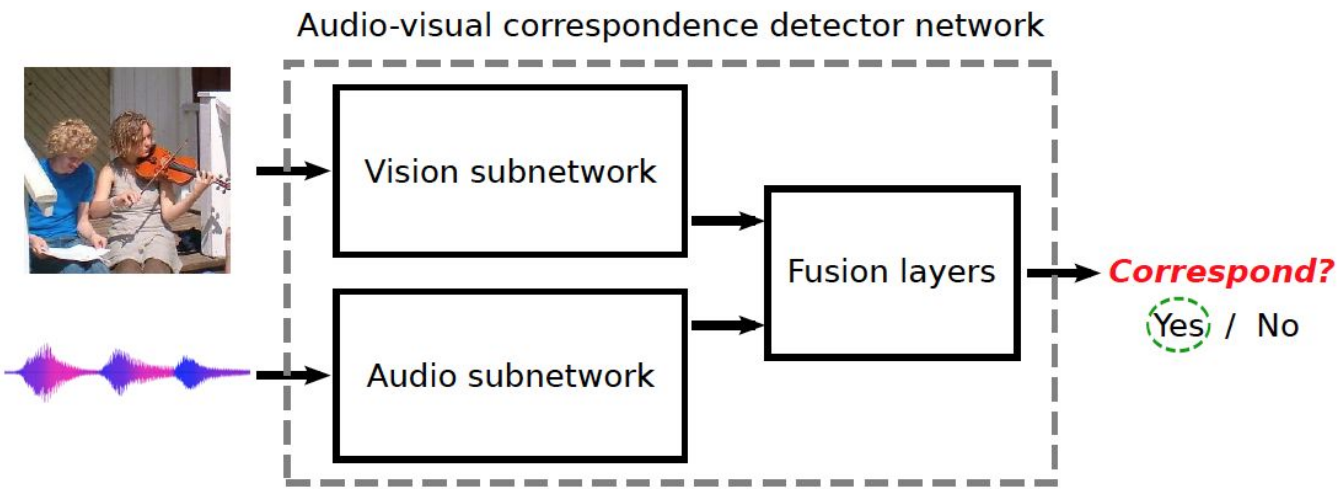


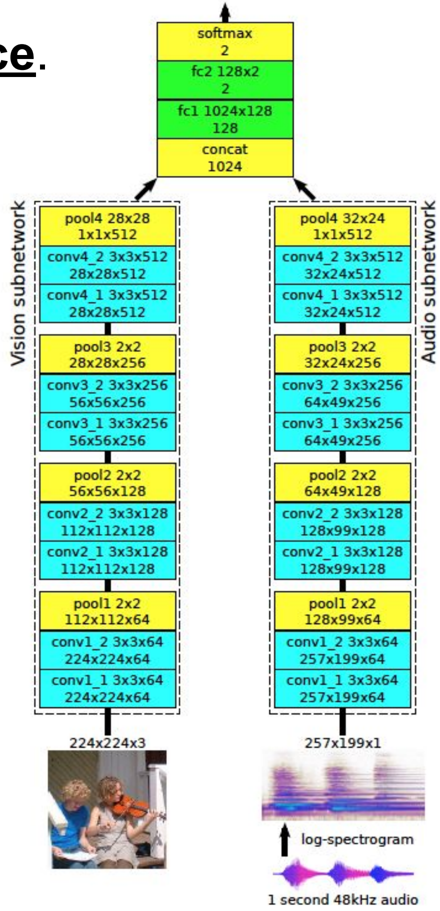


Audio

Audio & Visual Feature Learning

Audio and visual features learned by assessing **<u>correspondence</u>**.





Arandjelović, Relja, and Andrew Zisserman. ["Look, Listen and Learn."]() *ICCV 2017*

