Topic 13

Audio-Visual Scene Understanding

Multi-Sensory Perception

- Vision and audition arguably receive the most information
  - Survival
  - Communication
  - Information Preservation

AV Complementary

- Vision
  - Color, brightness, shape, texture
  - Spatial
  - Temporal

  - Accurate and detailed
  - Subject to lighting conditions, occlusions and view angles

- Audition
  - Loudness, frequency, timbre, location
  - Spatial
  - Temporal

  - Ambiguous
  - 360 degrees
  - Requires less attention, almost 24/7

AV Mutual Enhancement

- ## Vision enhances audition
  - – Lip reading
  - – Speech production of infants

- ## Audition enhances vision
  - – Visual search faster with spatial audio
  - – Sound track enhances emotion of movies

AV Conflicts & Compromises

- Ventriloquism Effect
  - https://youtu.be/aYTJJDjjNSY

- Sound Induced Flash Illusion
  - https://www.illusionsindex.org/i/sound-induced-flash

- McGurk Effect
  - /ba/ (a) + /ga/ (v)= /da/ (av)
  - https://youtu.be/jtsfidRq2tw

AV Scene Understanding

- How to build computational models to learn audio-visual representations?

- How to design algorithms to achieve better scene understanding than those with single modalities?

- How to use one modality to learn, infer and generate the other modality?

Some Tasks

- Cross-Modal Supervision

- Audio-Visual Association

- Visually Informed Audio Source Separation

- Cross-Modal Generation

Cross-Modal Supervision

Aytar, et al., "SoundNet: Learning sound representations from unlabeled video," NIPS 2016

- 2M YouTube videos.
- Visual Recognition Networks provide labels for training SoundNet.
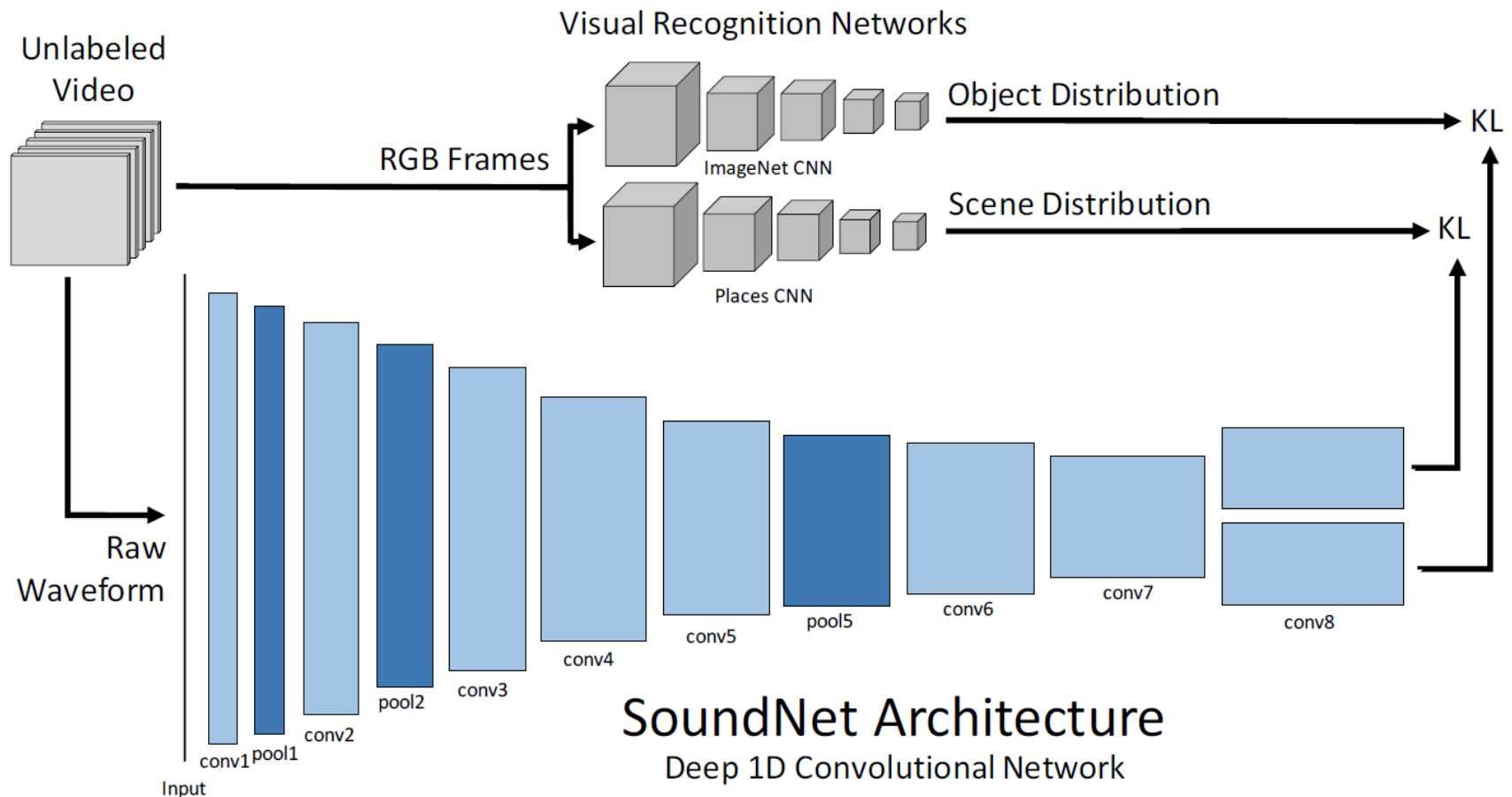
Sound Classification Results

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%.

Method	Accuracy on	
	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.

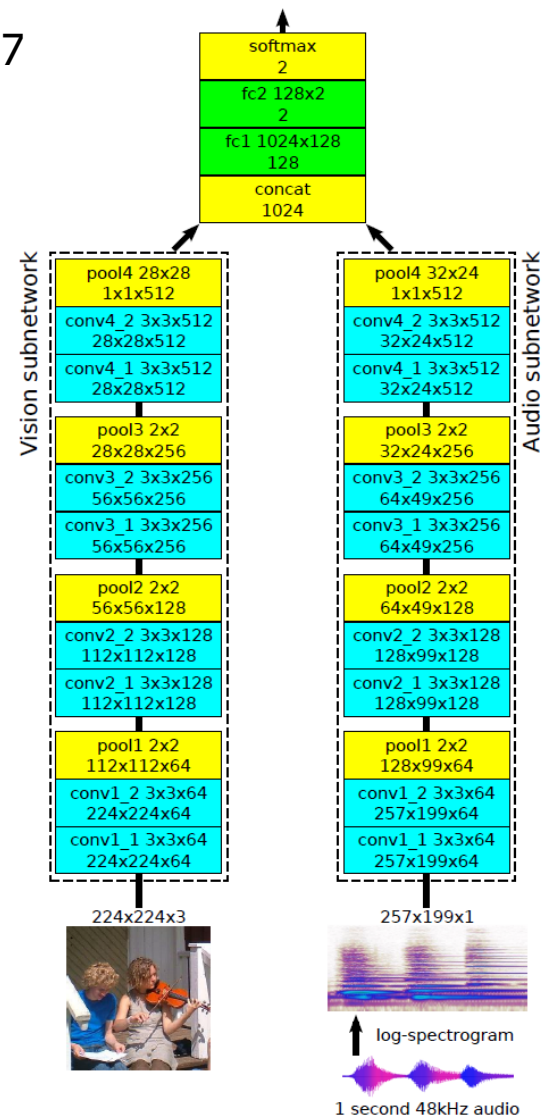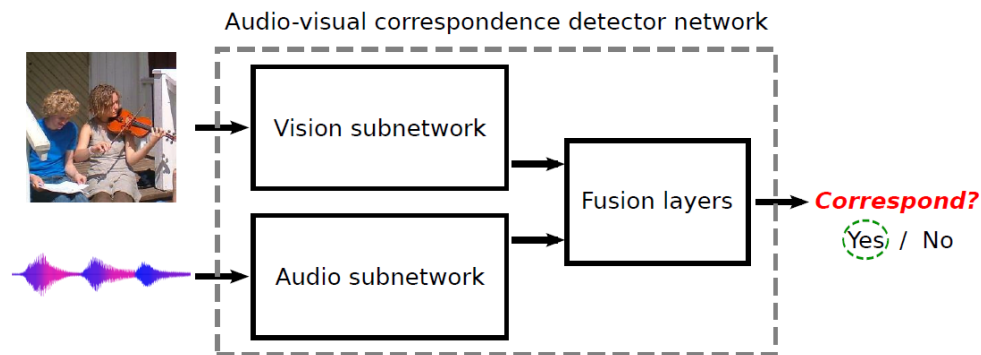

Baby Talk   Bubbles   Cheering   Bird Chirps

Figure 6: **What emerges in sound hidden units?** We visualize some of the hidden units in the last hidden layer of our sound representation by finding inputs that maximally activate a hidden unit. Above, we illustrate what these units capture by showing the corresponding video frames. No vision is used in this experiment; we only show frames for visualization purposes only.