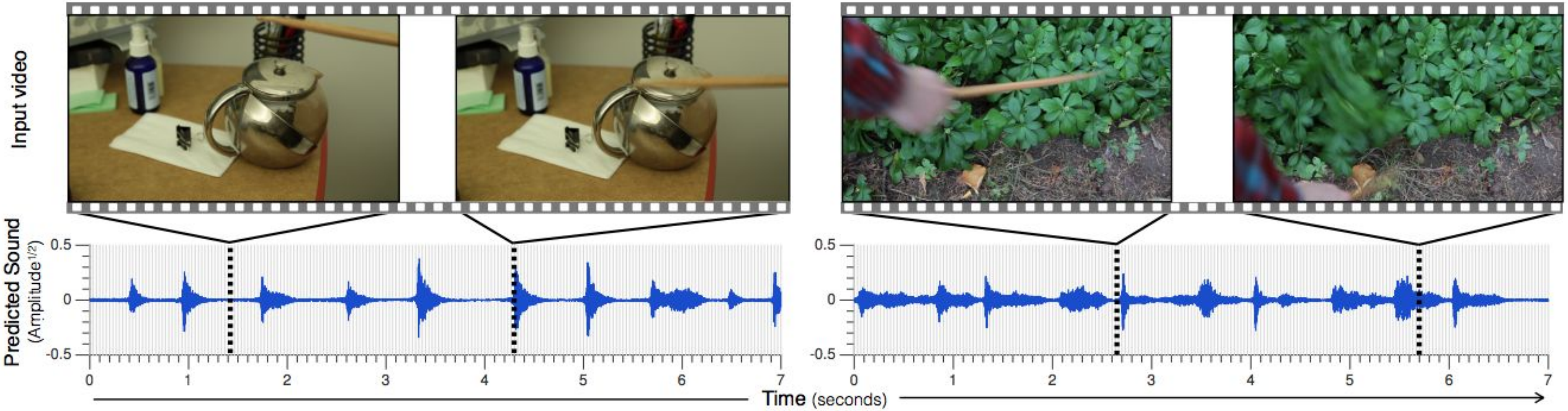
Audio & Vision

- Feature Learning
- **Cross-modal retrieval**
- Cross-modal Translation

Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Cross-modal Retrieval

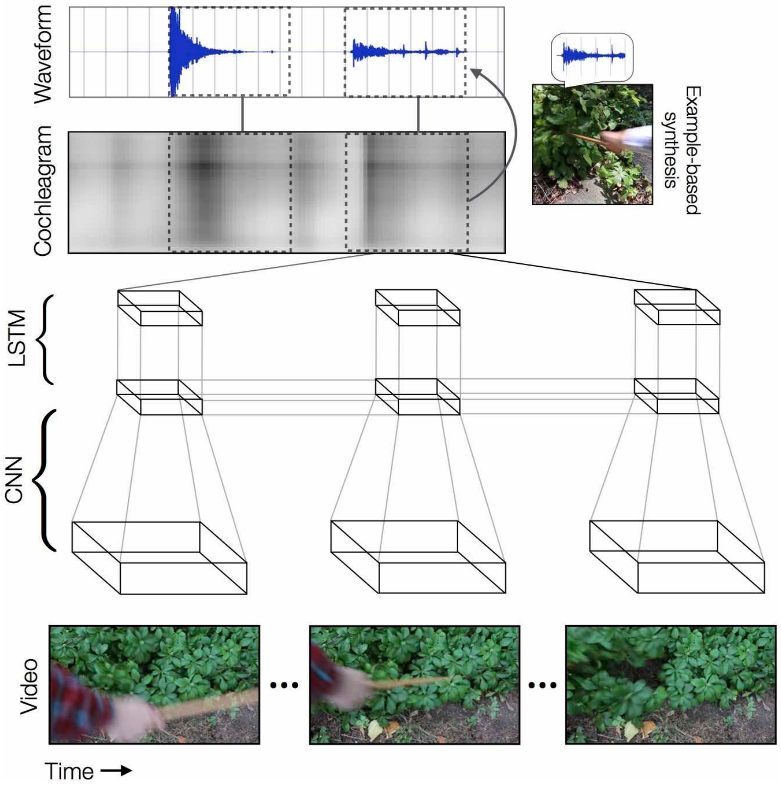
Learn synthesized sounds from videos of people hitting objects with a drumstick.



Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. ["Visually indicated sounds."]() CVPR 2016.