Audio-Visual Association

Arandjelovic & Zisserman, "Look, Listen and Learn", ICCV 2017

- Learn to classify match vs. non-match
- Flickr-SoundNet
  - Unlabelled, unconstrained
  - out of 2M videos, used 500K * 10s videos
- Kinetics-Sounds
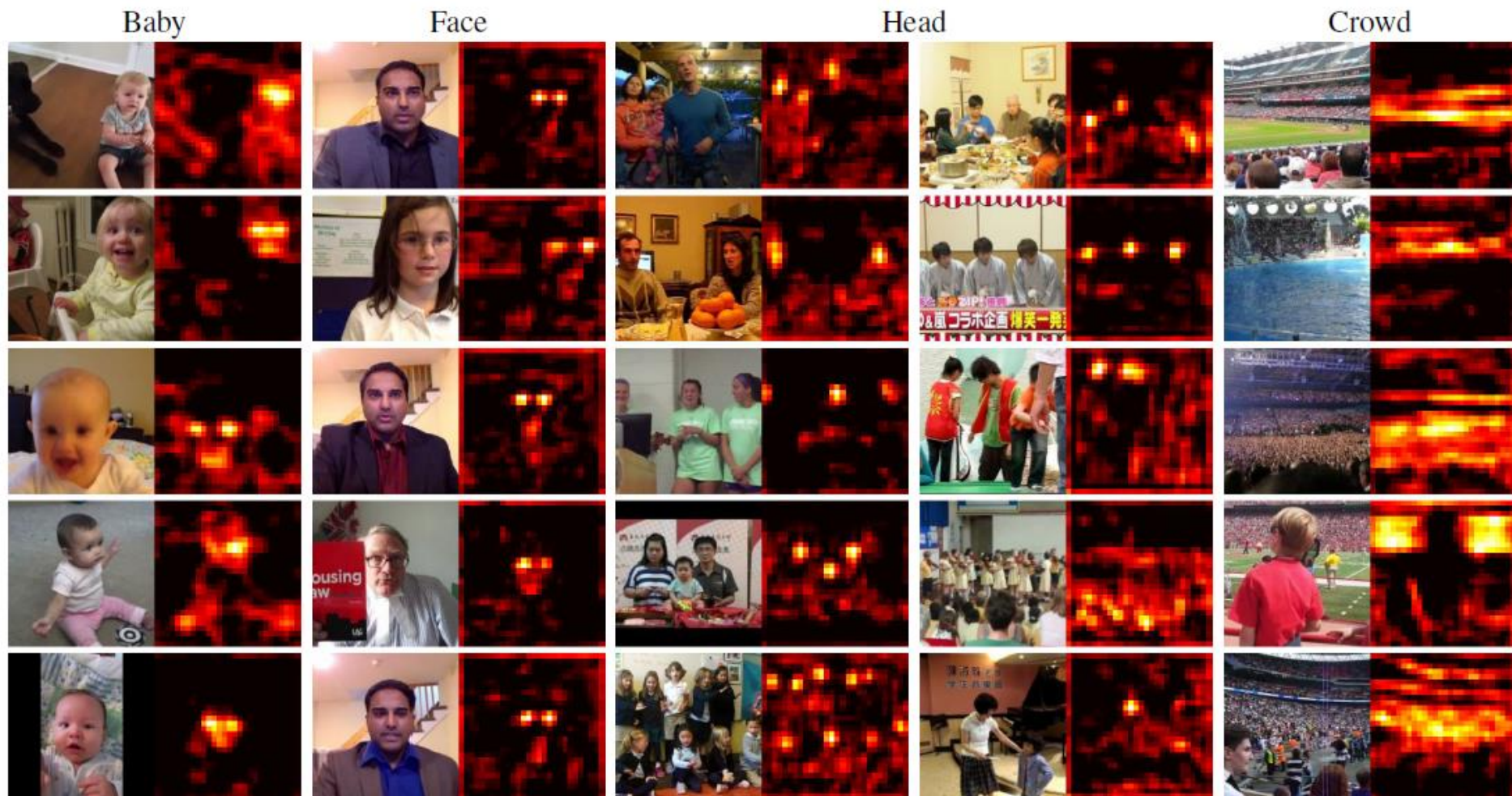 - 19K * 10s videos with human action annotations

Audio-visual correspondence detector network

Vision subnetwork

softmax 2	
fc2 128x2 2	
fc1 1024x128 128	
concat 1024	

Vision subnetwork

pool4 28x28 1x1x512
conv4_2 3x3x512 28x28x512
conv4_1 3x3x512 28x28x512
pool3 2x2 28x28x256
conv3_2 3x3x256 56x56x256
conv3_1 3x3x256 56x56x256
pool2 2x2 56x56x128
conv2_2 3x3x128 112x112x128
conv2_1 3x3x128 112x112x128
pool1 2x2 112x112x64
conv1_2 3x3x64 224x224x64
conv1_1 3x3x64 224x224x64

224x224x3

Audio subnetwork

pool4 32x24 1x1x512
conv4_2 3x3x512 32x24x512
conv4_1 3x3x512 32x24x512
pool3 2x2 32x24x256
conv3_2 3x3x256 64x49x256
conv3_1 3x3x256 64x49x256
pool2 2x2 64x49x128
conv2_2 3x3x128 128x99x128
conv2_1 3x3x128 128x99x128
pool1 2x2 128x99x64
conv1_2 3x3x64 257x199x64
conv1_1 3x3x64 257x199x64