Cross-modal Retrieval

Not end-to-end



Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Cross-modal Retrieval

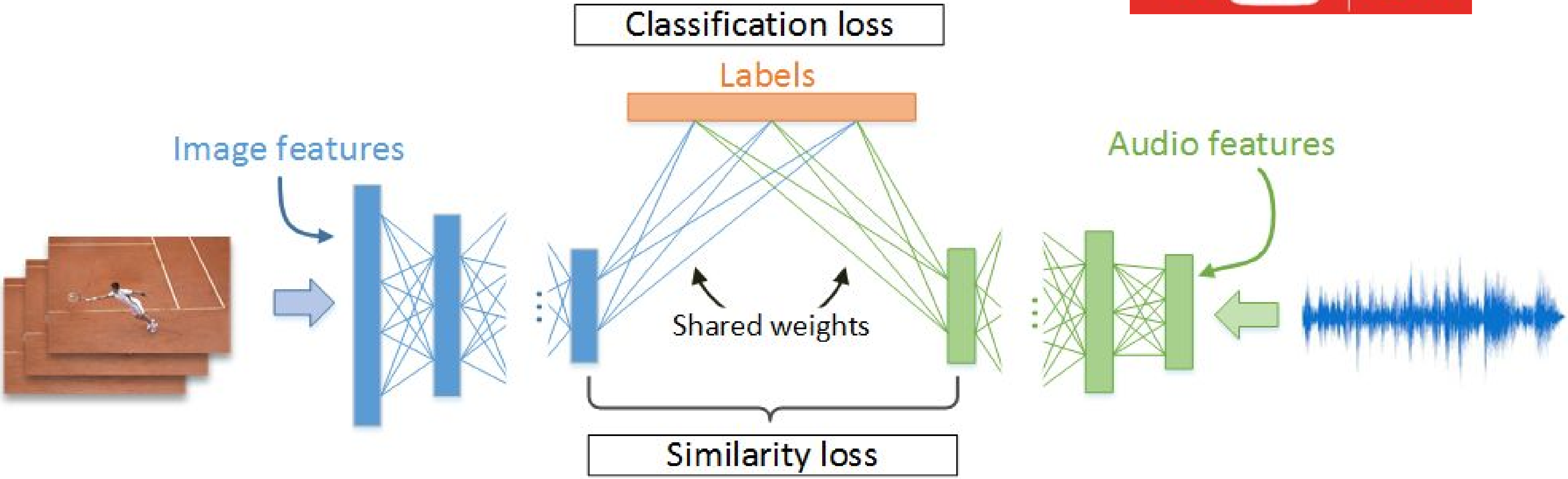
The Greatest Hits Dataset

Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. ["Visually indicated sounds."]() CVPR 2016.

Cross-modal Retrieval

[Paper draft]

YouTube 8M



Surís, Didac, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. "Cross-modal Embeddings for Video and Audio Retrieval." arXiv preprint arXiv:1801.02200 (2018).

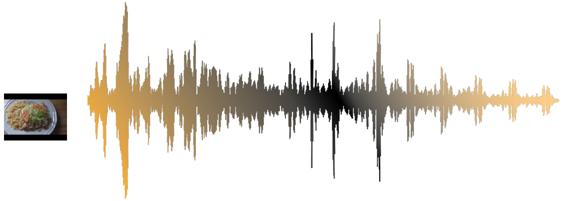
Cross-modal Retrieval

Video sonorization

Visual feature

Best match

Audio feature



Surís, Didac, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. ["Cross-modal Embeddings for Video and Audio Retrieval."]() arXiv preprint arXiv:1801.02200 (2018).

Cross-modal Retrieval

Audio coloring

Visual feature

Audio feature

Surís, Didac, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. ["Cross-modal Embeddings for Video and Audio Retrieval."]() arXiv preprint arXiv:1801.02200 (2018).