257x199x1

log-spectrogram

1 second 48kHz audio

Learnt Visual Concepts

Figure 6. **Learnt human-related visual concepts and semantic heatmaps (Flickr-SoundNet).** Each mini-column shows five images that most activate a particular unit of the 512 in pool4 of the vision subnetwork, and the corresponding heatmap (for more details see Figures 3 and 4). Column titles are a subjective names of concepts the units respond to.
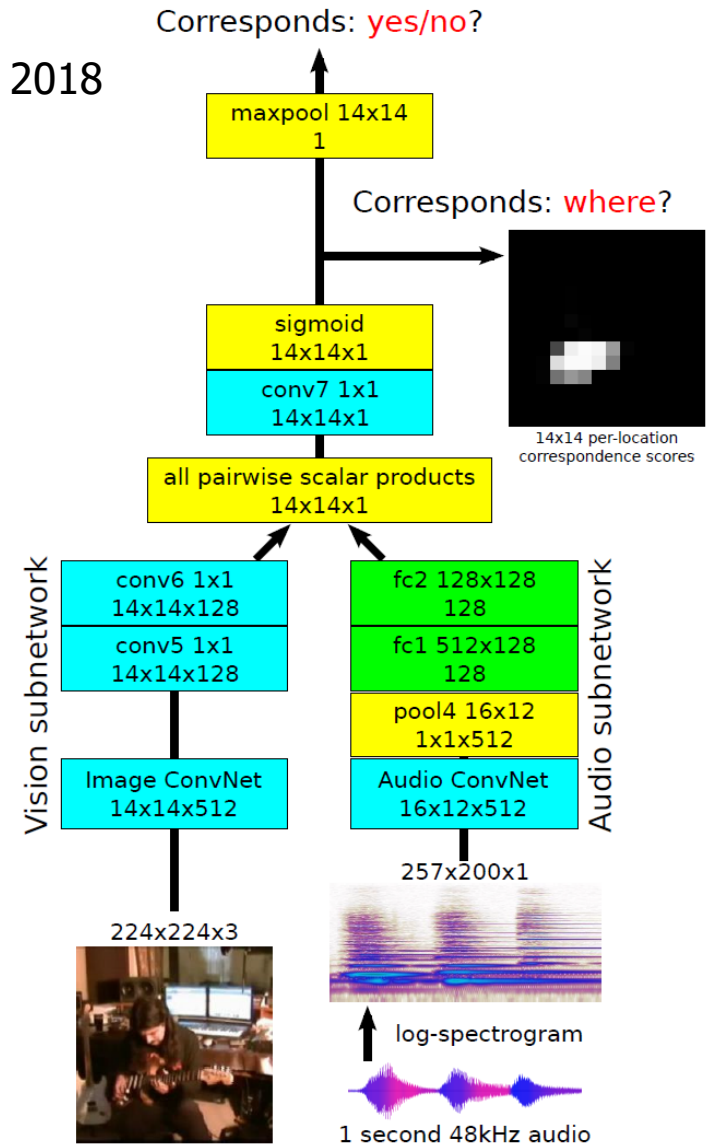
Learnt Audio Concepts
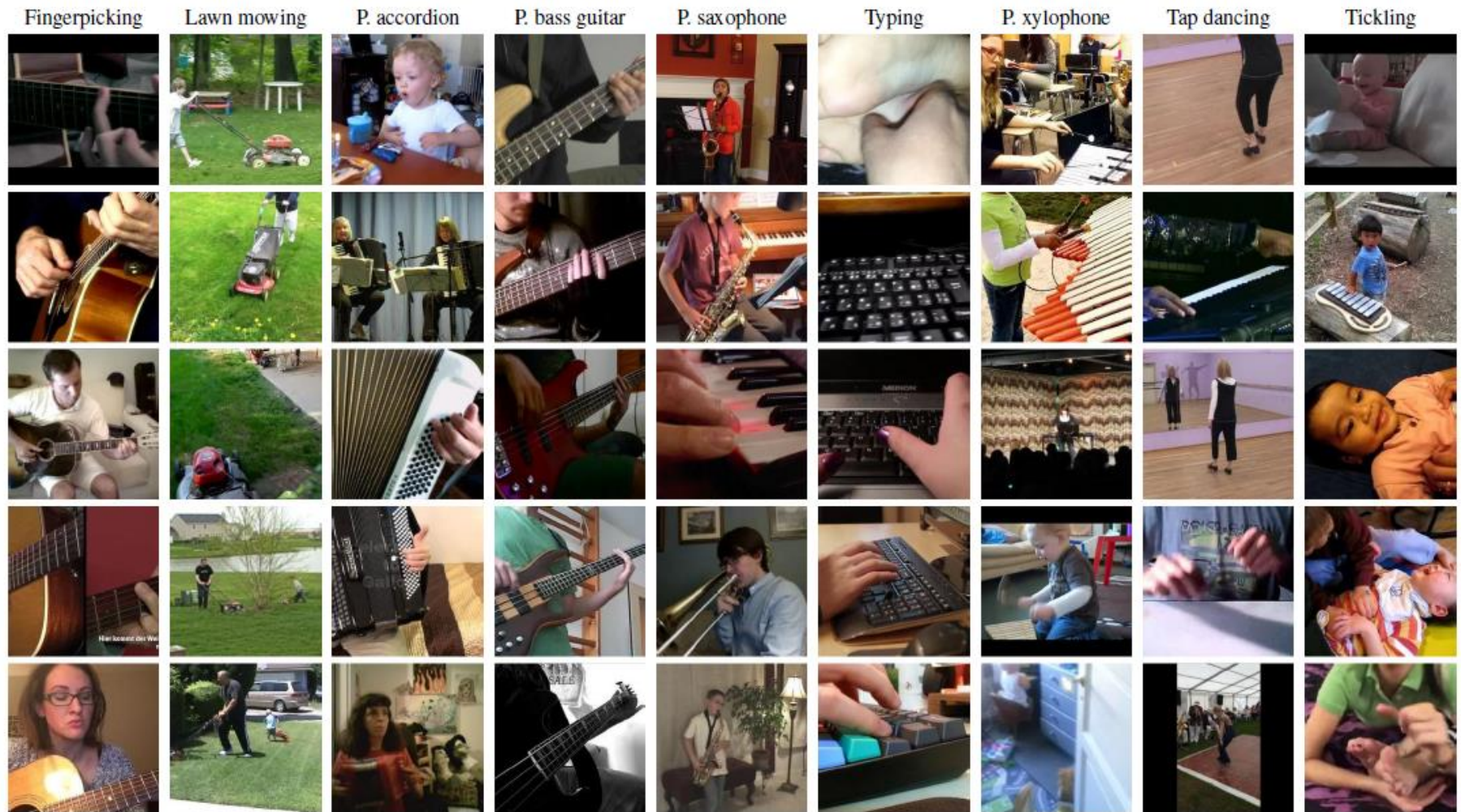
Figure 7. **Learnt audio concepts (Kinetics-Sounds).** Each column shows five sounds that most activate a particular unit in pool4 of the audio subnetwork. Purely for visualization purposes, as it is hard to display sound, the frame of the video that is aligned with the sound is shown instead of the actual sound form, but we stress that no vision is used in this experiment. Videos come from the Kinetics-Sounds test set and the network was trained on the Kinetics-Sounds train set. The top row shows the dominant action label for the unit ("P." stands for "playing").

Localizing Objects That Sound

Arandjelovic & Zisserman, "Objects that sound", ECCV 2018

- Vision network continues to operate at the 14*14 resolution
- Similarities between audio and all vision embeddings reveal the location

Corresponds: yes/no?

maxpool 14x14
1

Corresponds: where?

sigmoid
14x14x1

conv7 1x1
14x14x1

14x14 per-location correspondence scores

all pairwise scalar products
14x14x1

Vision subnetwork

conv6 1x1
14x14x128

conv5 1x1
14x14x128

Image ConvNet
14x14x512

224x224x3

Audio subnetwork

fc2 128x128
128

fc1 512x128
128

pool4 16x12
1x1x512

Audio ConvNet
16x12x512

257x200x1

log-spectrogram

1 second 48kHz audio

Localization Results

Visually Informed Source Separation

Zhao et al., "The sound of pixels", ECCV 2018

Sound generation procedure, i.e., test time
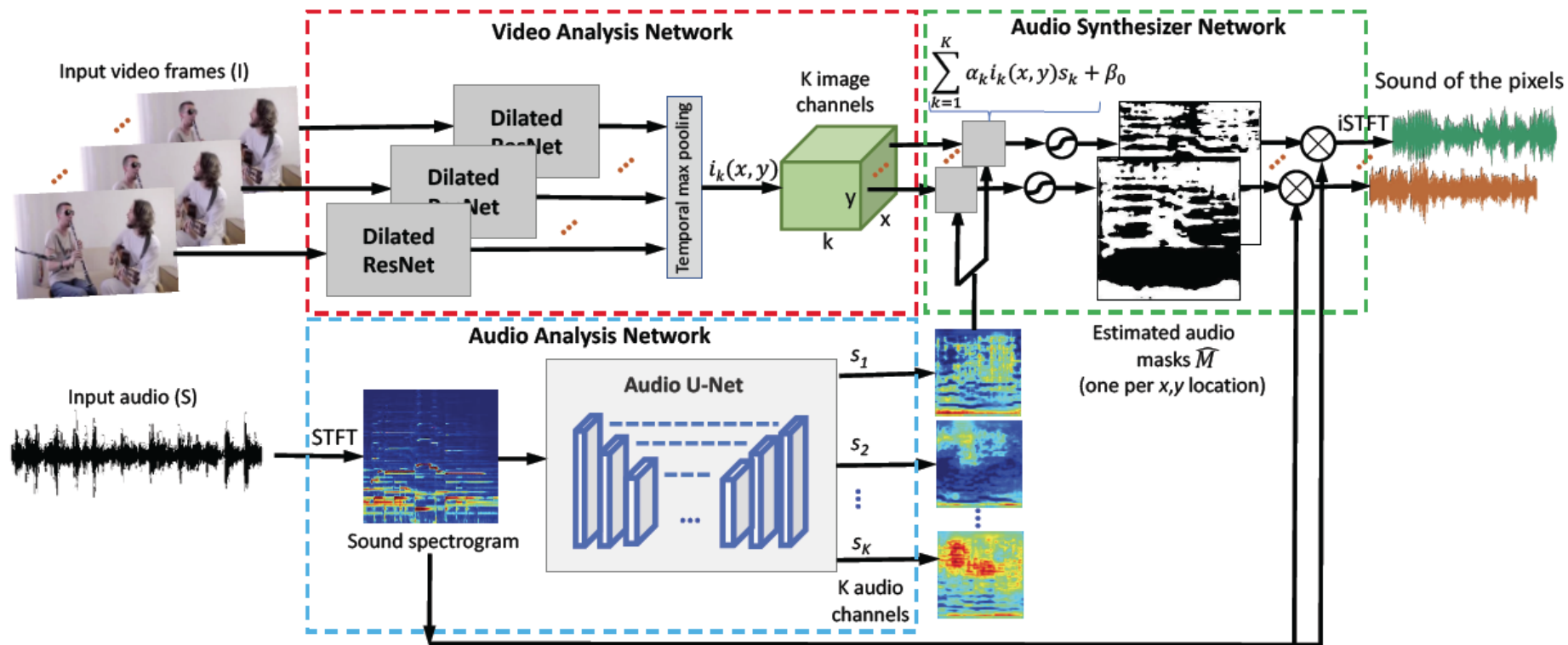
Training

Demos

http://sound-of-pixels.csail.mit.edu/
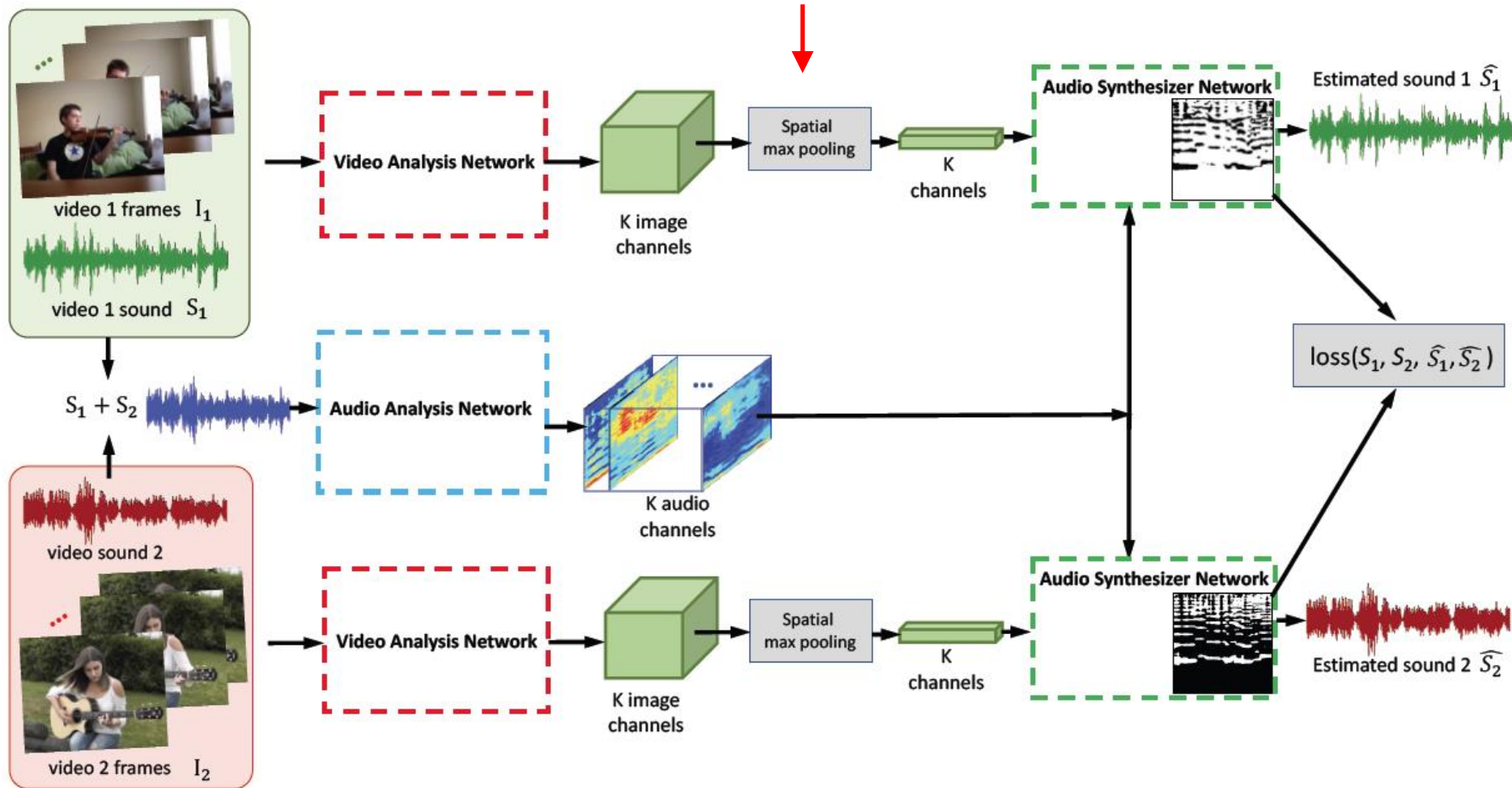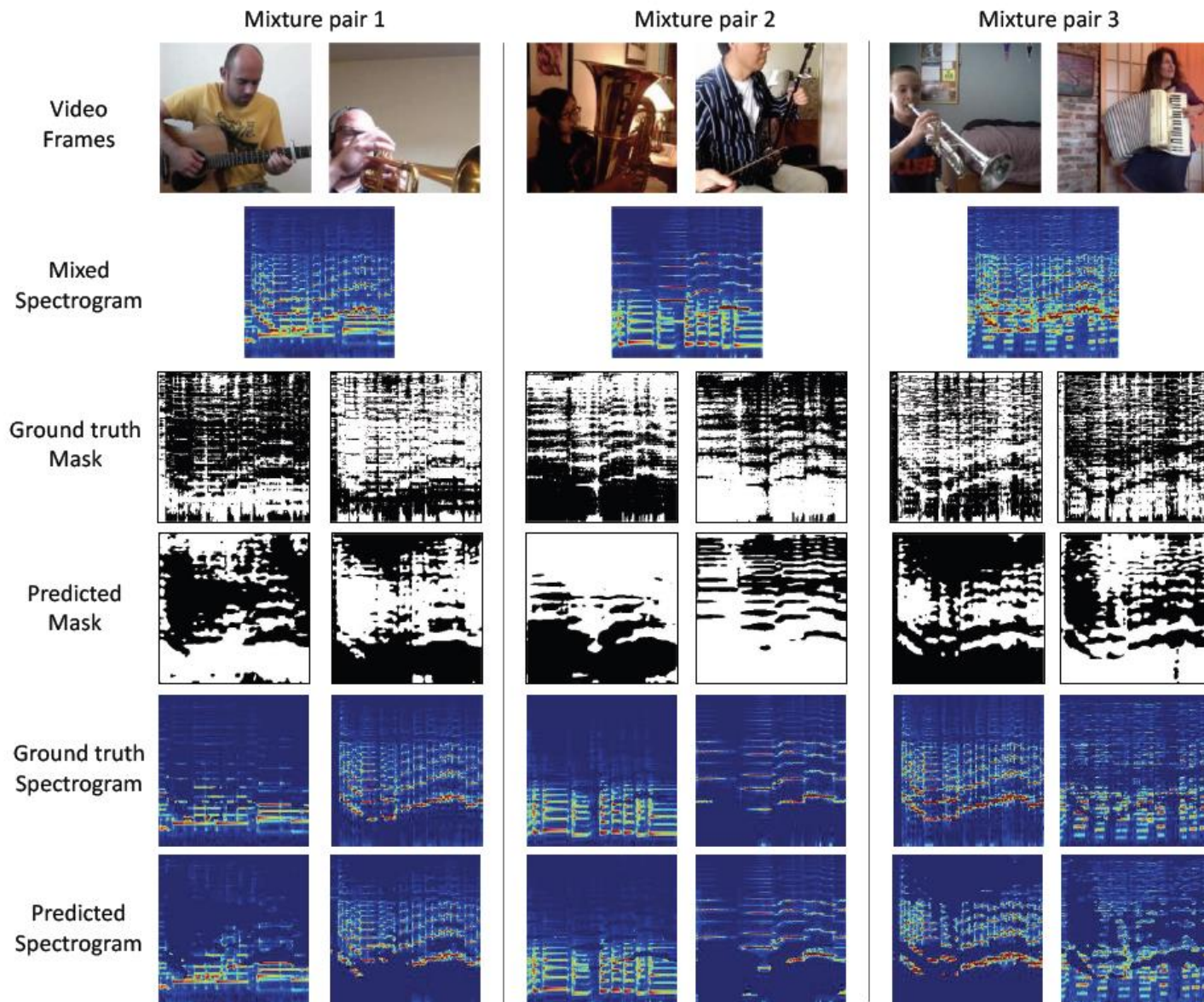
Visually Informed Source Separation