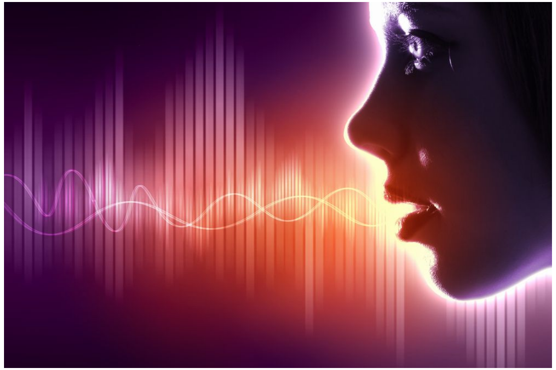
Audio & Vision

- Feature Learning
- Cross-modal retrieval
- **Cross-modal Translation**

Audio & Vision

Vision

Video

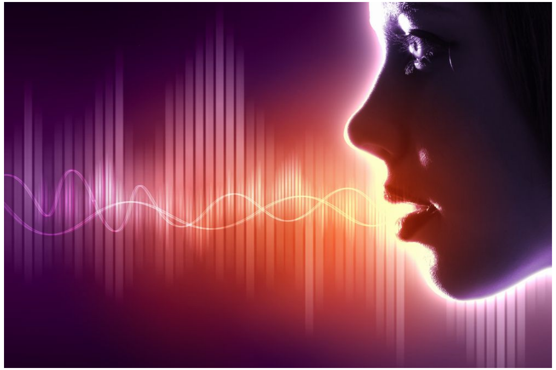


Speech

Audio & Vision

Vision

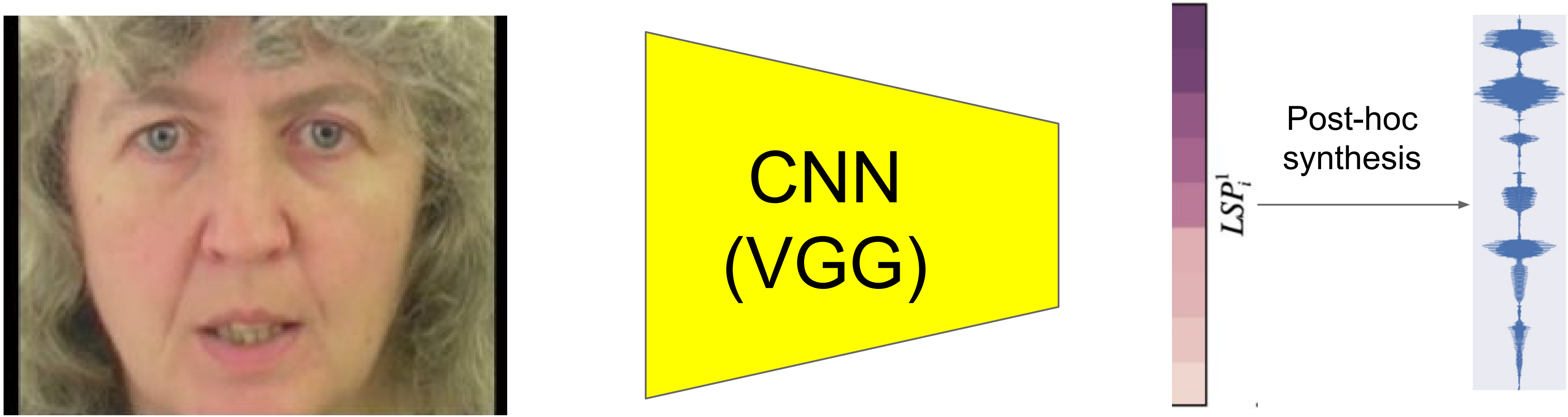
Video



Speech

Ephrat, Ariel, Tavi Halperin, and Shmuel Peleg. "Improved speech reconstruction from silent video." In ICCV 2017 Workshop on Computer Vision for Audio-Visual Media. 2017.

Speech Generation from Video

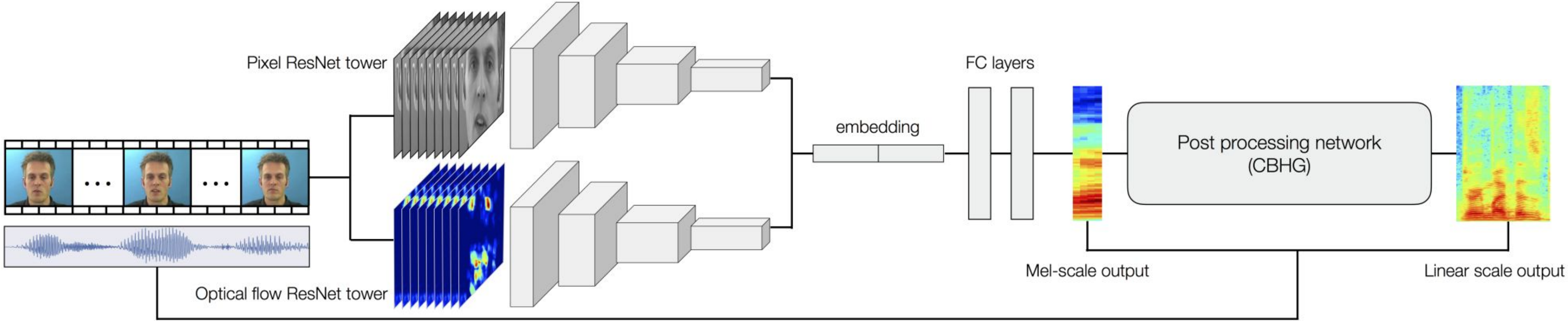


Frame from a silent video

Audio feature

Ephrat et al. [Vid2speech: Speech Reconstruction from Silent Video.]() *ICASSP 2017*