Zhao et al., "The sound of motions", 2019.
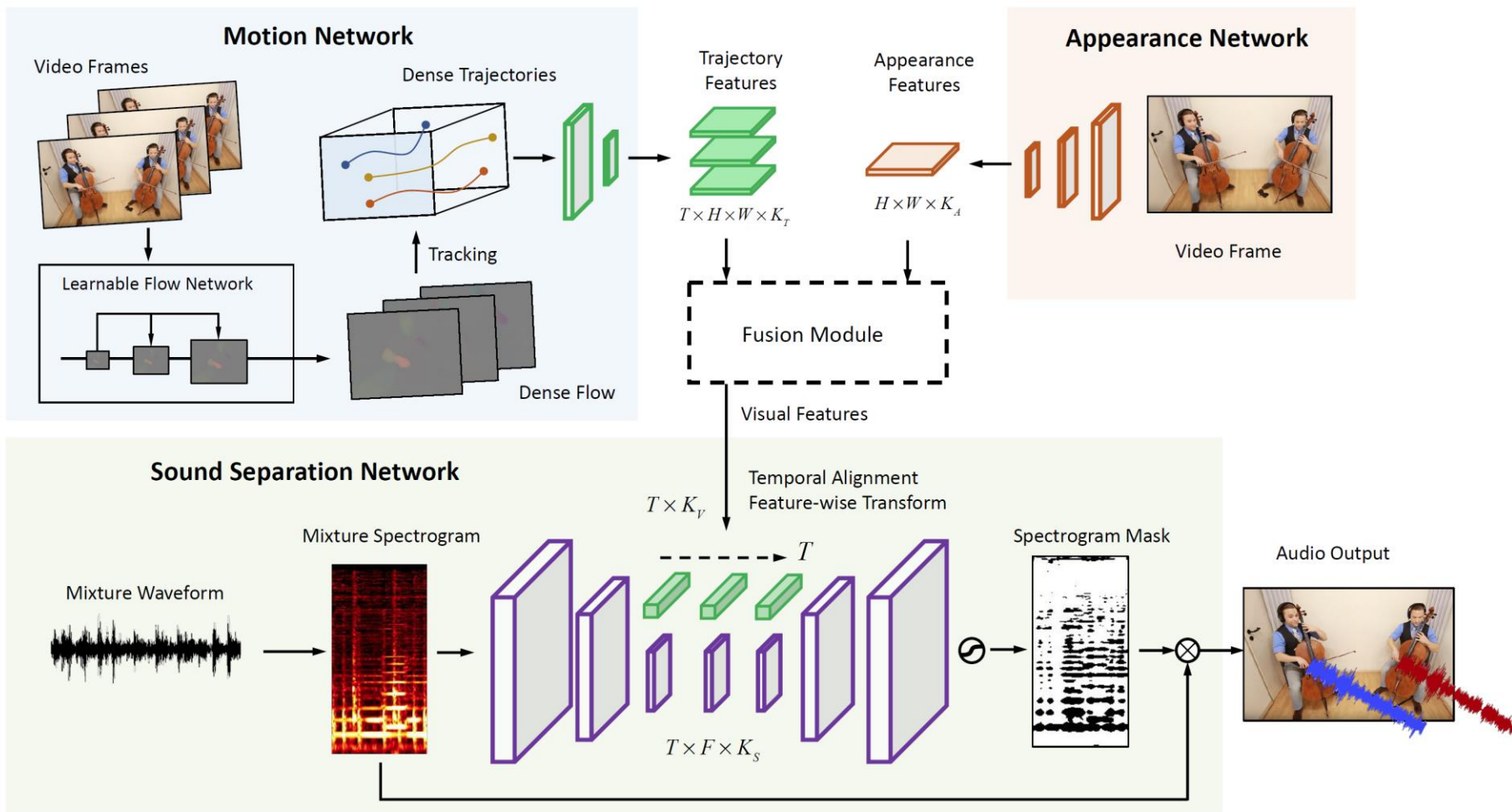
AV Speech Separation

Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," SIGGRAPH 2018.
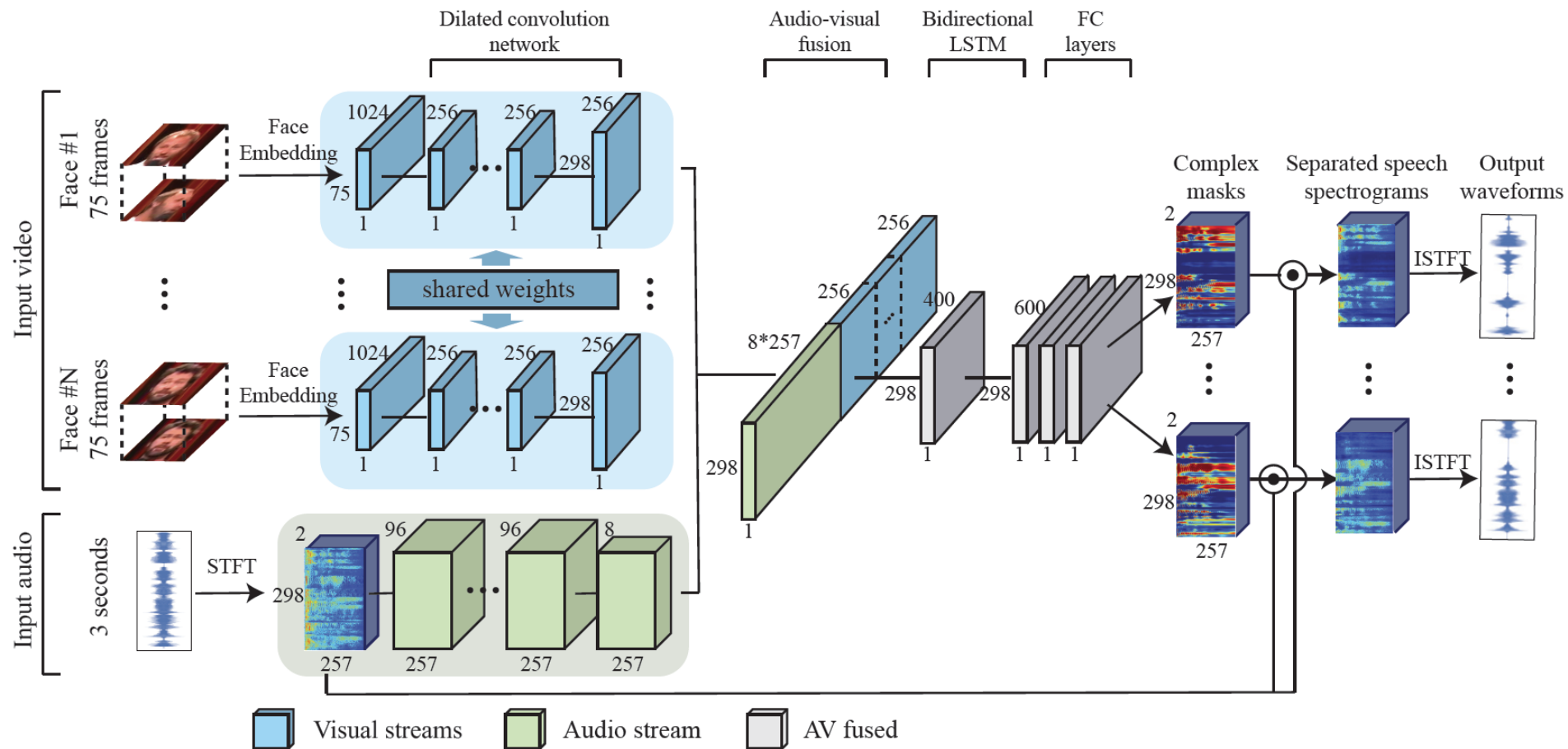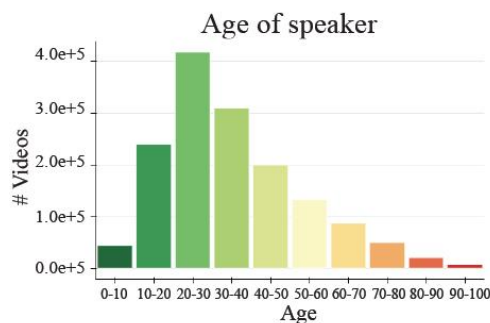
Training Data

4700 hours of video segments from 150,000 distinct speakers!!!