Speech Generation from Video



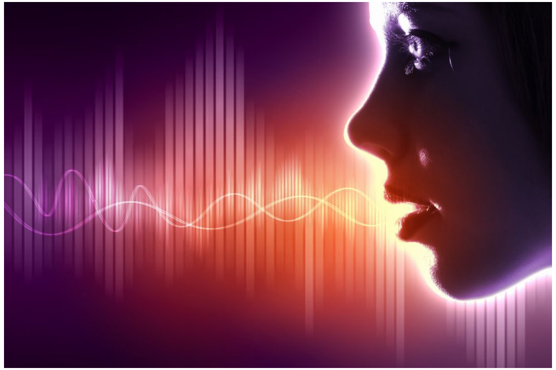
Ephrat, Ariel, Tavi Halperin, and Shmuel Peleg. "Improved speech reconstruction from silent video." In ICCV 2017 Workshop on Computer Vision for Audio-Visual Media. 2017.

Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." BMVC 2017.

Audio & Vision

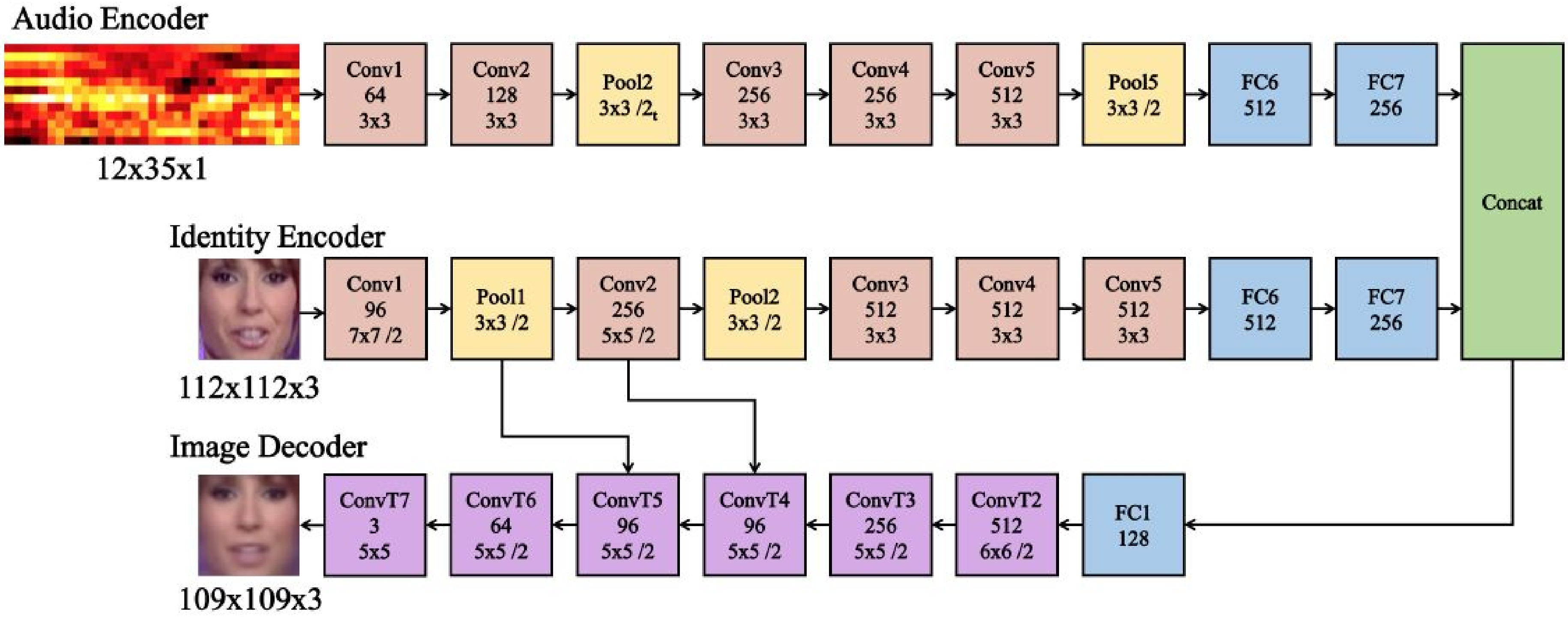
Vision

Video



Speech

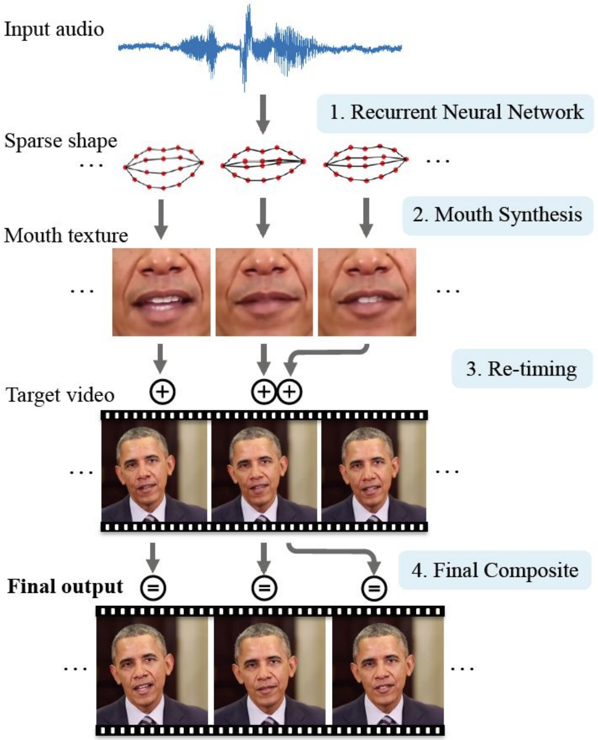
Speech to Video Synthesis (mouth)



Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. ["You said that?."]() BMVC 2017.

Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." SIGGRAPH 2017

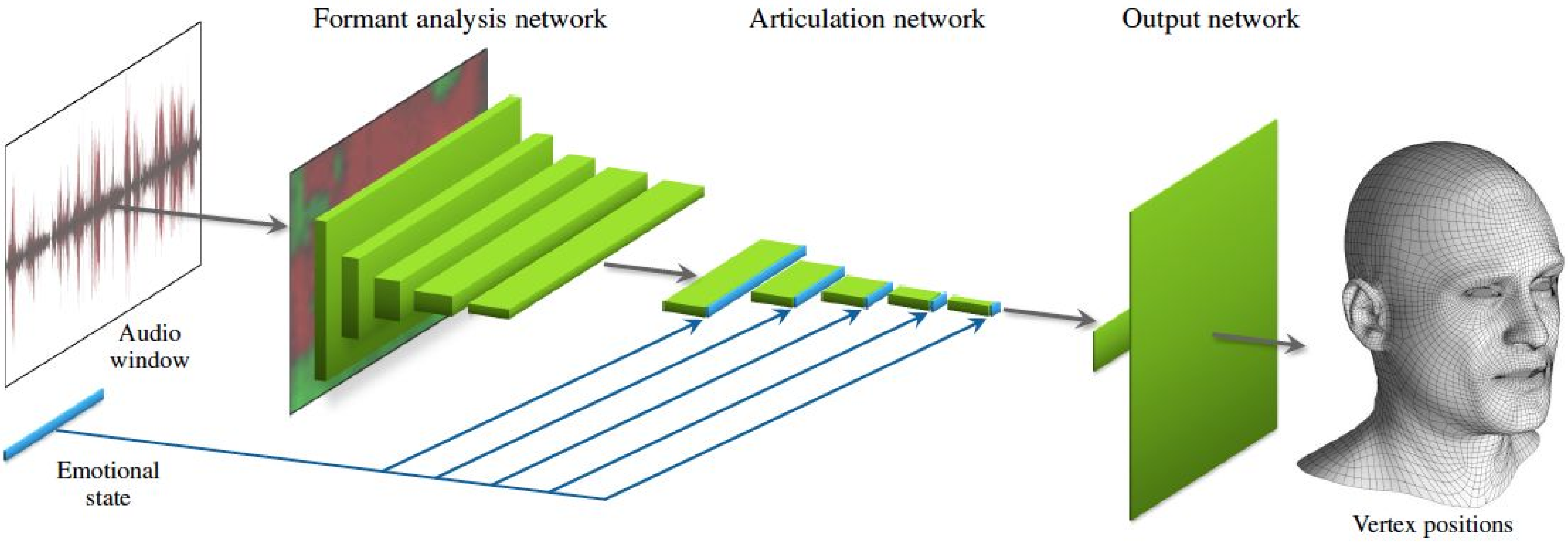
Speech to Video Synthesis (mouth)



Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. ["Synthesizing Obama: learning lip sync from audio."]() SIGGRAPH 2017.