(a) Online videos of talks and lectures we collected

(b) Video segments with localized speakers and clean speech (which comprise our dataset)
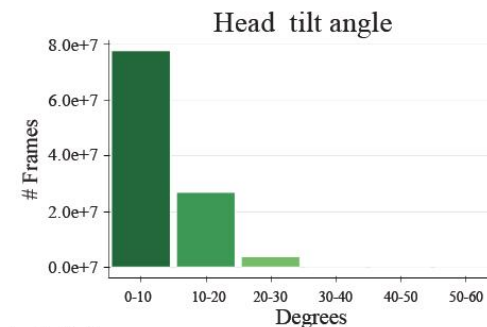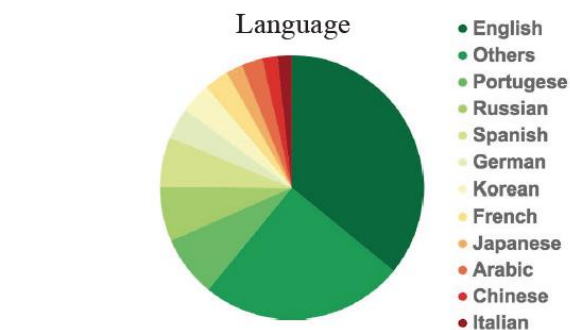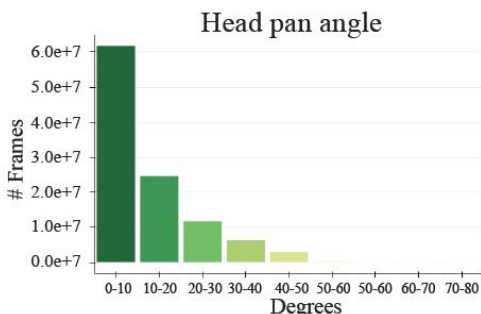
(c) Dataset statistics

Results and Demos

Mandarin (Enhancement)			
	Gabbay et al. [2017]	Hou et al. [2018]	Ours
PESQ	2.25	2.42	**2.5**
STOI	-	0.66	**0.71**
SDR	-	2.8	**6.1**

TCD-TIMIT (Separation)		
	Gabbay et al. [2017]	Ours
SDR	0.4	**4.1**
PESQ	2.03	**2.42**

CUAVE (Separation)		
	Casanovas et al. [2010]	Pu et al. [2017]	Ours
SDR	7	6.2	**12.6**

https://ai.googleblog.com/2018/04/looking-to-listen-audio-visual-speech.html

Source Permutation Problem
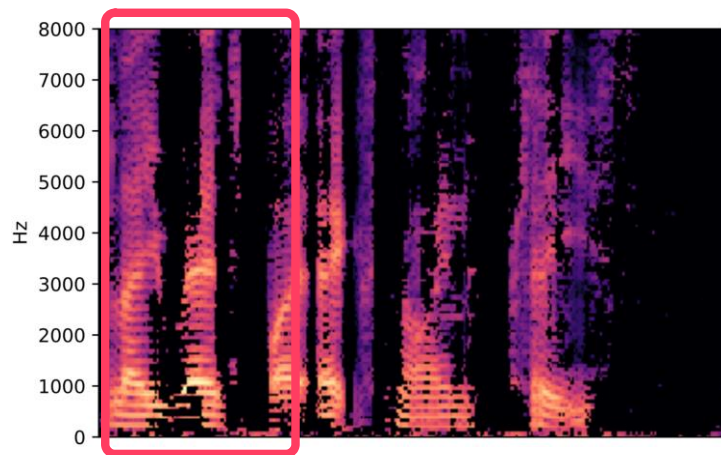
Ground-Truth1
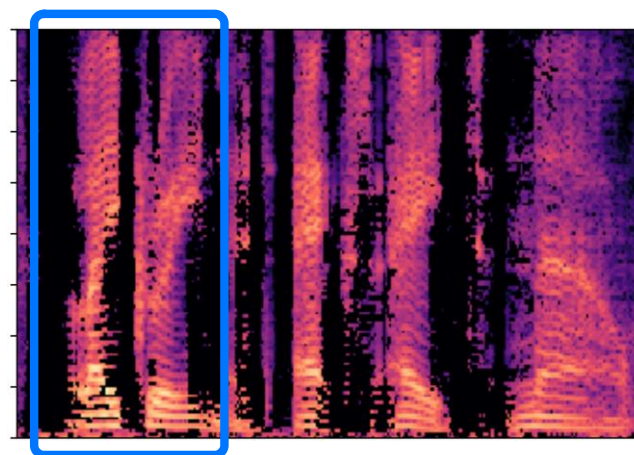
Ground-Truth2
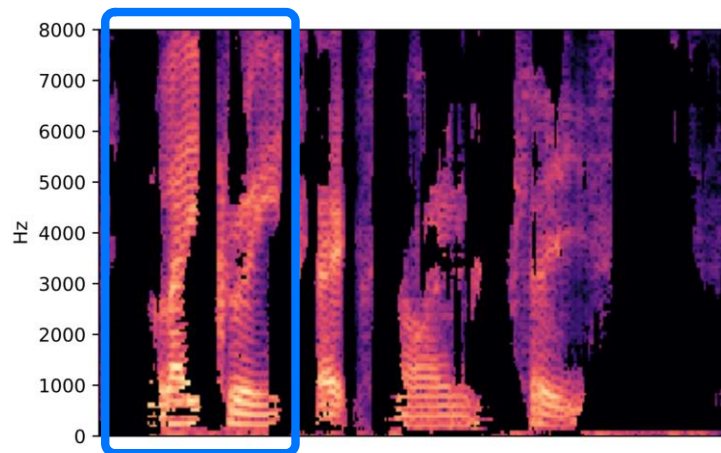
Deep Clustering1
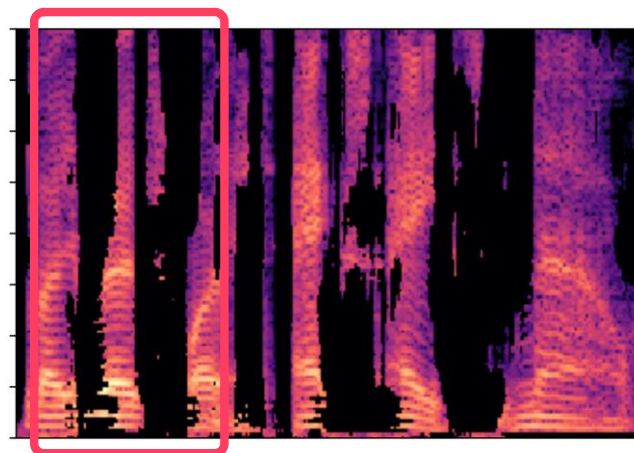
Deep Clustering2

Proposed Approach

Lu et al., "Listen and look: Audio-visual matching assisted speech source separation", SPL 2018.

- Model audio-visual matching between sound fluctuation and lip movement
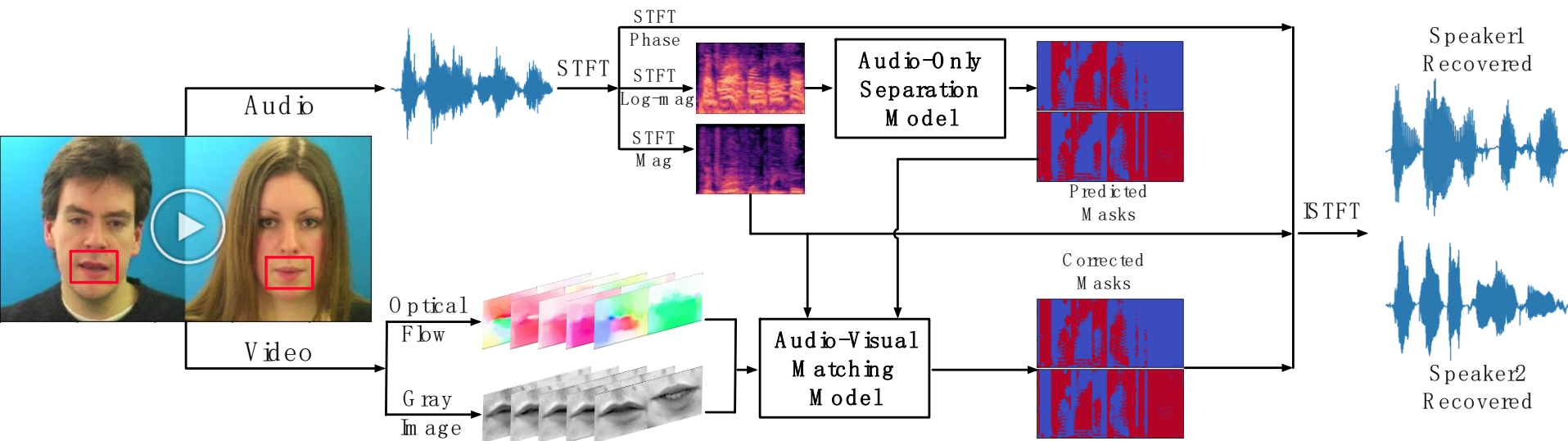
Audio-Visual Matching Network
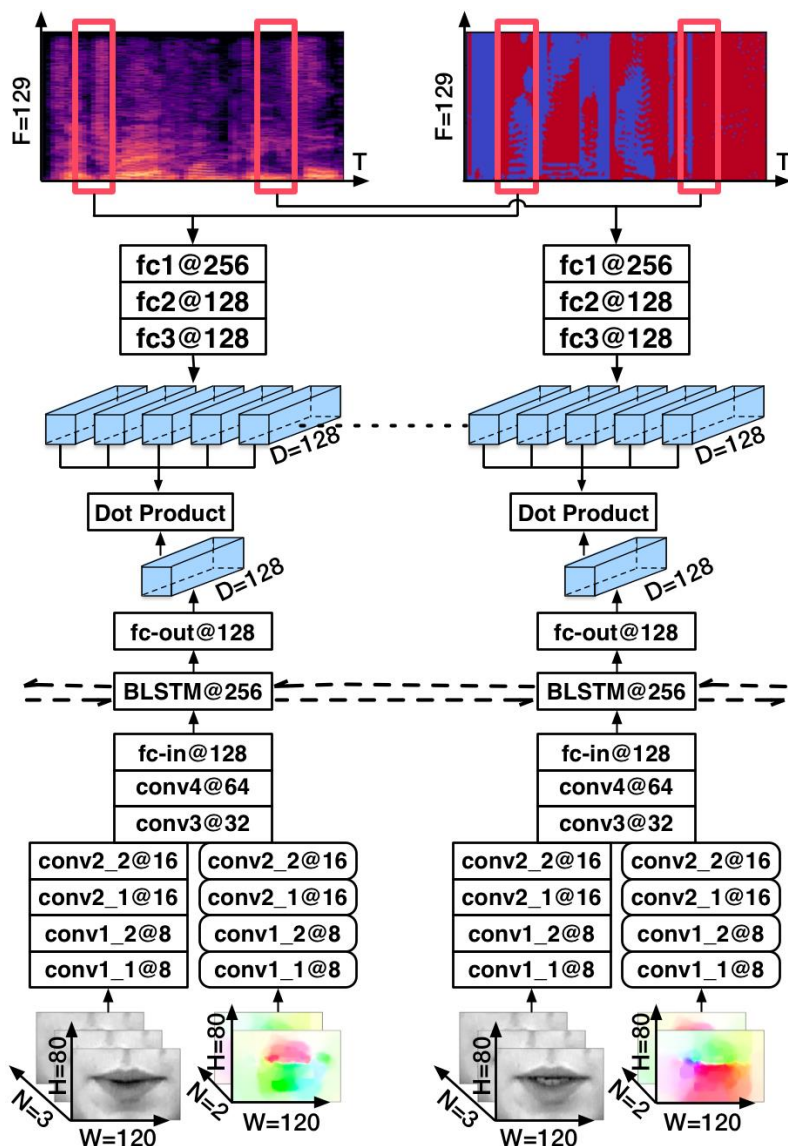
- Audio Network
  o MLP: fc1 ~ fc3

- Visual Network
  o Optical flow + gray image
  o Separate and merge
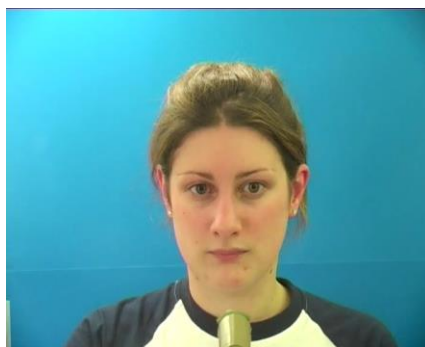 o CNN + LSTM

- Embedding
  o Similarity measure

Challenging Examples

Challenging Examples

Ground-truth

Deep Clustering
(audio-only)

Proposed
(audio-visual)

Challenging Examples

Challenging Examples

Ground-truth
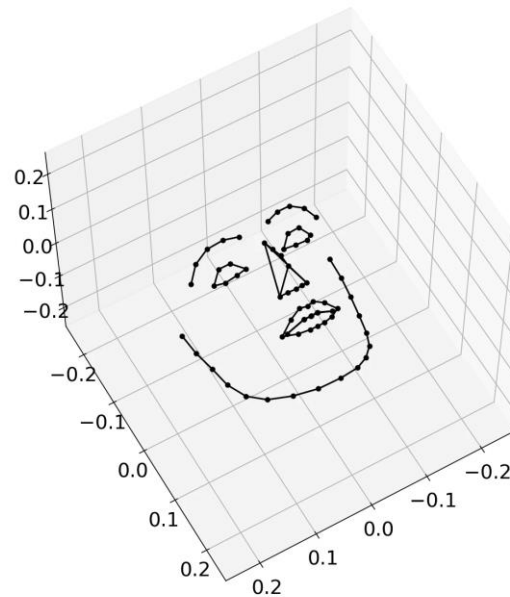
Deep Clustering
(audio-only)

Proposed
(audio-visual)

Talking Face Generation

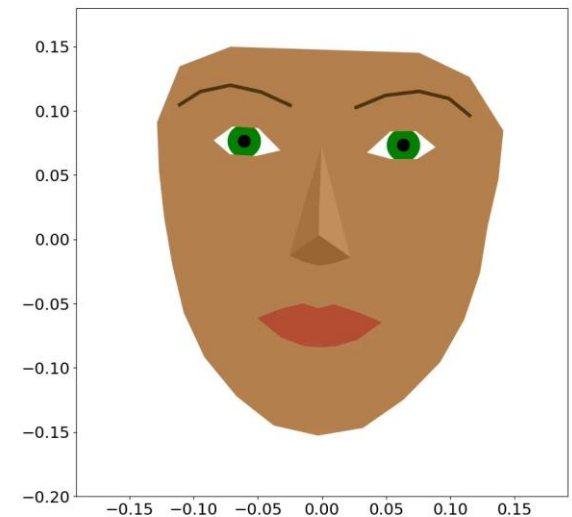□ Real-time face landmark generation from unheard speech of unknown speakers

[Eskimez et al, LVA/ICA'18]
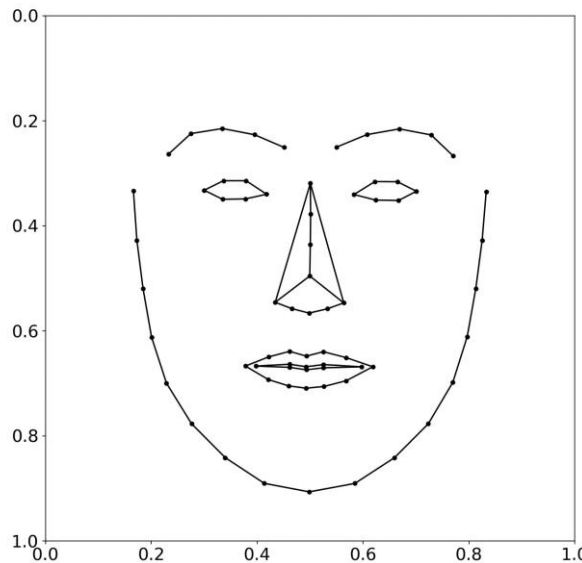
LSTM predicting 2D landmarks

1-D CNN predicting PCA coefficients of 3D landmarks
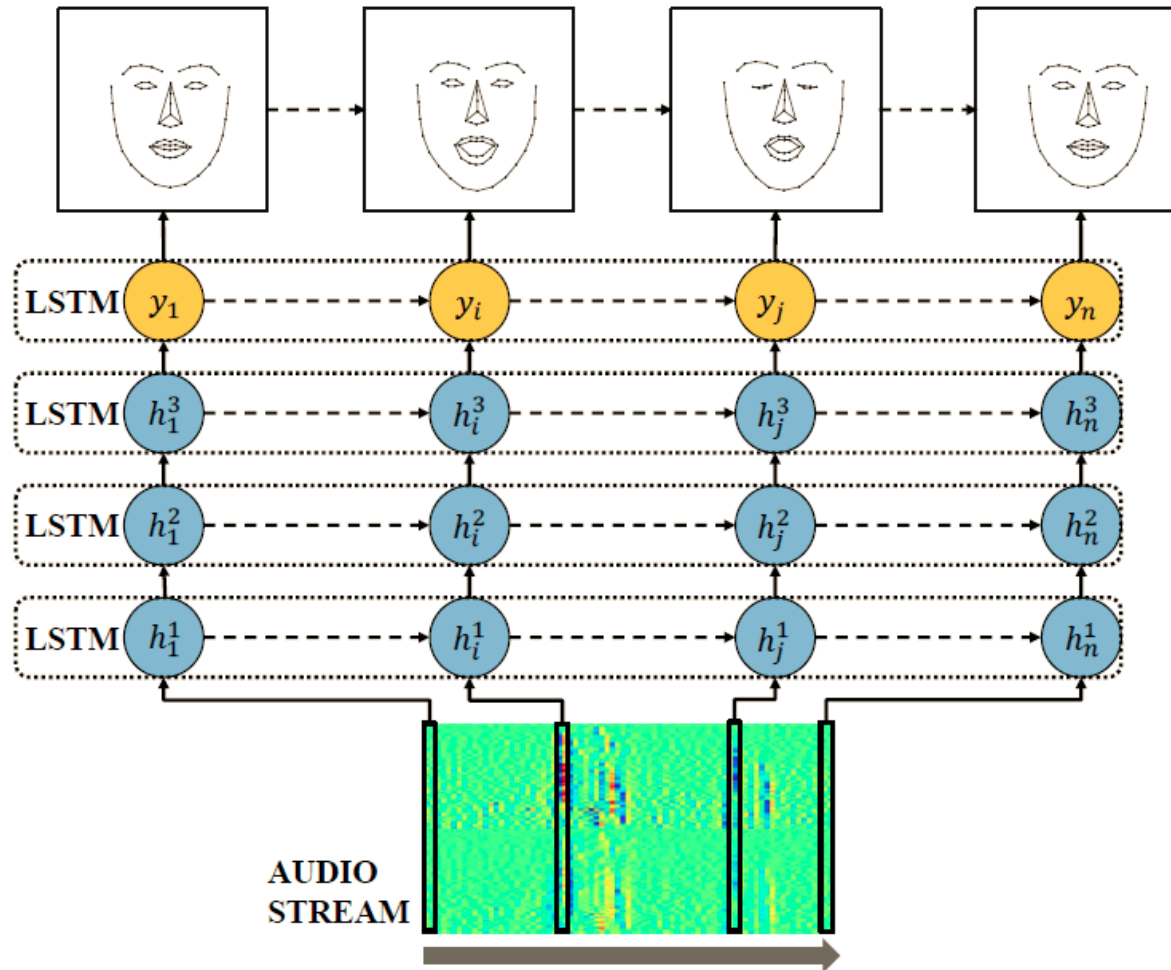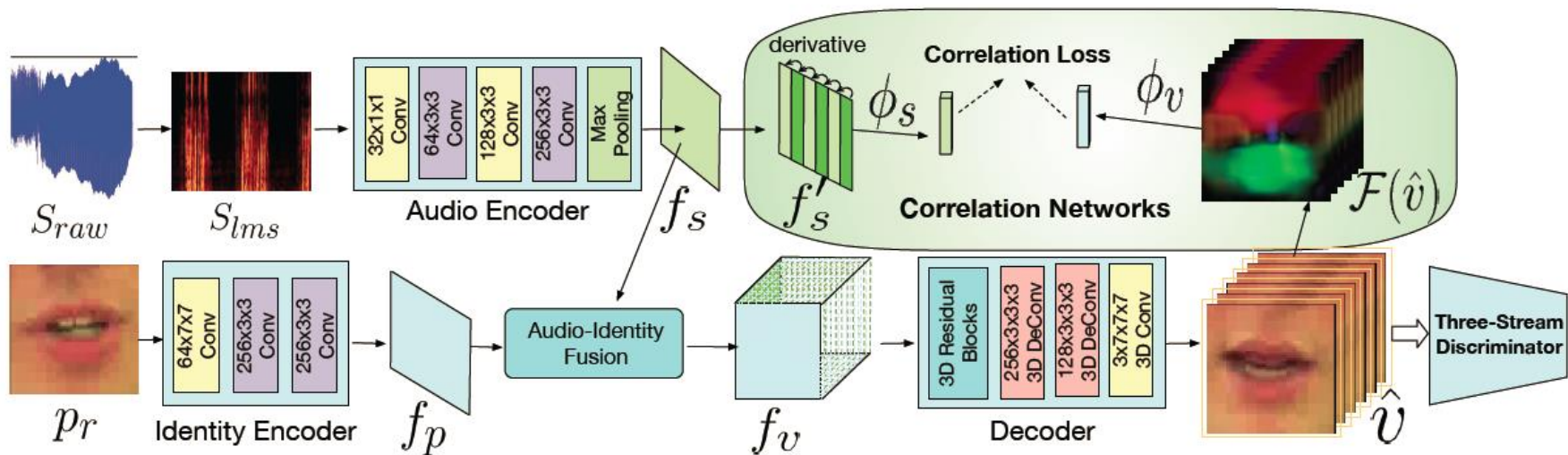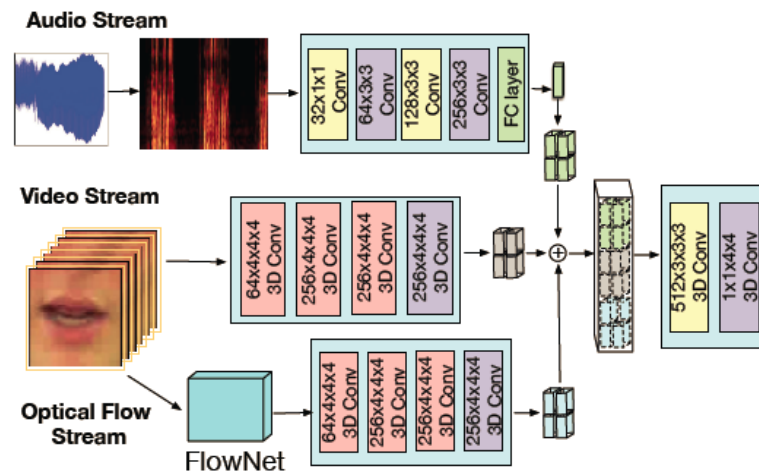
Network Structure

Photo-Realistic Generation

Fig. 5: Three-stream GAN discriminator illustration

https://www.youtube.com/watch?v=7IX_sIL5v0c