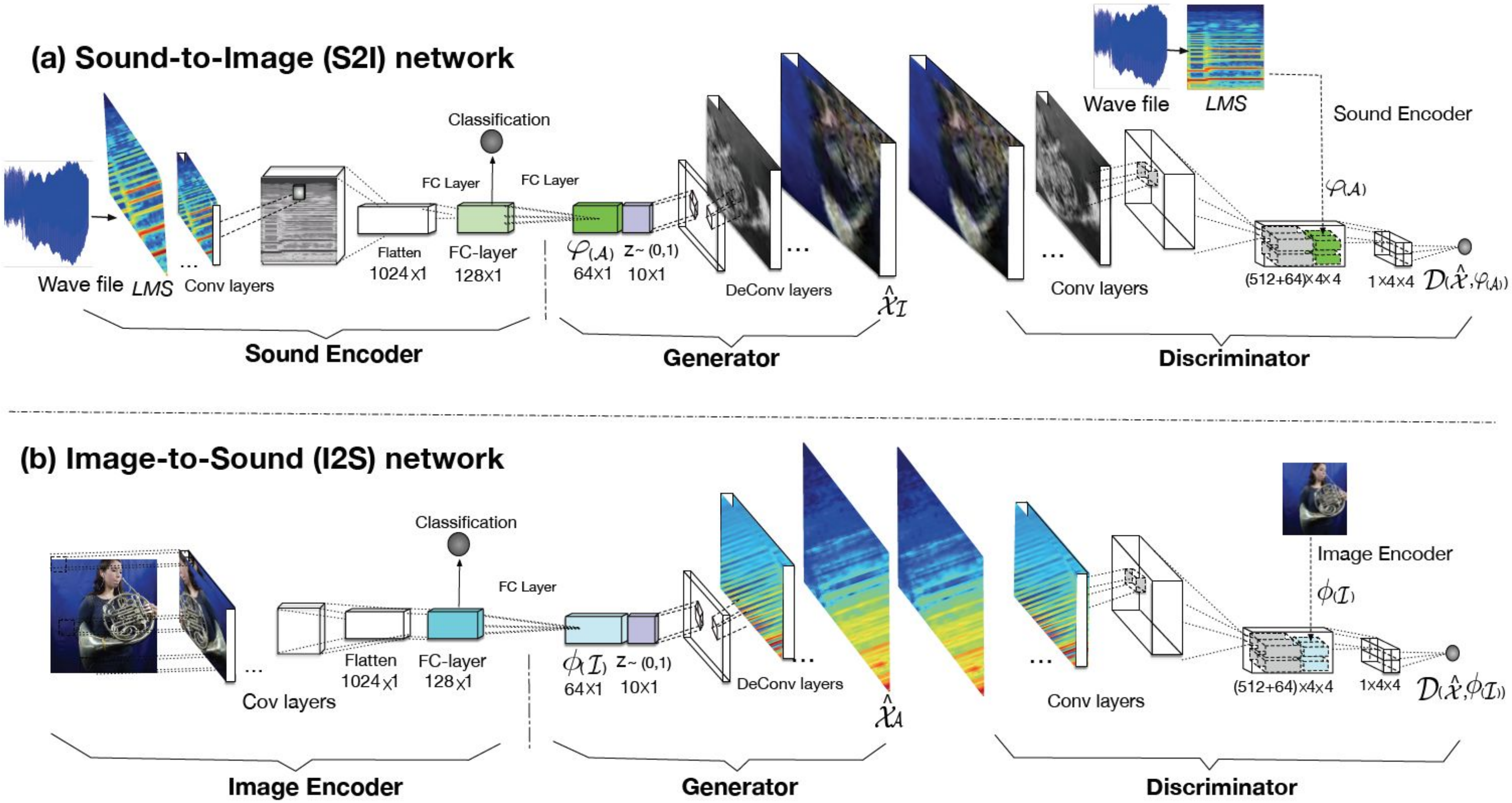
Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." SIGGRAPH 2017

Speech to Video Synthesis (pose & emotion)



Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. ["Audio-driven facial animation by joint end-to-end learning of pose and emotion."]() SIGGRAPH 2017

Audio & Visual Generation



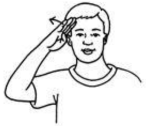
L. Chen, S. Srivastava, Z. Duan and C. Xu. Deep Cross-Modal Audio-Visual Generation. ACM Multimedia Thematic Workshops 2017.

Speech2Signs (under work)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory
and Communications
Image Processing Group

TALP UPC

Hello → ASR → "Hello" → NMT → SLPA → Video Generator

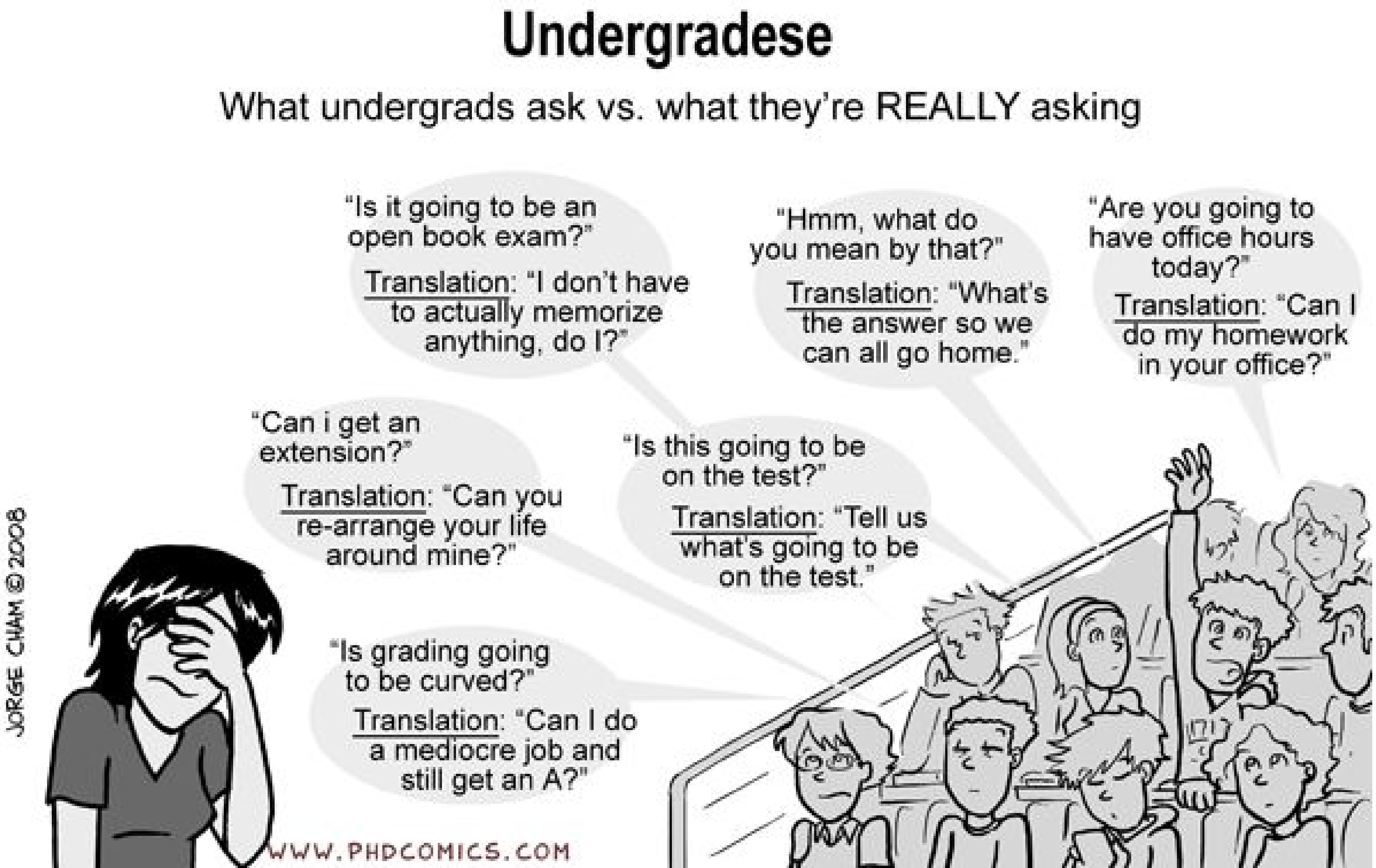


Caffe2 facebook research

Audio & Vision

- Feature Learning
- Cross-modal retrieval
- Cross-modal Translation

Questions ?


Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?" Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?" Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?" Translation: "Can I do my homework in your office?"
"Can i get an extension?" Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?" Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?" Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM