Tutorial:

One Perceptron to Rule Them All

ACM ICMR 2020

Dublin, Ireland, 26-29 October 2020

Part IV: Sound & Vision

Xavier Giro-i-Nieto

@DocXavi
xavier.giro@upc.edu

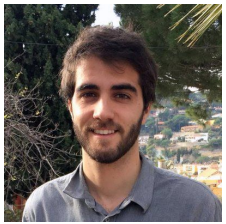
Associate Professor
Universitat Politècnica de Catalunya
Barcelona Supercomputing Center

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
IDEAI
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Acknowledgments

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Amanda Duarte



Dídac Surís

Amaia Salvador

Margarita Geleta

Cristina Puntí

Universitat Pompeu Fabra Barcelona

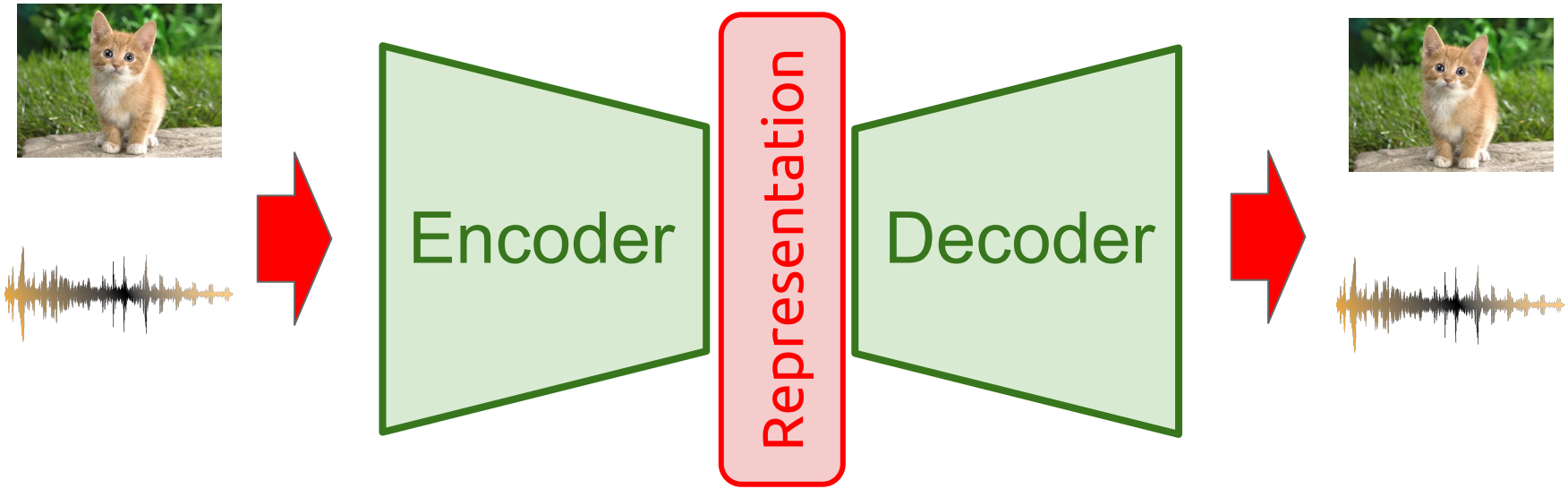
MTG Music Technology Group

DOLBY

Jordi Pons

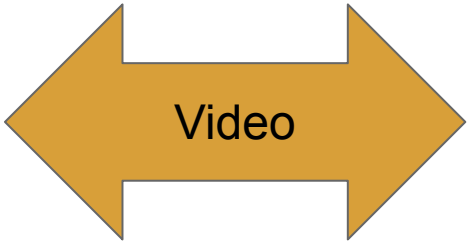
Outline

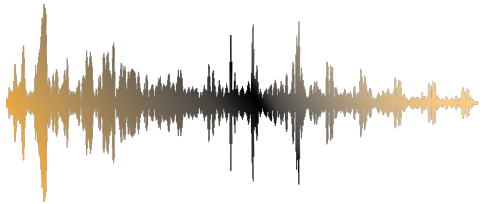
1. **Motivation**
2. Feature Learning
3. Cross-modal Translation
4. Embodied AI


Encoder
Representation
Decoder

Self-supervised Learning

Vision

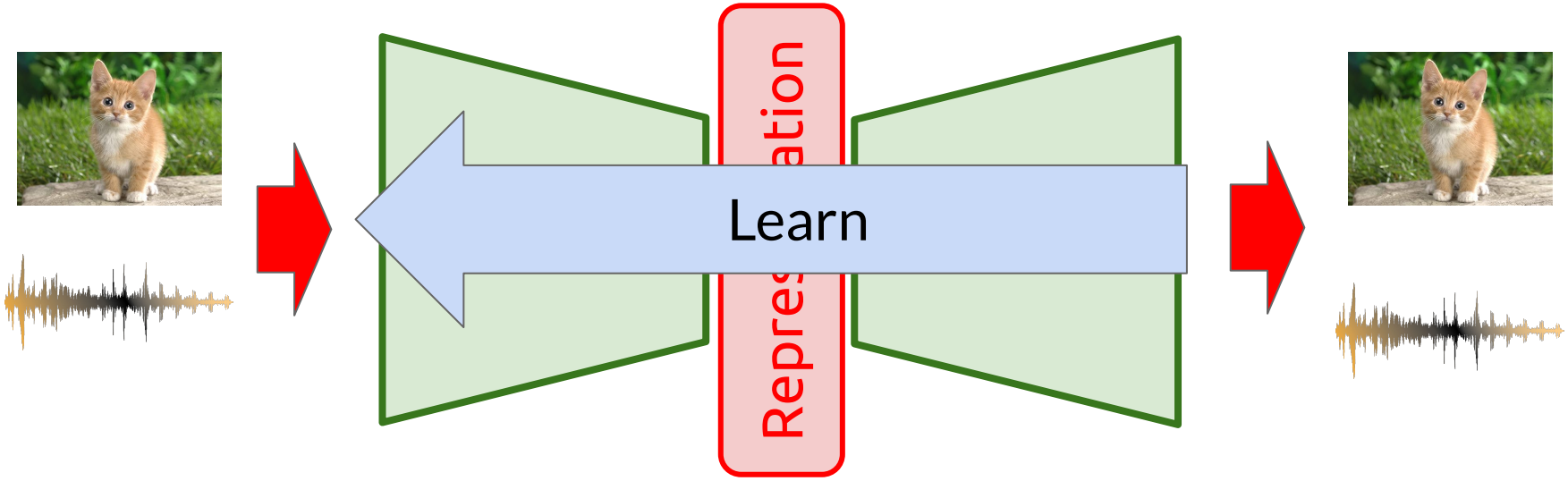




Audio

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

Self-supervised Learning



Outline

1. Motivation
2. **Feature Learning**
3. Cross-modal Translation

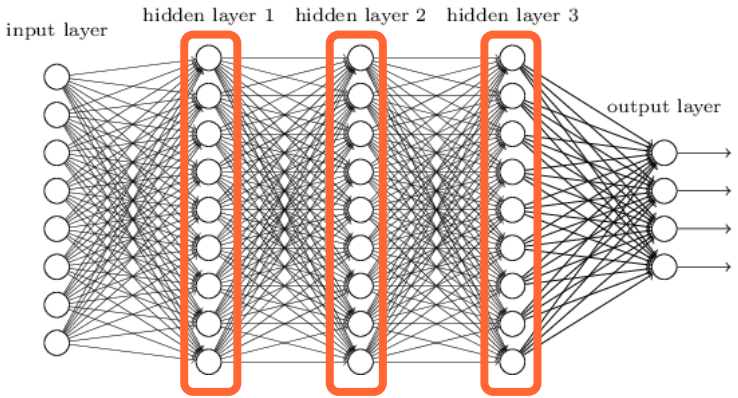
Self-supervised Feature Learning

Self-supervised feature learning is a form of unsupervised learning where the raw data provides the supervision.

- A **pretext (or surrogate) task** must be designed.
- By defining a proxy loss, the NN **learns representations**, which should be valuable for the actual downstream task.

Unlabeled data
(X)

Representations learned without labels



Reference: Andrew Zisserman (PAISS 2018)

Self-supervised Feature Learning

Generative / Predictive

Loss measured in the output space

Contrastive

Loss measured in the representation space

Source: Ankesh Anand, "Contrastive Self-Supervised Learning" (2020)"

Outline

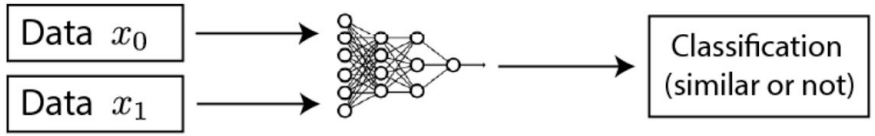
1. Motivation
2. Feature Learning
    a. **Generative / Predictive Methods**
    b. Contrastive Methods
3. Cross-modal Translation
4. Embodied AI

Self-supervised Feature Learning

Generative / Predictive

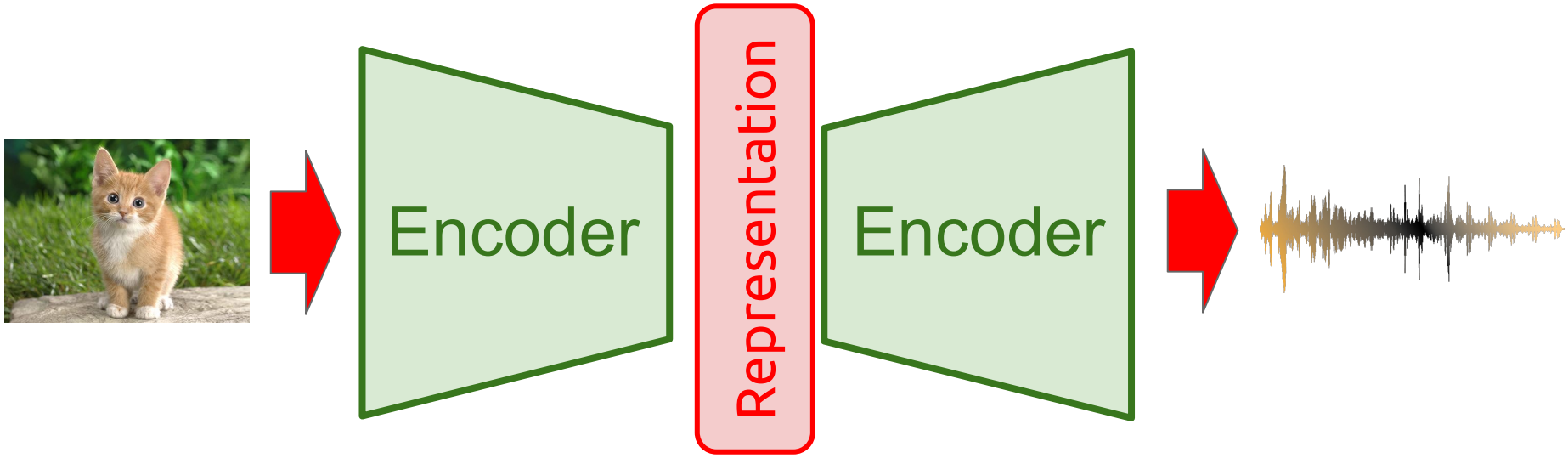
Loss measured in the output space

Contrastive



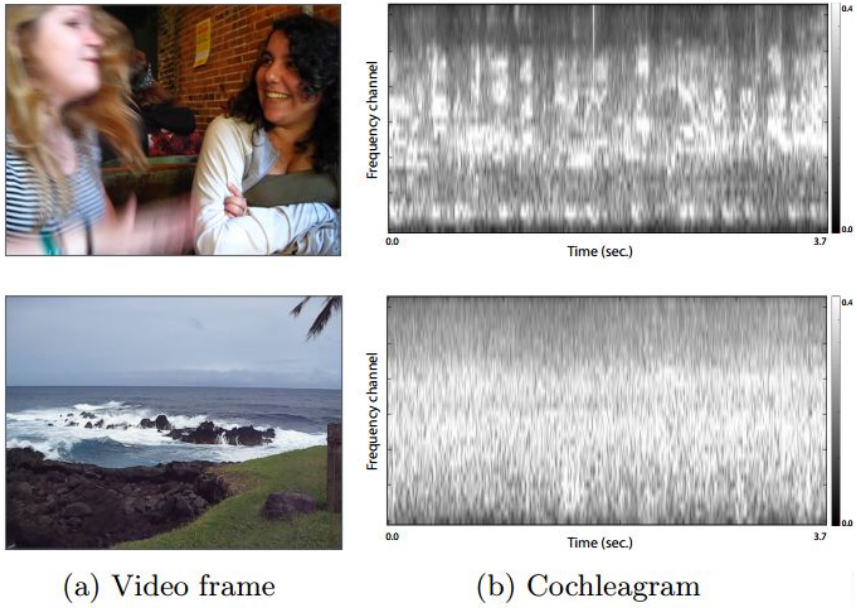
Loss measured in the representation space

Source: Ankesh Anand, "Contrastive Self-Supervised Learning" (2020)"


Encoder
Representation
Encoder

Prediction of Audio Features (stats)

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

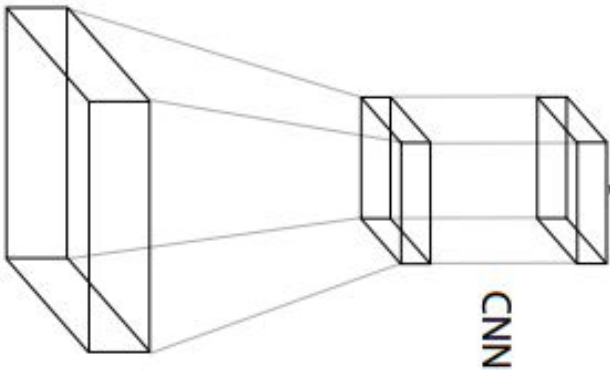


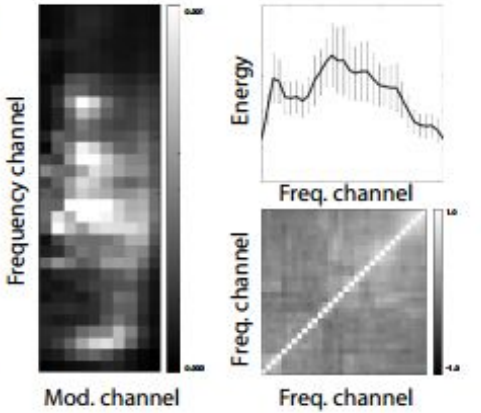
(a) Video frame (b) Cochleagram

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

Prediction of Audio Features (stats)

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



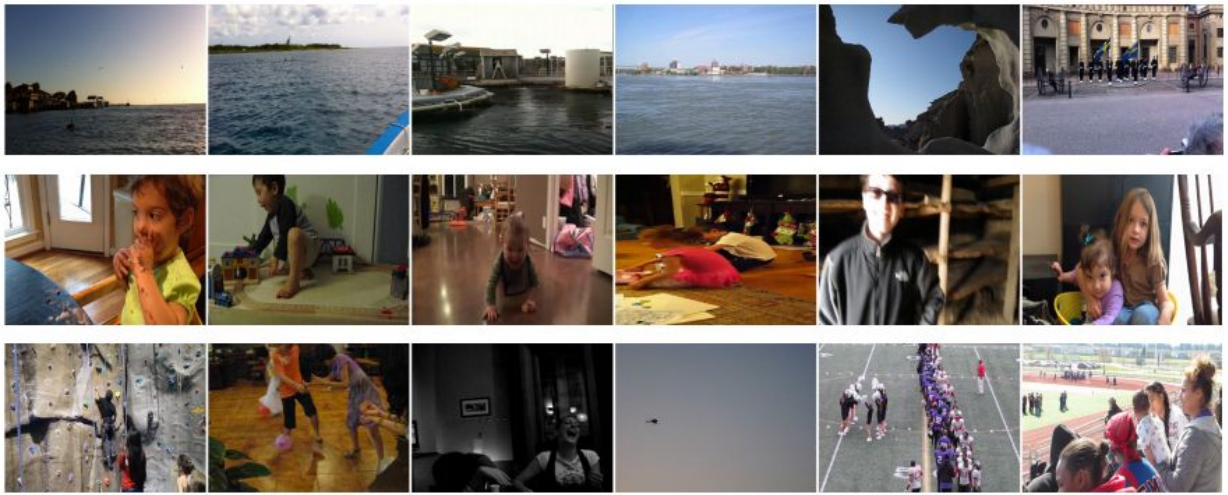


Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

Prediction of Audio Features (stats)

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

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



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

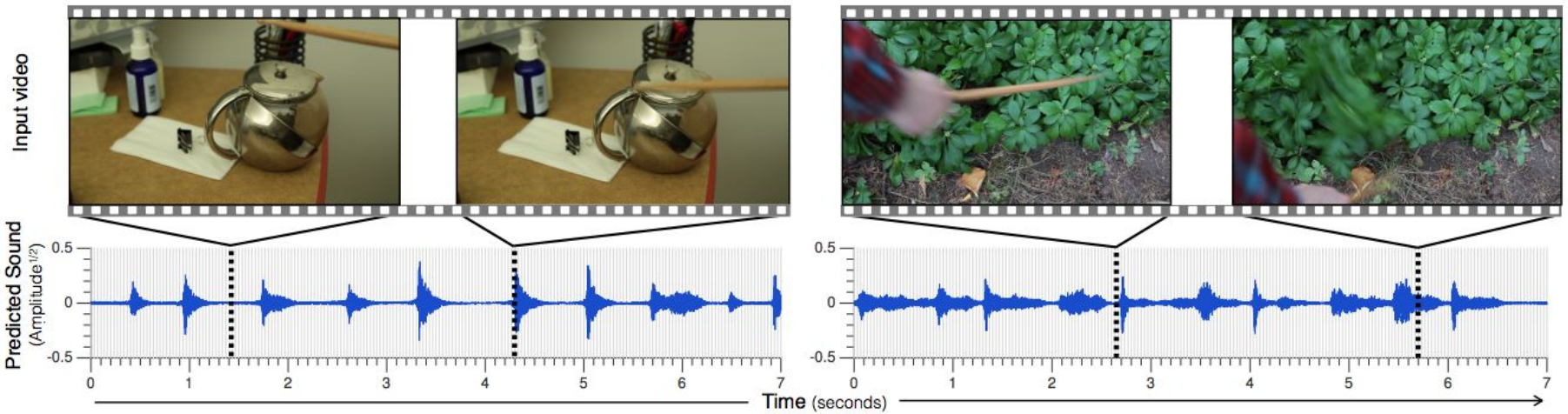
Prediction of Audio Features (stats)

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

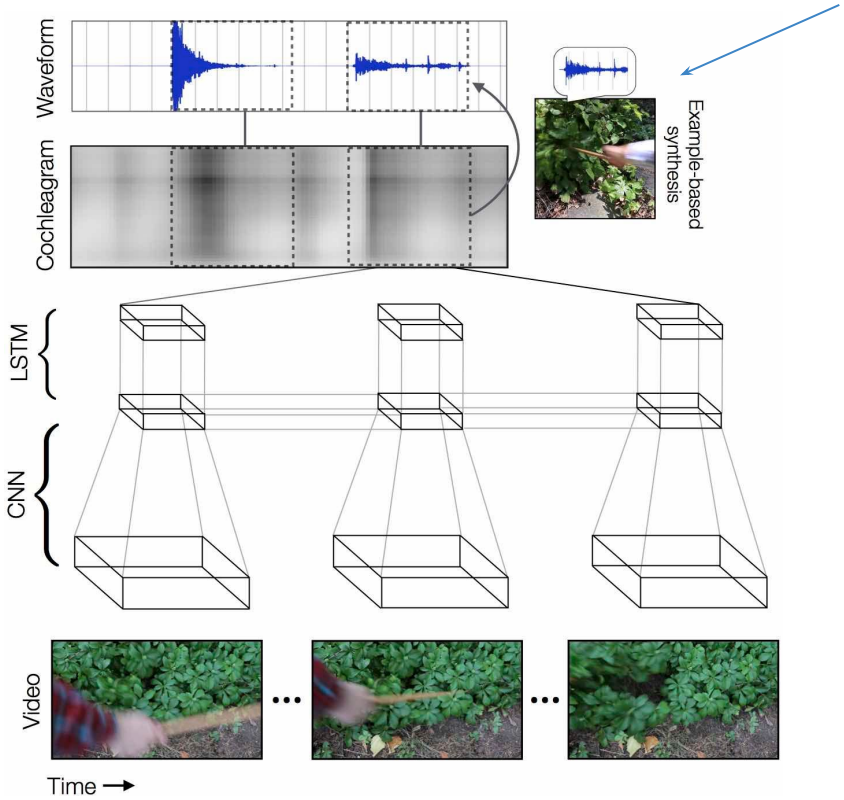
Prediction of Audio Features (cochleagram)

Video sonorization: Retrieve matching sounds for videos of people hitting objects with a drumstick.



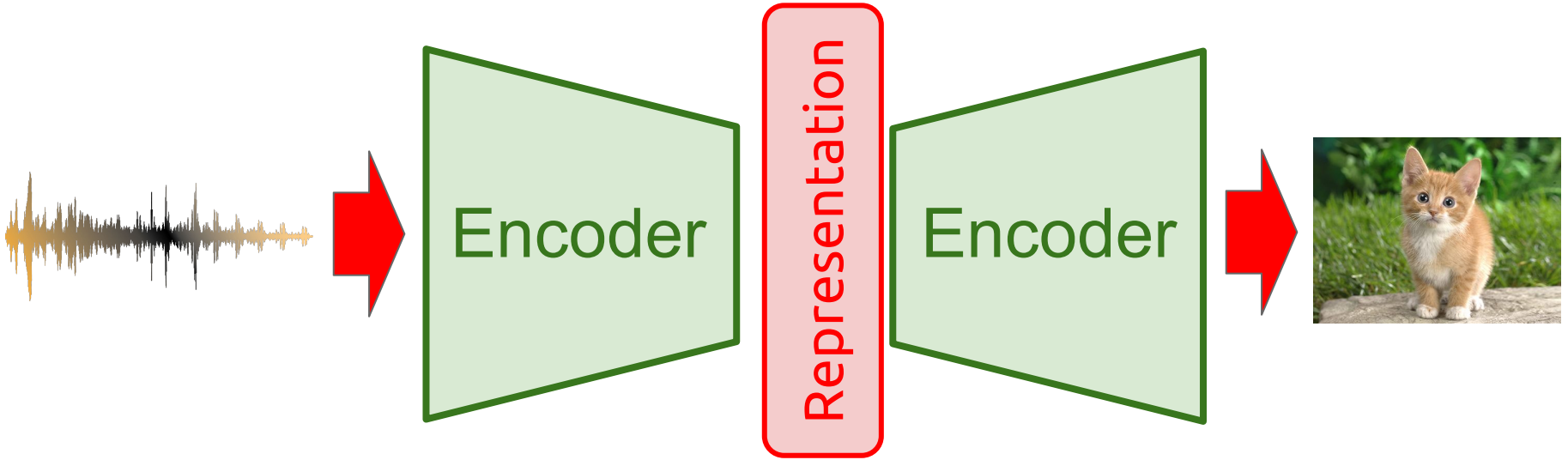
Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Prediction of Audio Features (cochleagram)



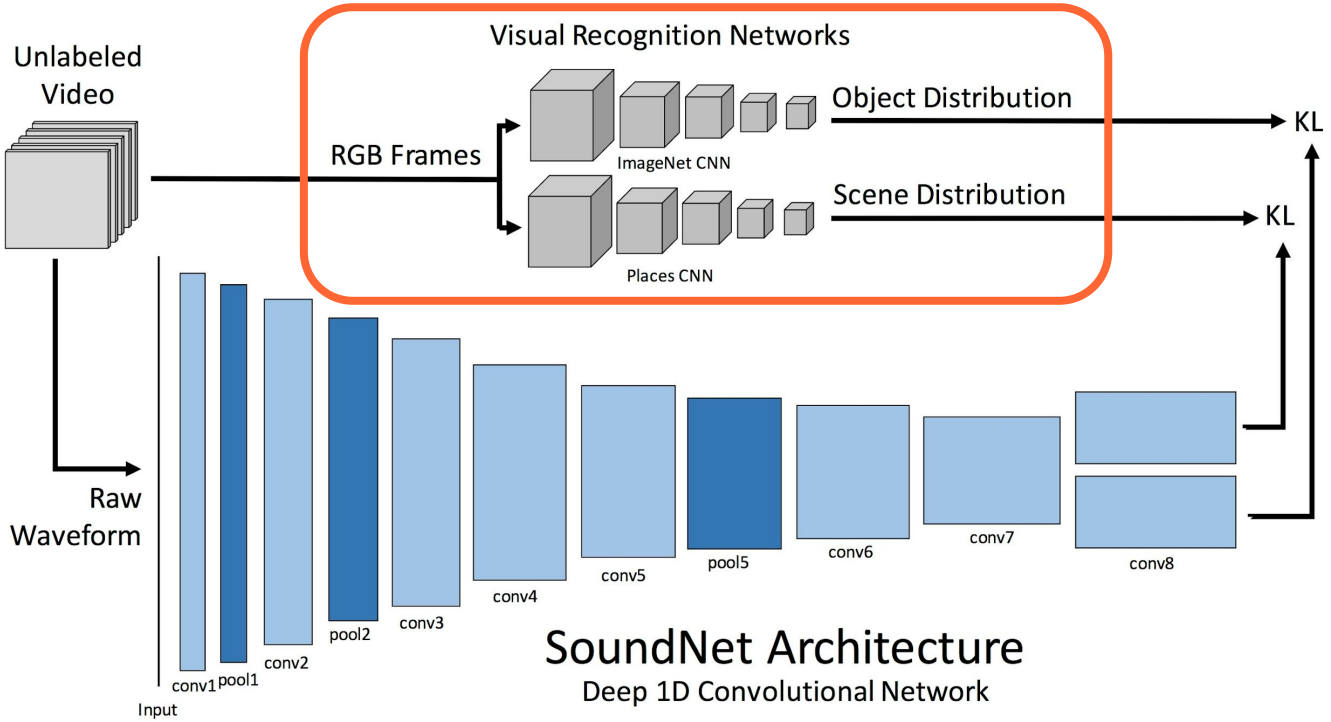
Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

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


Encoder
Representation
Encoder

Prediction of Image Labels (distillation)

Teacher network: Visual Recognition (object & scenes)



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

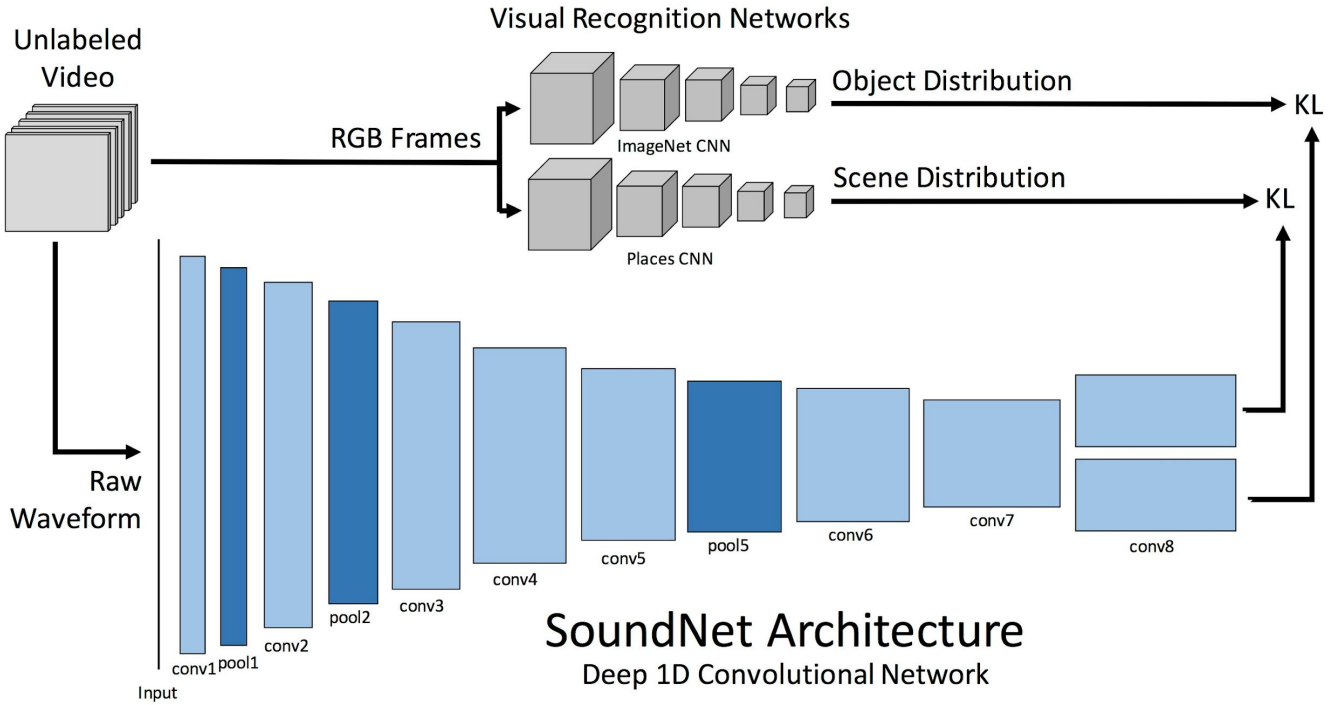
Predicted Objects and Scenes from Sound Only

(Videos are blurred so you can try to recognize yourself!)

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

Prediction of Image Labels (distillation)

Learned audio features are good for environmental sound recognition.



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

Prediction of Image Labels (distillation)

Learned audio features are good for environmental sound recognition.

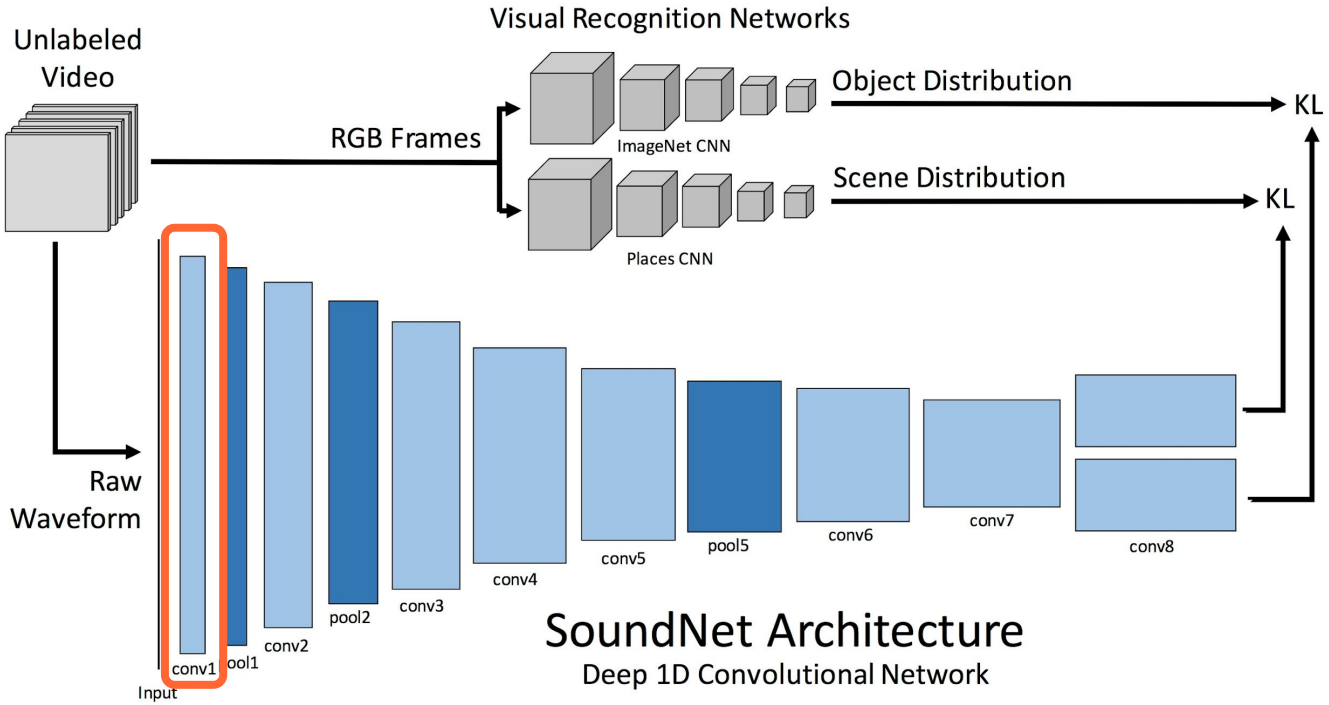
Method	Accuracy
RG [29]	69%
LTT [21]	72%
RNH [30]	77%
Ensemble [34]	78%
SoundNet	**88%**

Method	Accuracy on ESC-50	Accuracy on 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%

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

Prediction of Image Labels (distillation)

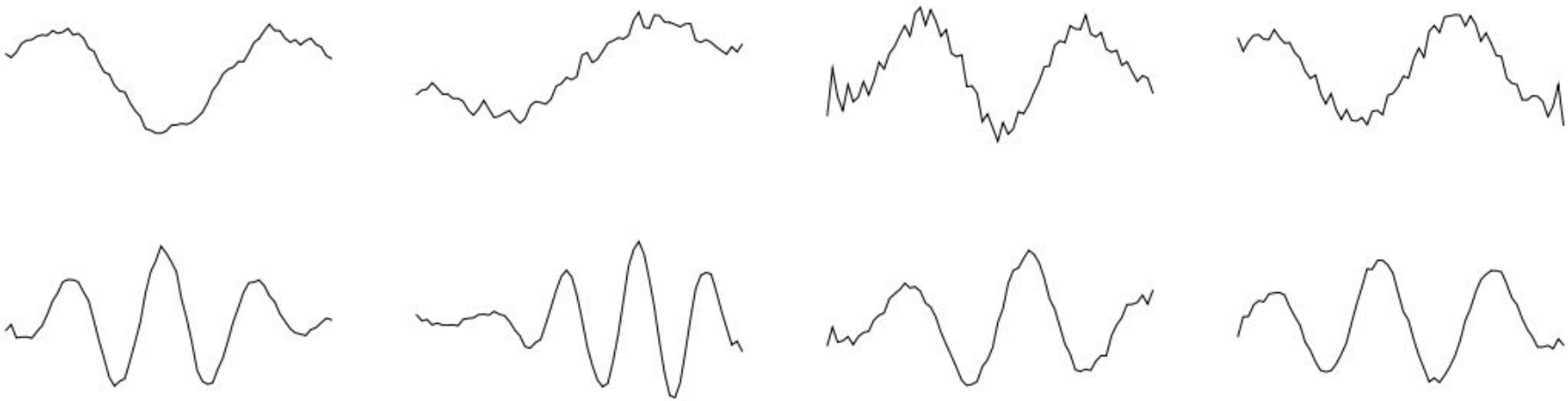
Visualization of the 1D filters over raw audio in conv1.



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

Prediction of Image Labels (distillation)

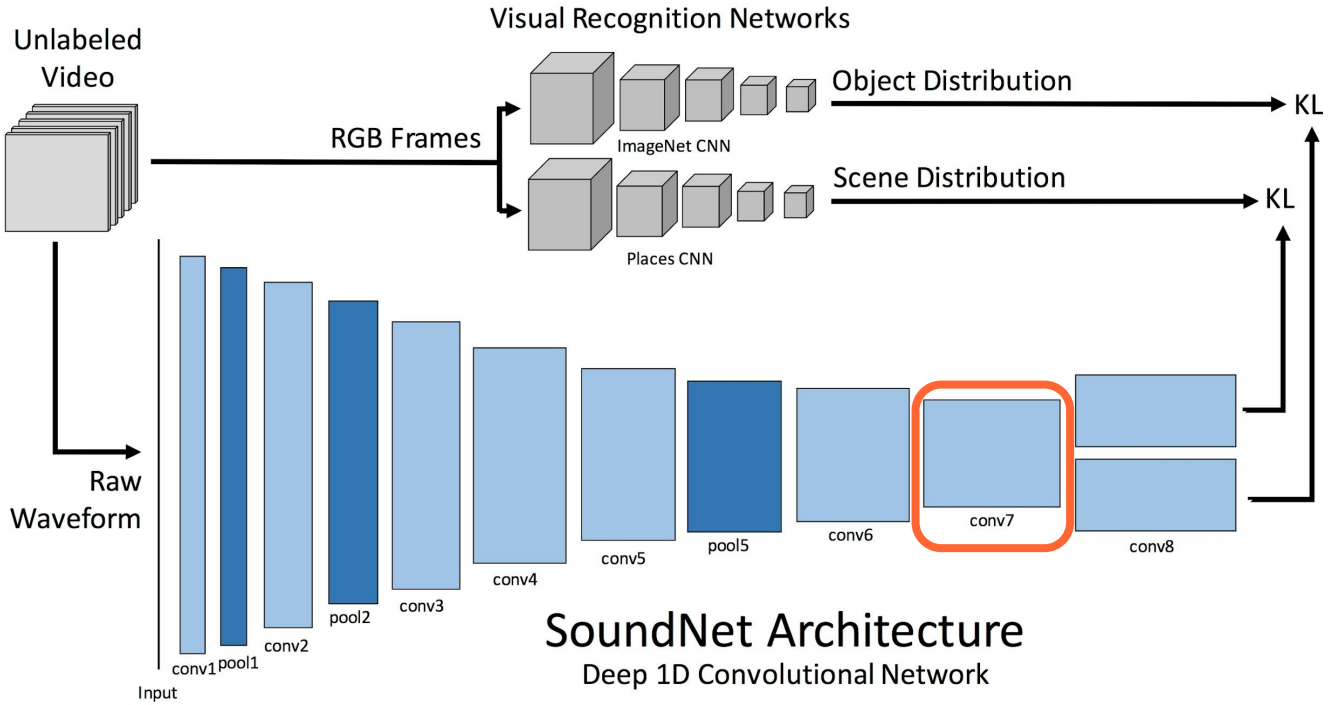
Visualization of the 1D filters over raw audio in conv1.



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

Prediction of Image Labels (distillation)

Visualize video frames 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.

Prediction of Image Labels (distillation)

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

Baby Talk

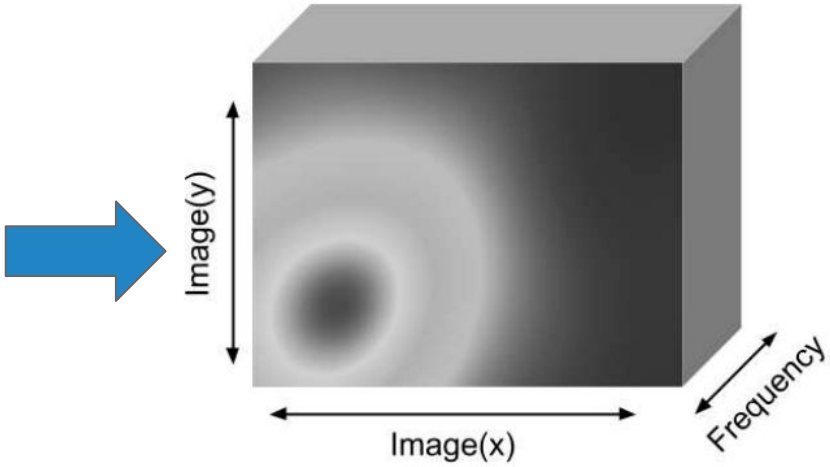
Bubbles

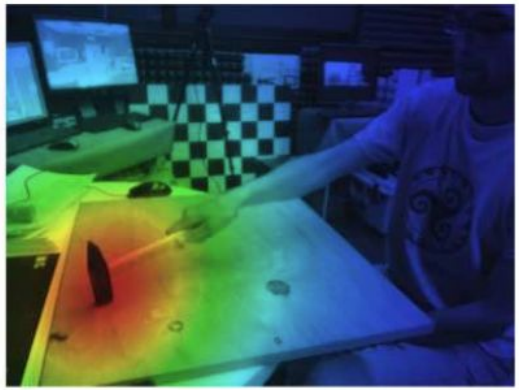
Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Prediction of Acoustic Images (distillation)

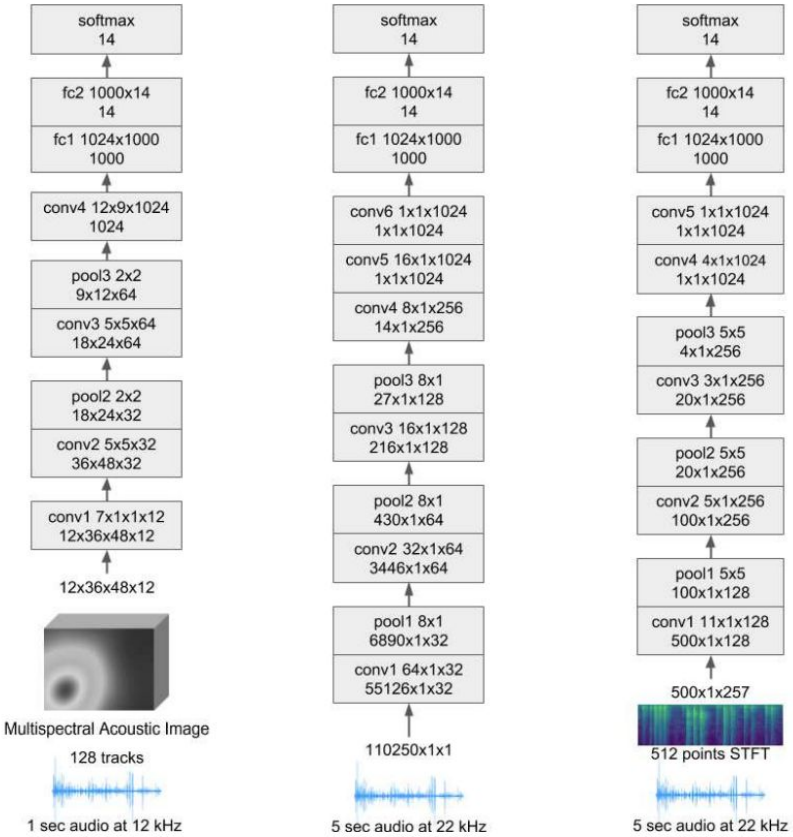
Acoustic images are aligned in space and synchronized in time during learning.

Acoustic-optical camera

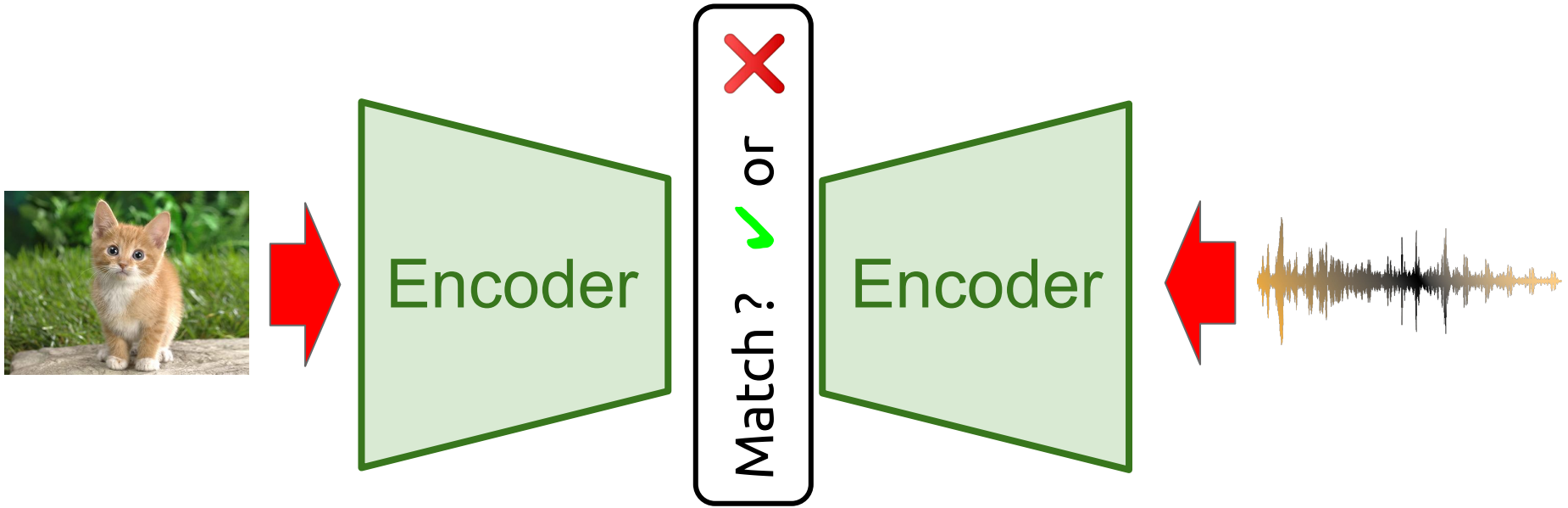




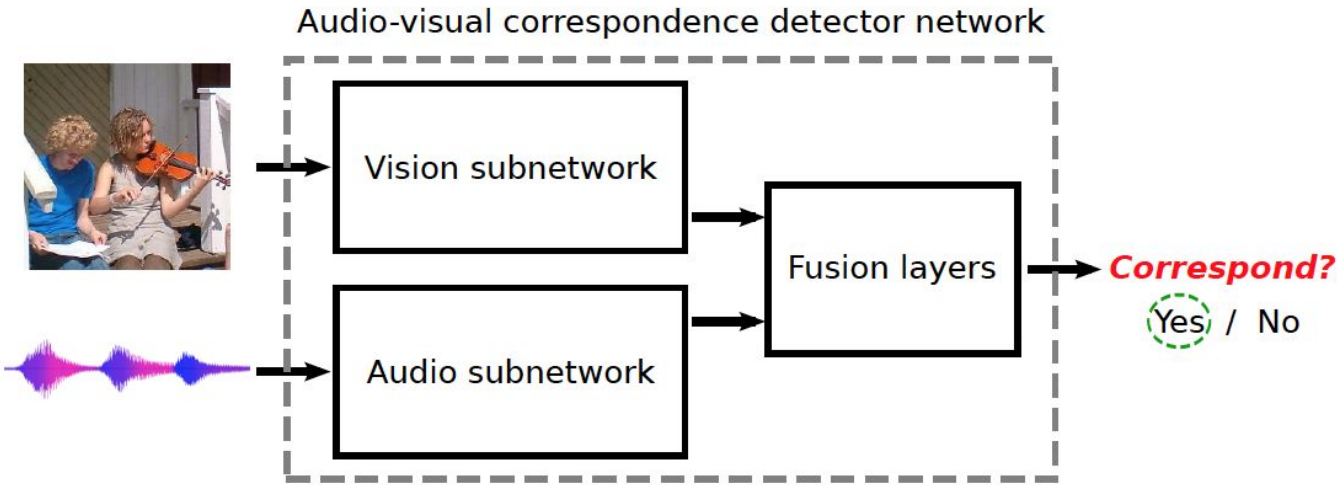
Perez, A., Sanguineti, V., Morerio, P., & Murino, V. (2020). Audio-visual model distillation using acoustic images. WACV 2020. [code]

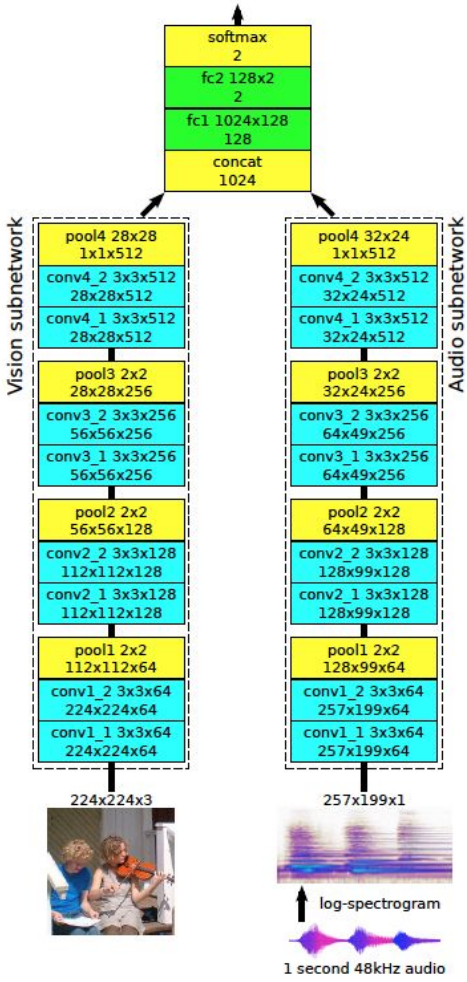


Perez, A., Sanguineti, V., Morerio, P., & Murino, V. (2020). Audio-visual model distillation using acoustic images. WACV 2020. [code]


Encoder
Match ? ✓ or ✗
Encoder

Binary Verification

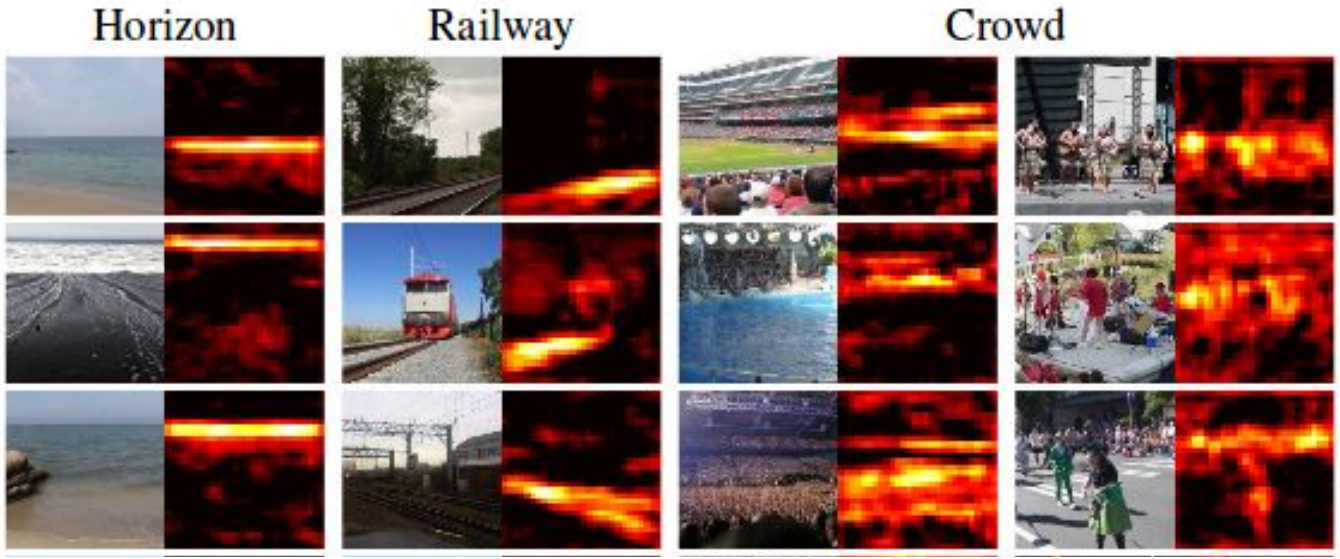


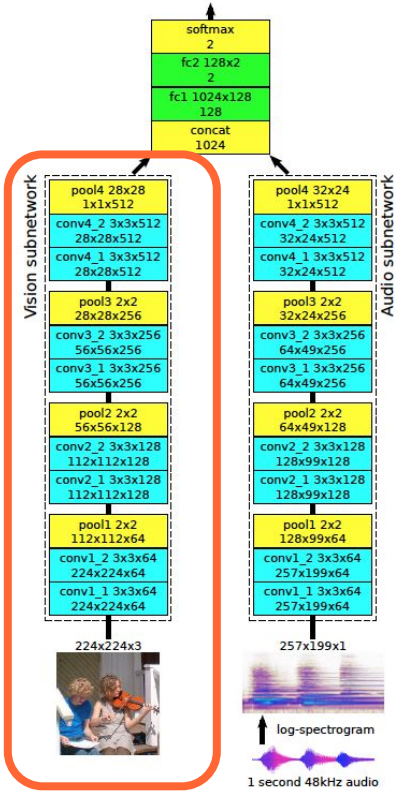


Arandjelović, Relja, and Andrew Zisserman. "Look, Listen and Learn." *ICCV 2017.*

Binary Verification

Most activated unit in *pool4* layer of the **visual network**.

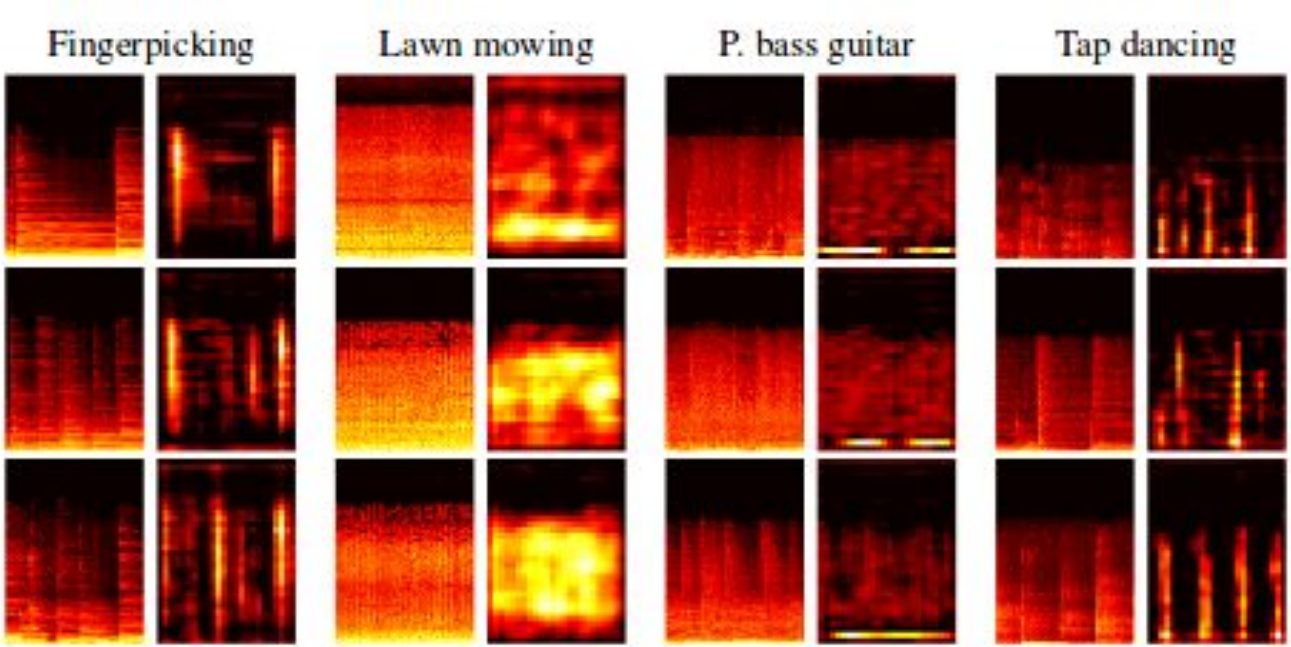


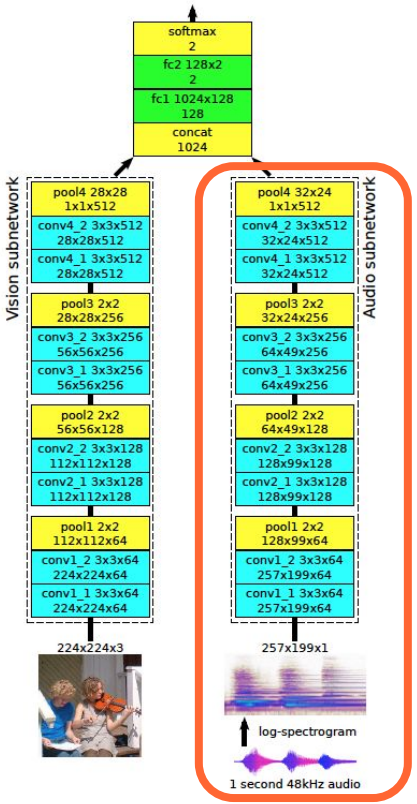


Arandjelović, Relja, and Andrew Zisserman. "Look, Listen and Learn." *ICCV 2017.*

Binary Verification

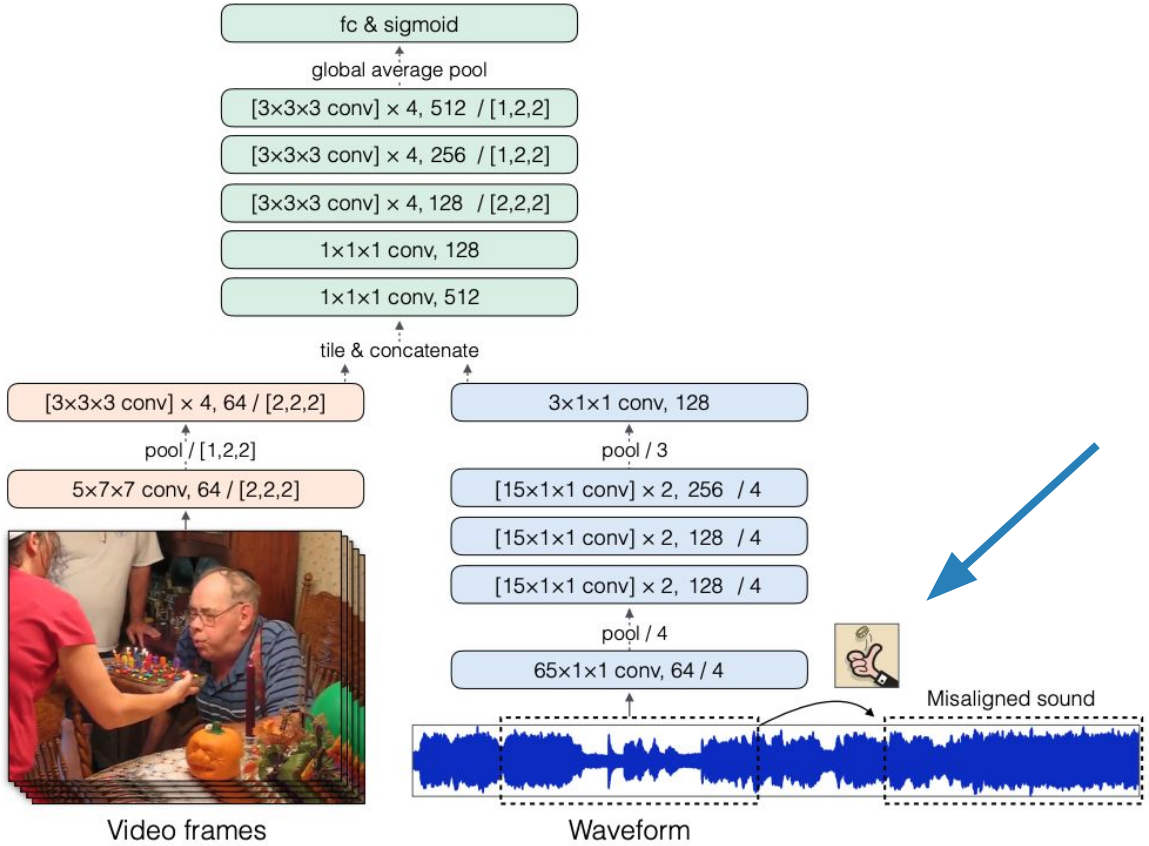
Most activated unit in *pool4* layer of the **audio network**





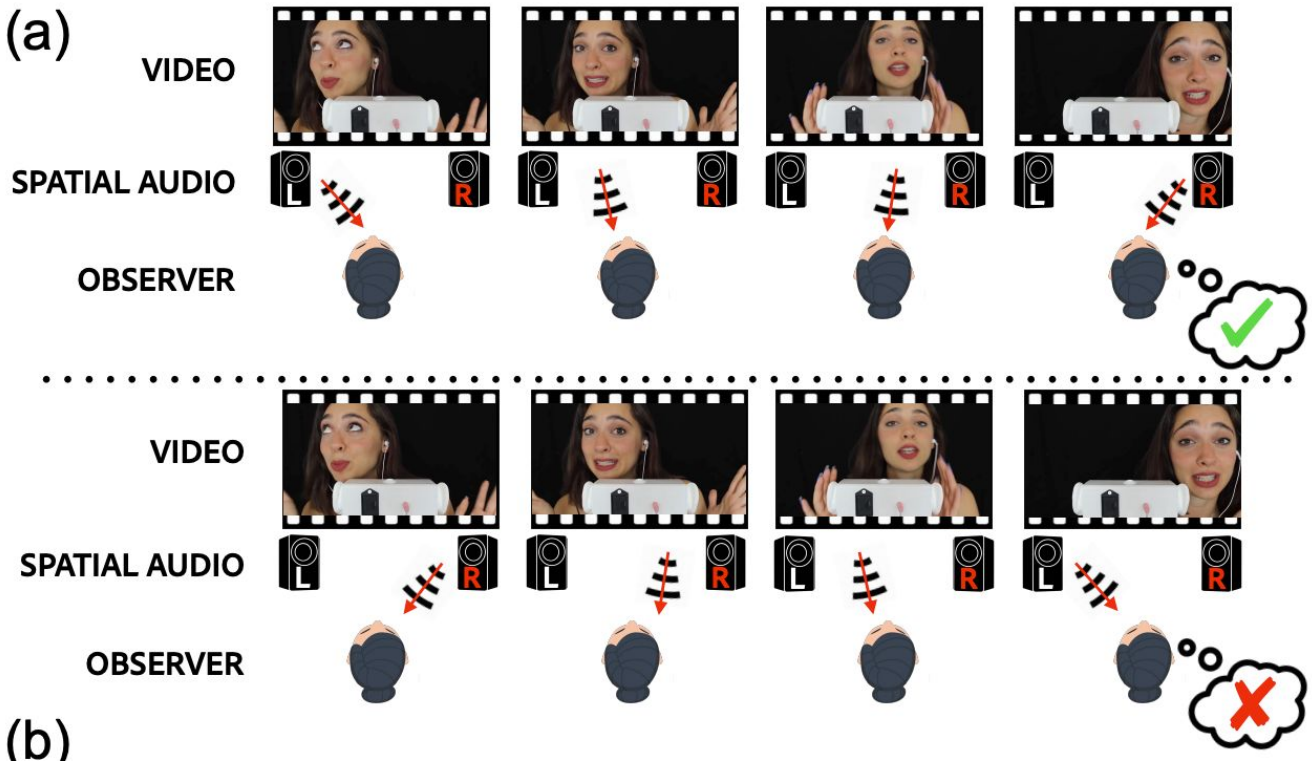
Arandjelović, Relja, and Andrew Zisserman. "Look, Listen and Learn." *ICCV 2017.*

Binary Verification



Owens, A., & Efros, A. A. Audio-visual scene analysis with self-supervised multisensory features. ECCV 2018.

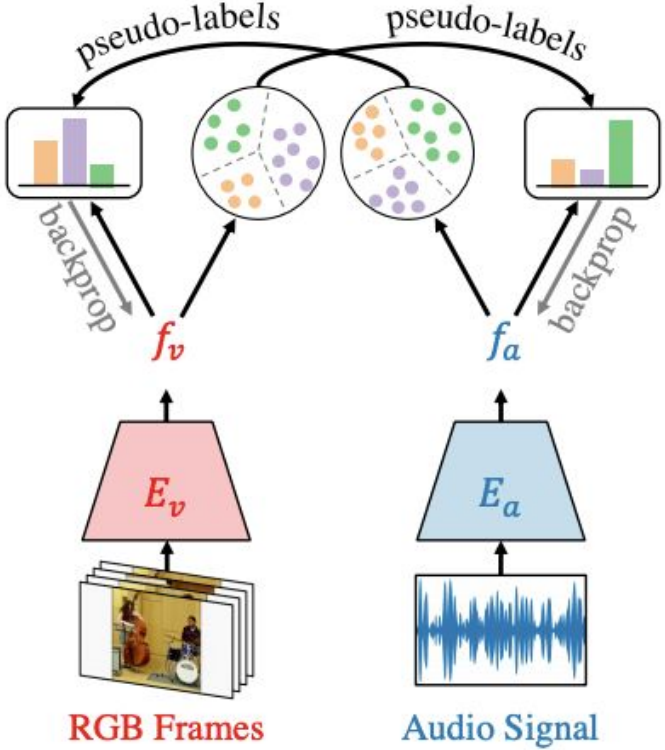
Binary Verification



Flipped right & left audio channels

Yang, Karren, Bryan Russell, and Justin Salamon. "Telling Left From Right: Learning Spatial Correspondence of Sight and Sound." CVPR 2020. [tweet]

Binary verification (clustering)

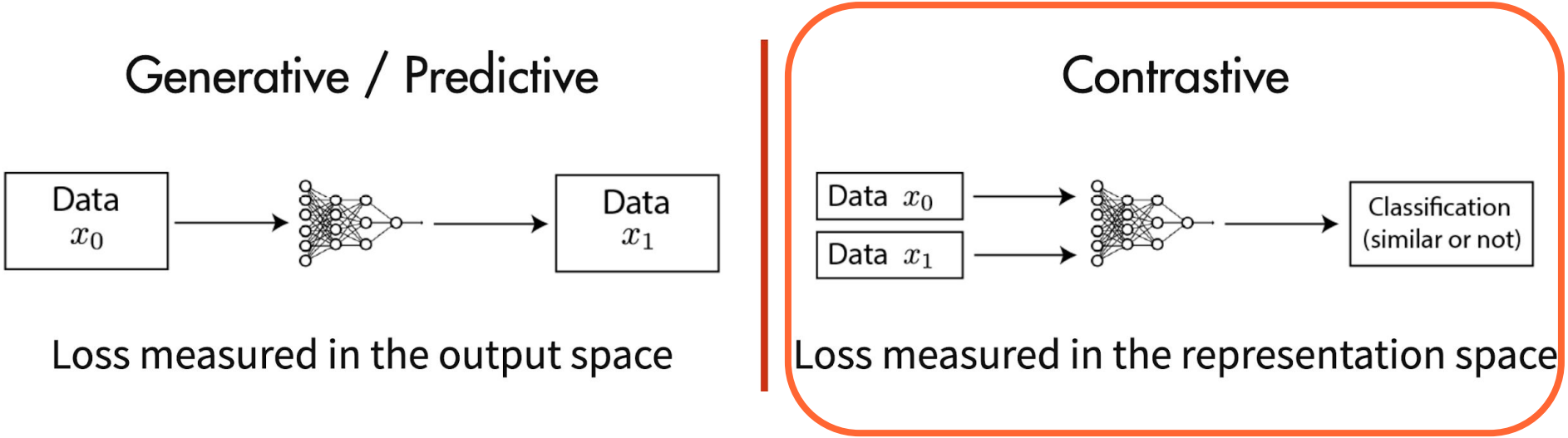


#XCD Alwassel, H., Mahajan, D., Torresani, L., Ghanem, B., & Tran, D. (2019). Self-Supervised Learning by Cross-Modal Audio-Video Clustering. arXiv preprint arXiv:1911.12667.

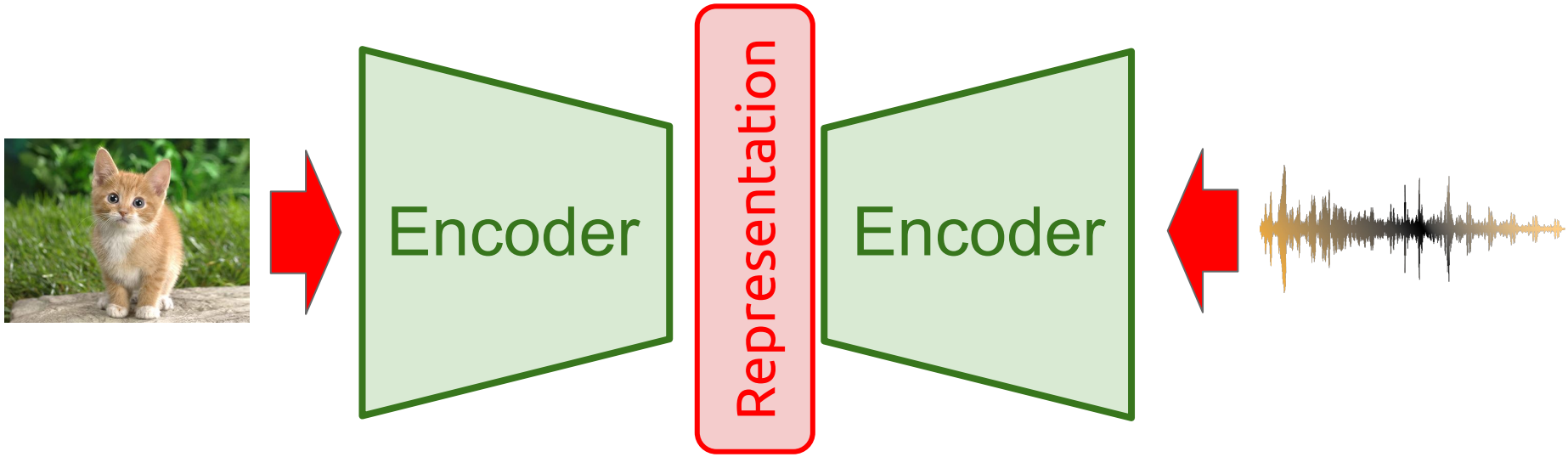
Outline

1. Motivation
2. Feature Learning
    a. Generative / Predictive Methods
    b. Contrastive Methods
3. Cross-modal Translation
4. Embodied AI

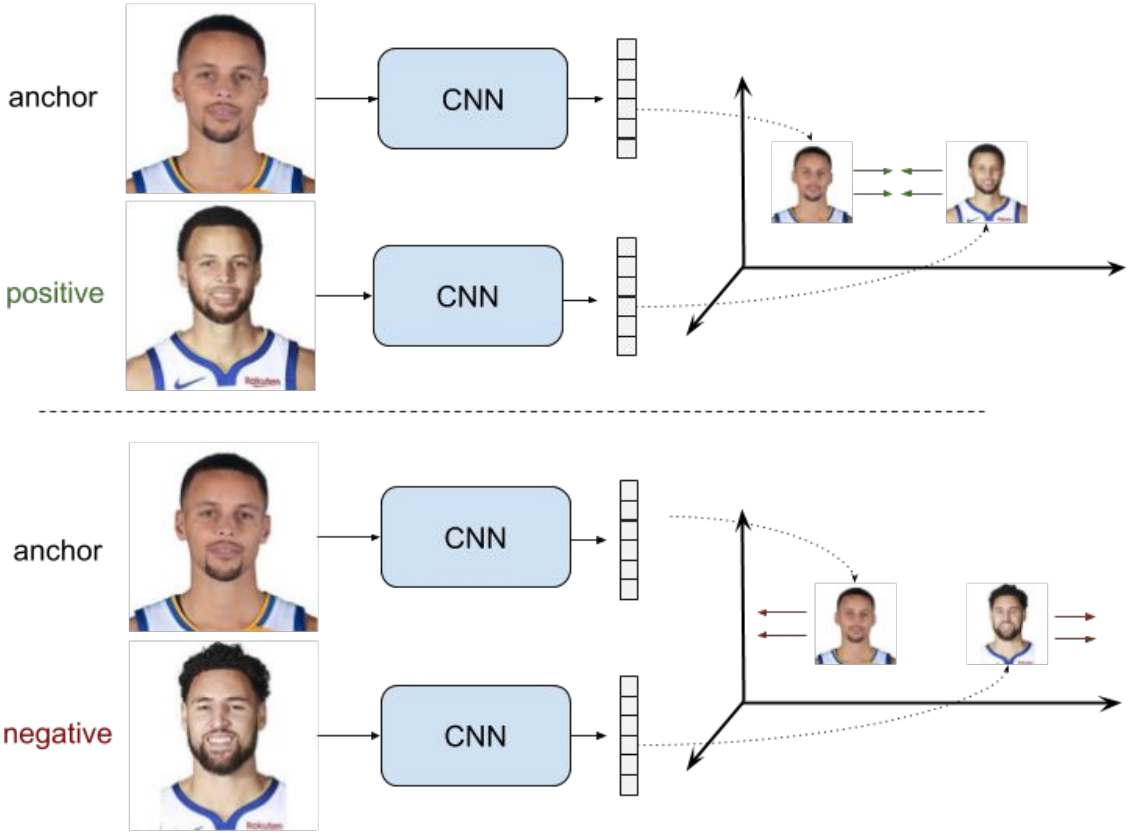
Self-supervised Feature Learning



Source: Ankesh Anand, "Contrastive Self-Supervised Learning" (2020)"

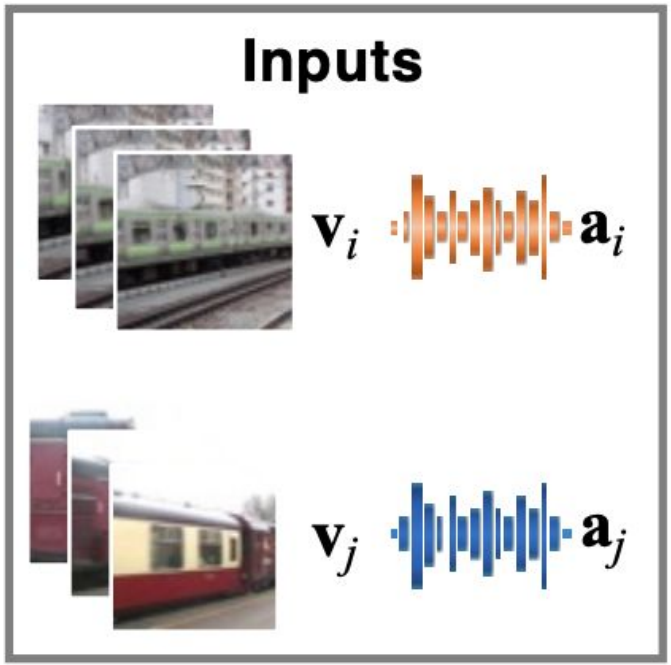

Encoder
Representation
Encoder

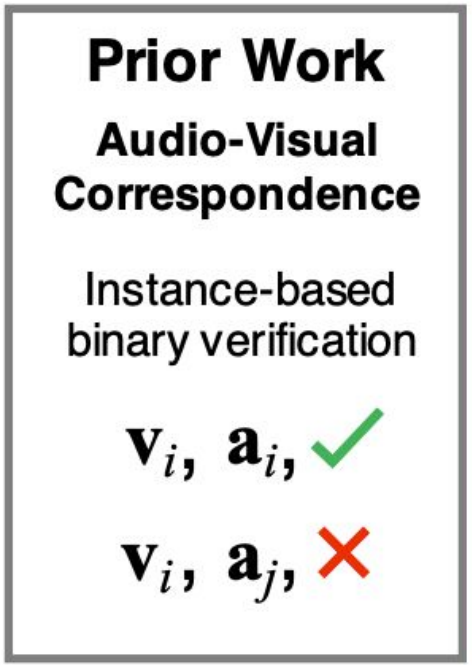
Contrastive Learning

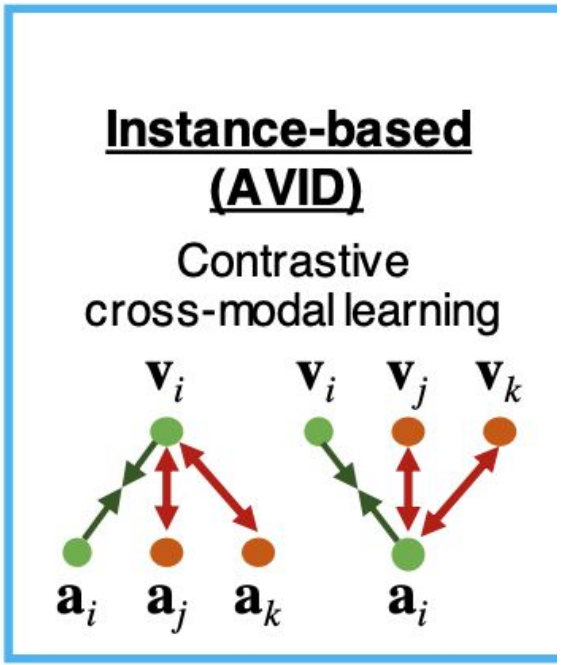


Source: Raul Gómez, "Understanding Ranking Loss, Contrastive Loss, Margin Loss, Triplet Loss, Hinge Loss and all those confusing names" (2019)

Contrastive Learning (cross-modal)

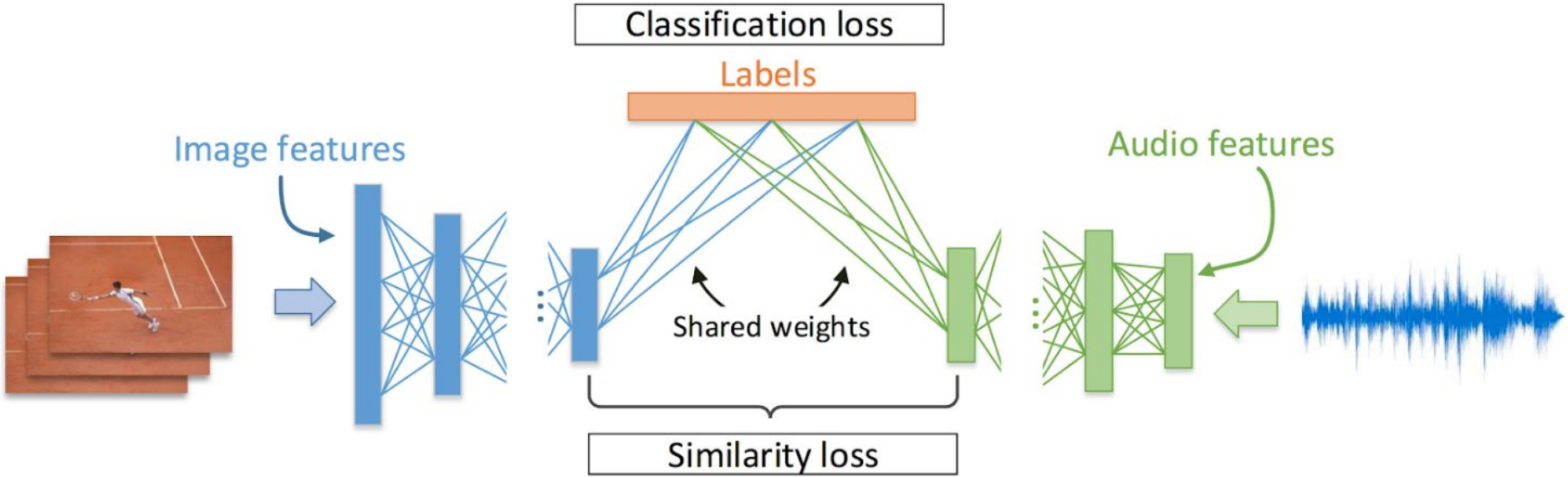






#AVID Morgado, Pedro, Nuno Vasconcelos, and Ishan Misra. "Audio-visual instance discrimination with cross-modal agreement." arXiv preprint arXiv:2004.12943 (2020). [code]

Contrastive Learning (cosine similarity+class)

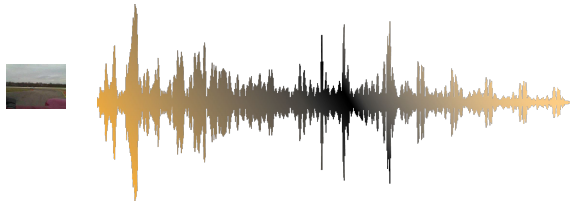


Amanda Duarte, Dídac Surís, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. "Cross-modal Embeddings for Video and Audio Retrieval." ECCV Women in Computer Vision Workshop 2018.

Contrastive Learning (cosine similarity+class)

Audio feature

Best match



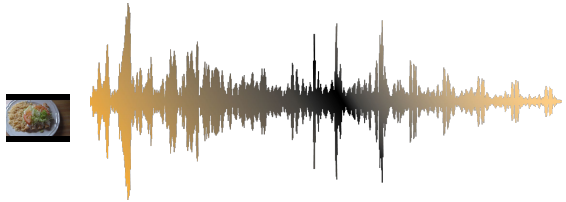
Amanda Duarte, Dídac Surís, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. "Cross-modal Embeddings for Video and Audio Retrieval." ECCV Women in Computer Vision Workshop 2018.

Contrastive Learning (cosine similarity+class)

Visual feature

Audio feature

Best match



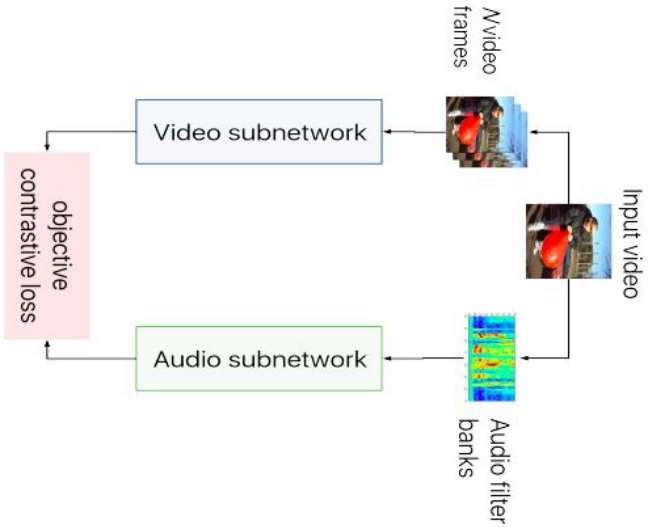
Amanda Duarte, Dídac Surís, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. "Cross-modal Embeddings for Video and Audio Retrieval." ECCV Women in Computer Vision Workshop 2018.

Contrastive Learning (pairwise L2 hinge loss)

Positive A/V pair

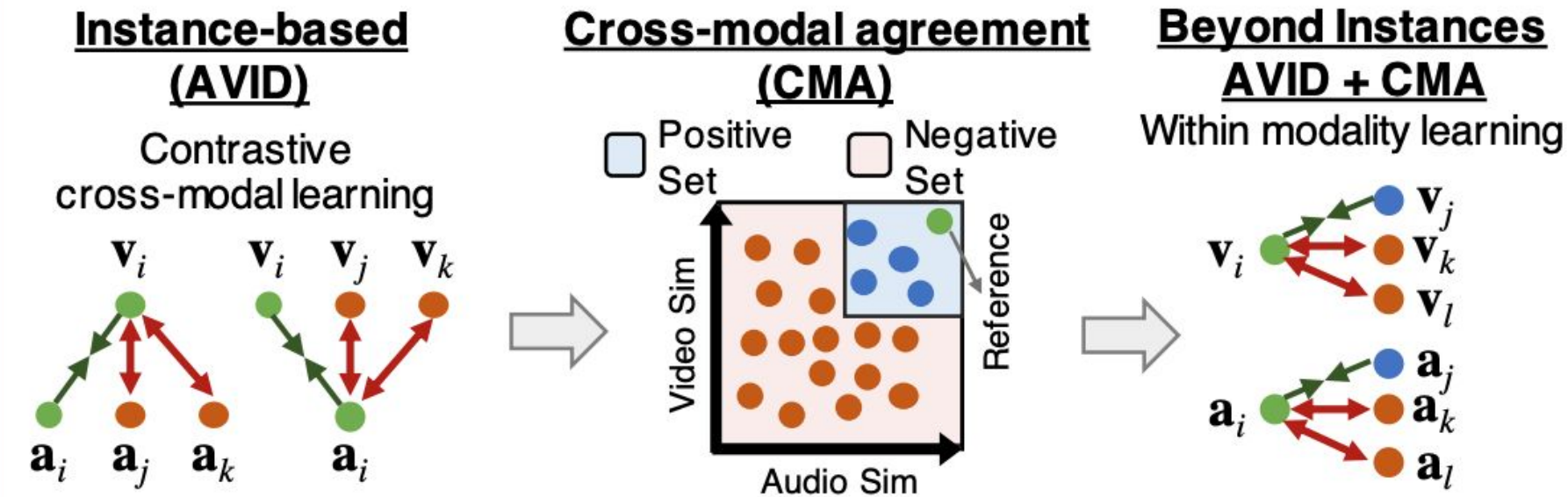
Negative A/V pair

$$E = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)})||f_v(v^{(n)}) - f_a(a^{(n)})||_2^2 + (1-y^{(n)})\max(\eta - ||f_v(v^{(n)}) - f_a(a^{(n)})||_2, 0)^2$$



#AVTS Korbar, Bruno, Du Tran, and Lorenzo Torresani. "Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization." NIPS 2018.

Contrastive Learning (within-modal)



#AVID #CMA Morgado, Pedro, Nuno Vasconcelos, and Ishan Misra. "Audio-visual instance discrimination with cross-modal agreement." arXiv preprint arXiv:2004.12943 (2020). [code]

Outline

1. Motivation
2. Feature Learning
3. **Cross-modal Translation**
    a. **Sound to Vision**
    b. Vision to Sound
4. Embodied AI

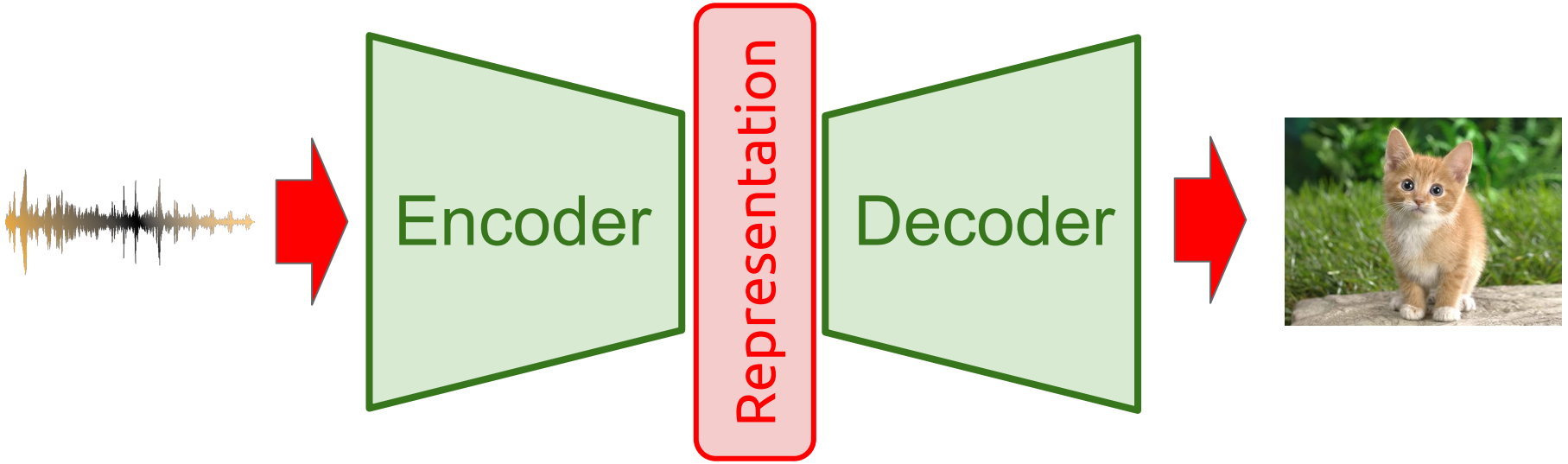
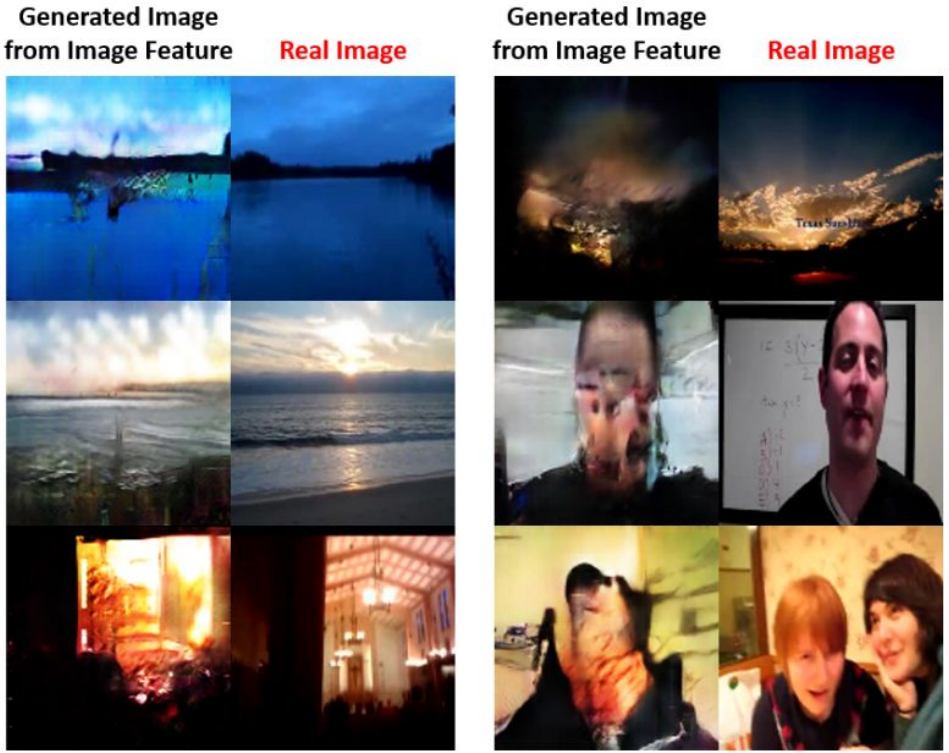

Encoder
Representation
Decoder

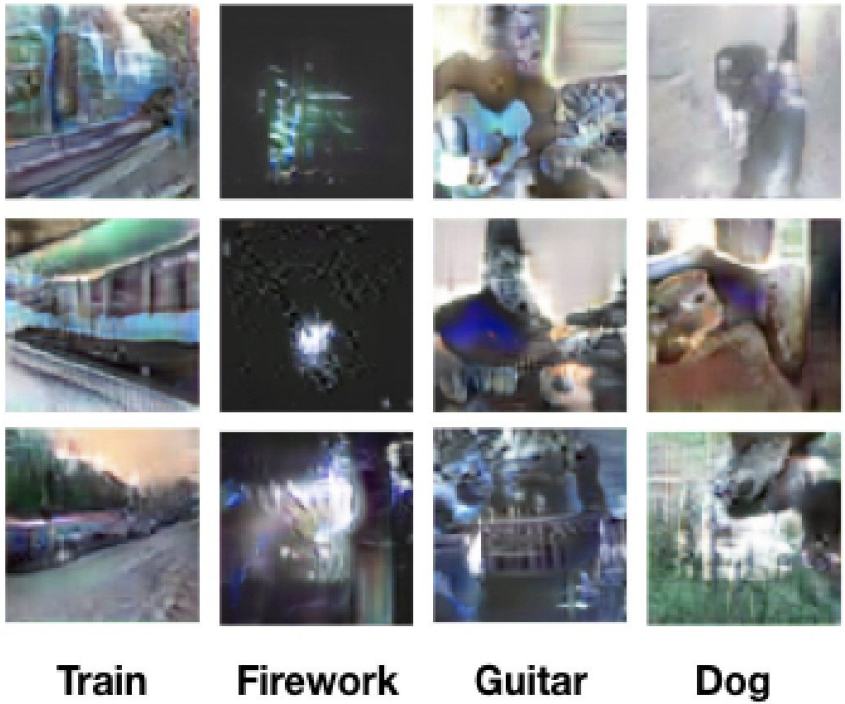
Image hallucination from sound



Lyu, Jeonghyun, Takashi Shinozaki, and Kaoru Amano. "Generating Images from Sounds Using Multimodal Features and GANs." (2018).

Image hallucination from sound

Conditional image generation based on StackGAN (stage I).

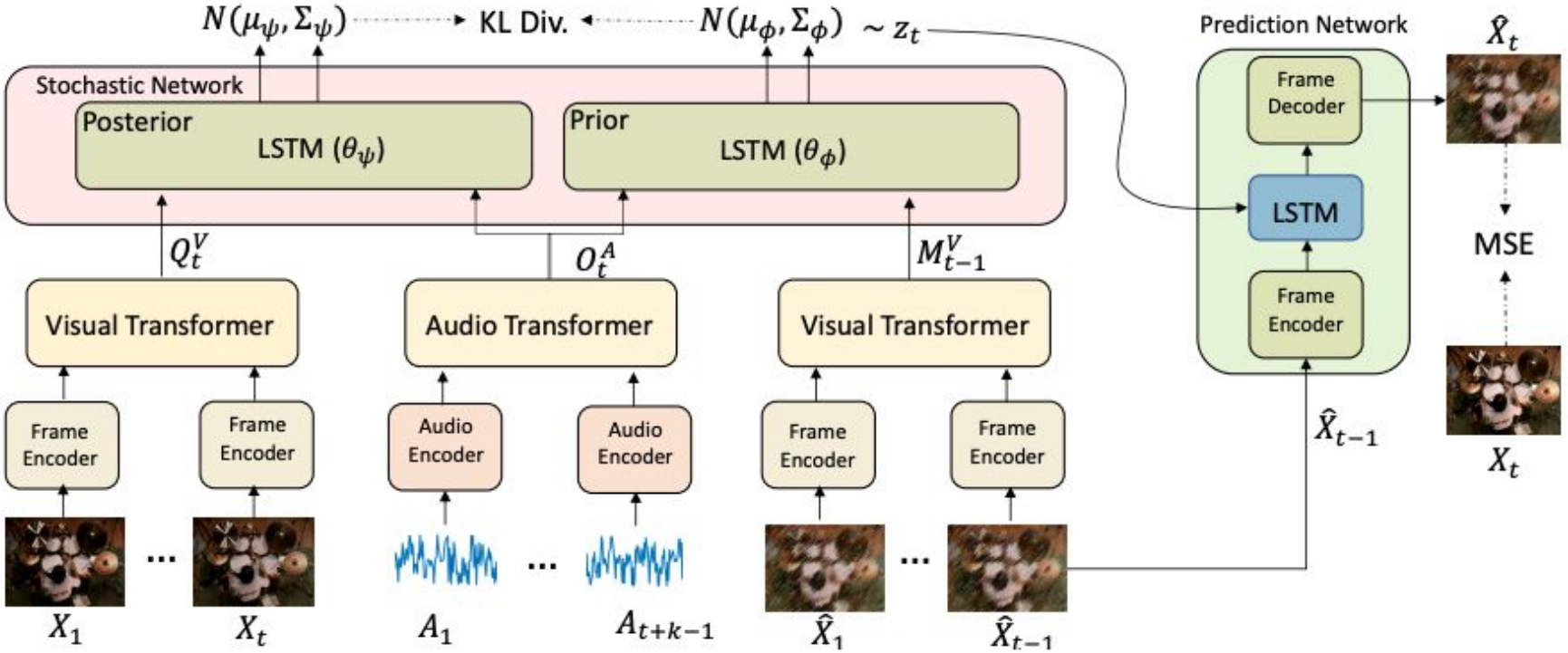


Chih Wen Lin, , "Generating Images from Audio" NeurIPS 2018 Creativity Workshop.

Image hallucination from sound

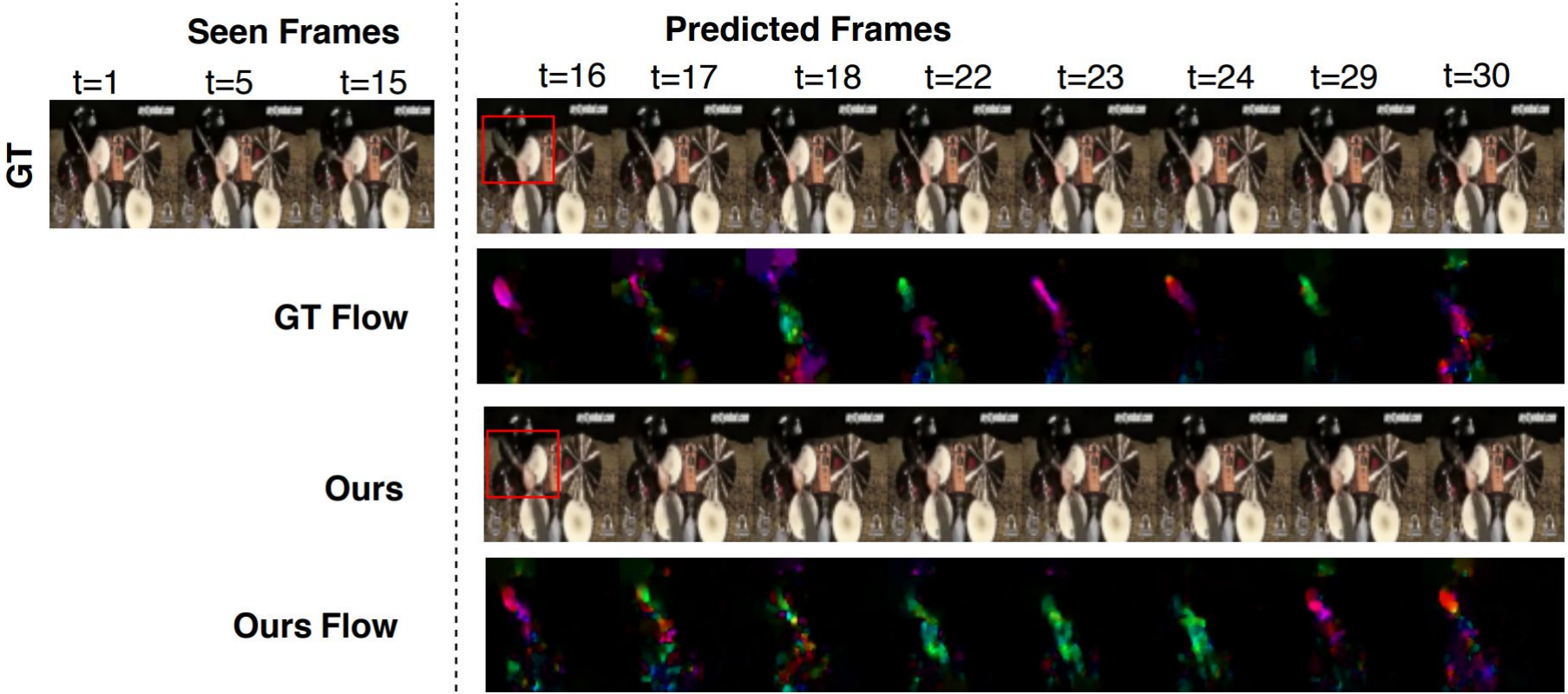
Wan, C. H., Chuang, S. P., & Lee, H. Y. (2019, May). Towards audio to scene image synthesis using generative adversarial network. ICASSP 2019.

Video hallucination from sound



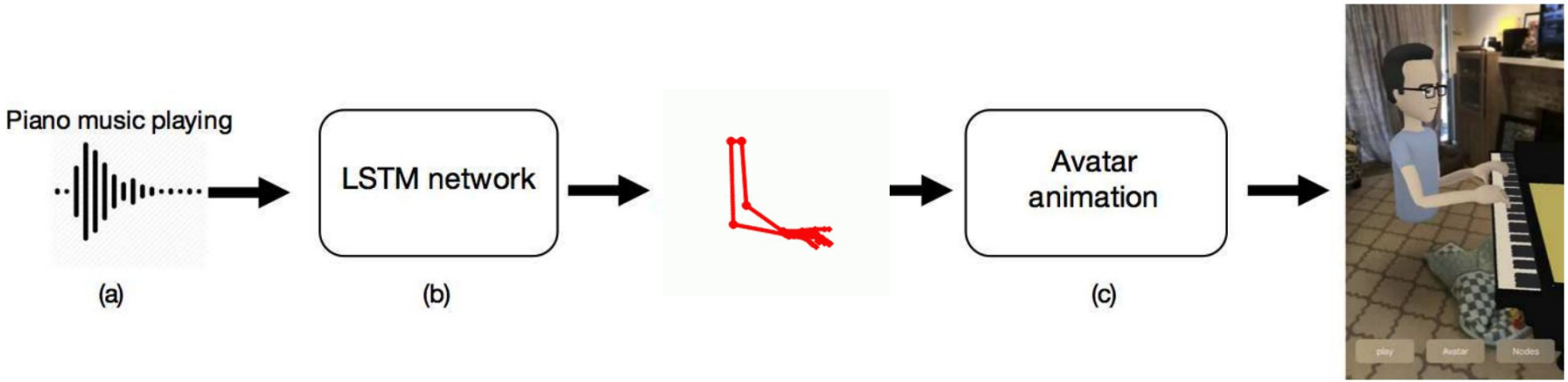
#Sound2Sight Cherian, Anoop, Moitreya Chatterjee, and Narendra Ahuja. "Sound2sight: Generating visual dynamics from sound and context." ECCV 2020.

Video hallucination from sound



#Sound2Sight Cherian, Anoop, Moitreya Chatterjee, and Narendra Ahuja. "Sound2sight: Generating visual dynamics from sound and context." ECCV 2020.

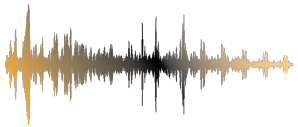
Avatar animation with music (skeletons)



Shlizerman, E., Dery, L., Schoen, H., & Kemelmacher-Shlizerman, I. (2018). Audio to body dynamics. CVPR 2018.

AR AVATAR
IPHONE RECORDING

Shlizerman, E., Dery, L., Schoen, H., & Kemelmacher-Shlizerman, I. (2018). Audio to body dynamics. CVPR 2018.



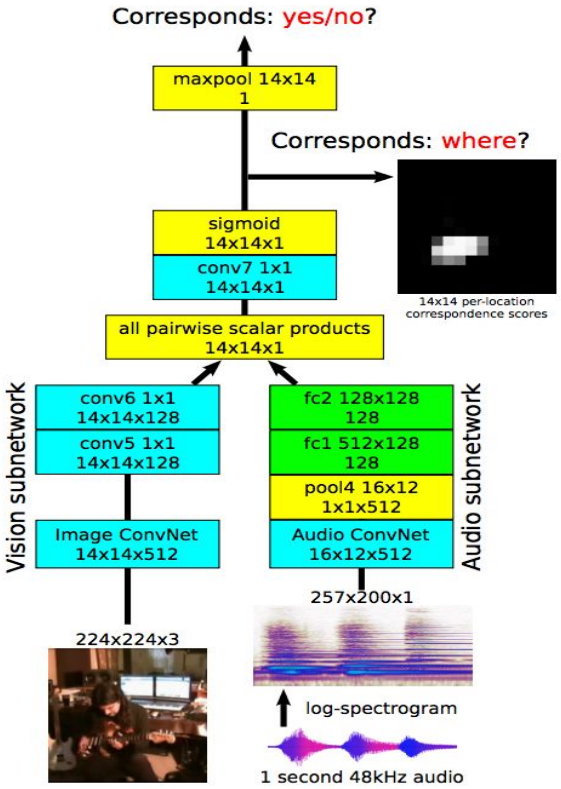
Encoder

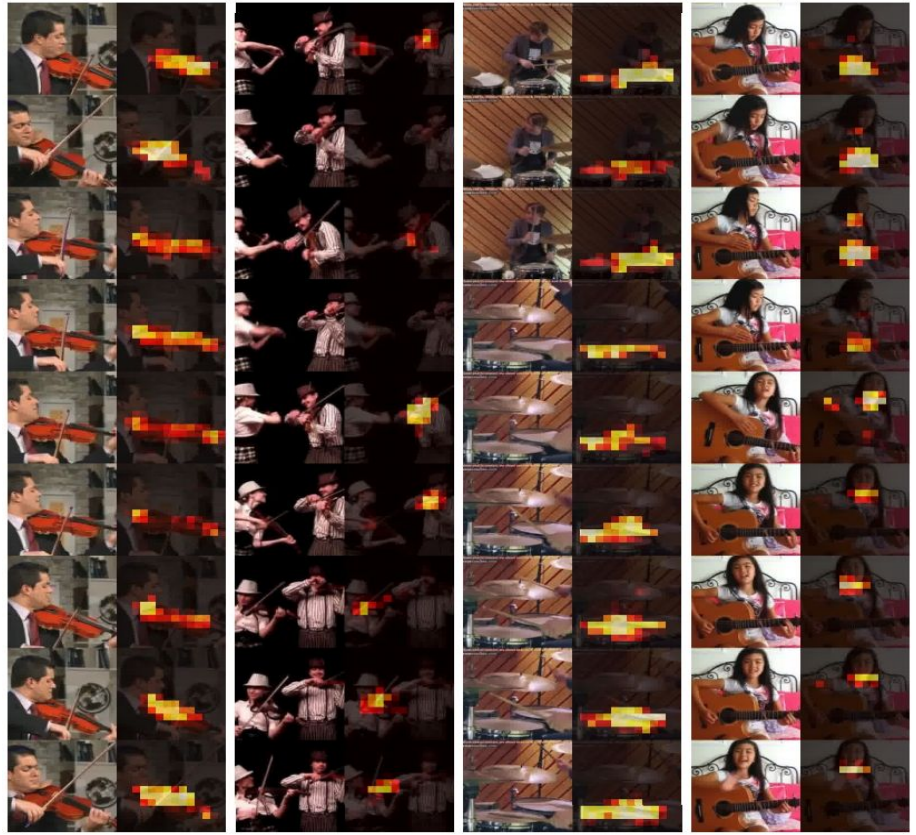
Encoder

Representation

Decoder

Sound Source Localization



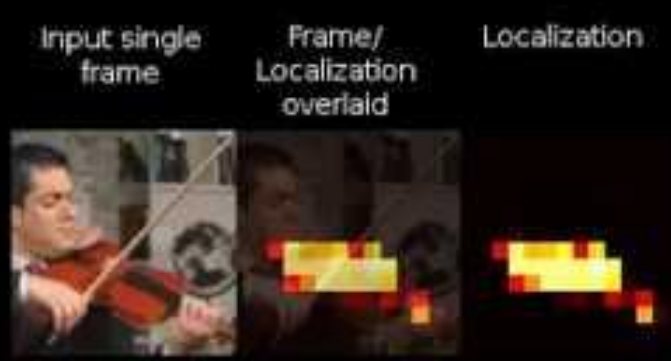


Arandjelović, Relja, and Andrew Zisserman. "Objects that Sound." ECCV 2018.

Objects that Sound

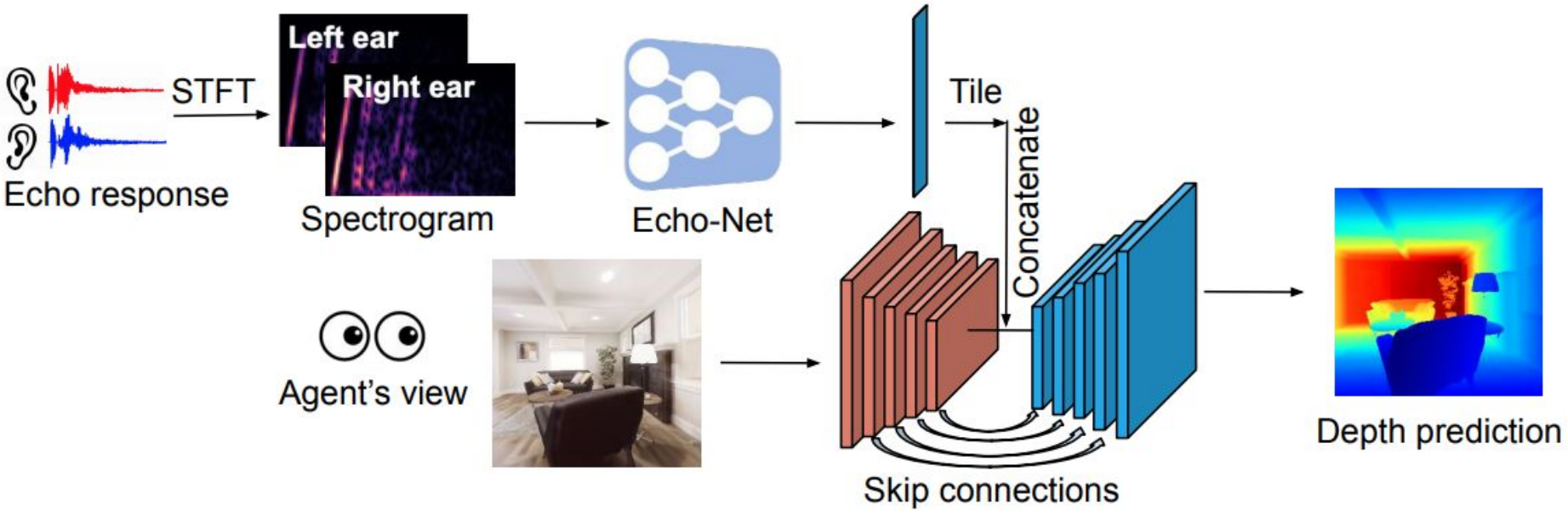
Relja Arandjelović[1] , Andrew Zisserman[1,2]
[1]DeepMind [2]University of Oxford

Frames are processed completely independently, motion information is not used, and there is no temporal smoothing

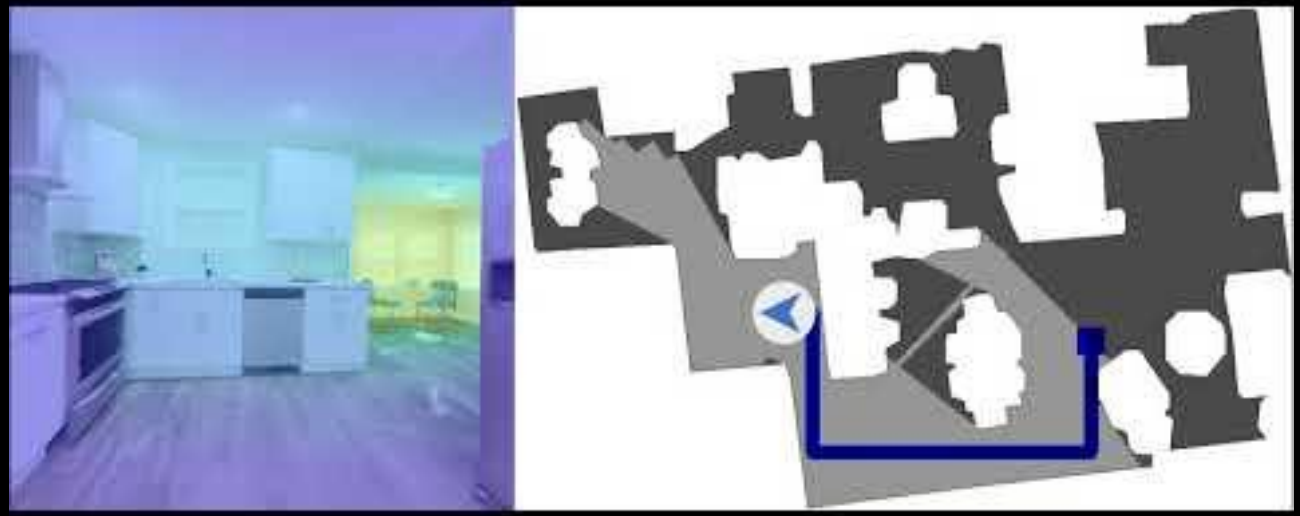


Arandjelović, Relja, and Andrew Zisserman. "Objects that Sound." ECCV 2018.

Depth Prediction by echoes



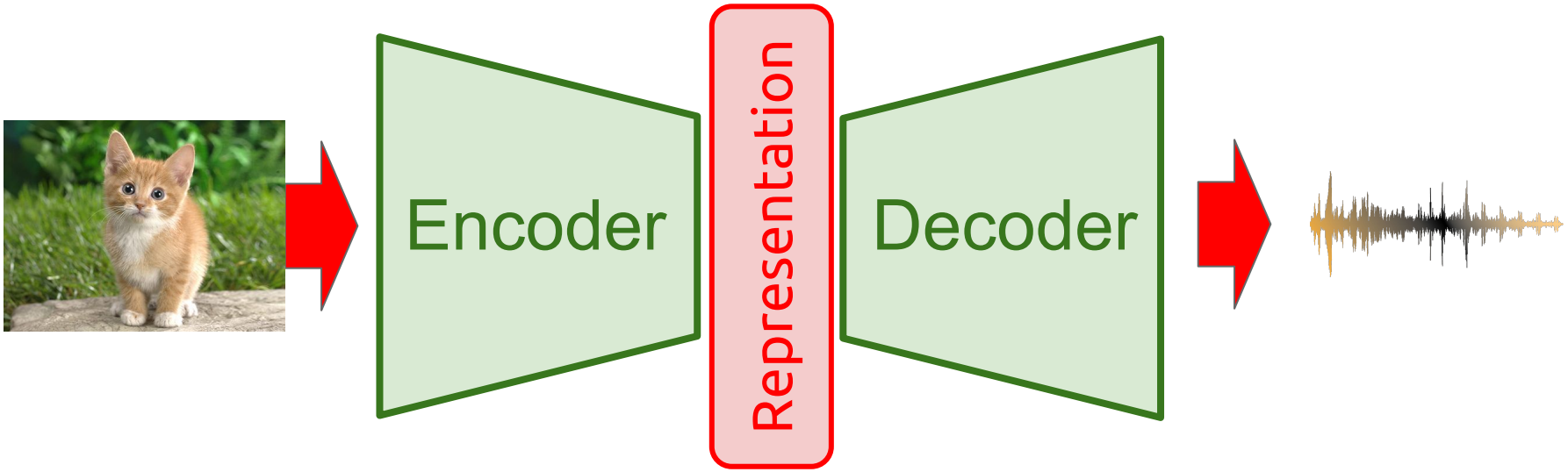
Gao, R., Chen, C., Al-Halah, Z., Schissler, C., & Grauman, K. (2020). VisualEchoes: Spatial Image Representation Learning through Echolocation. ECCV 2020.



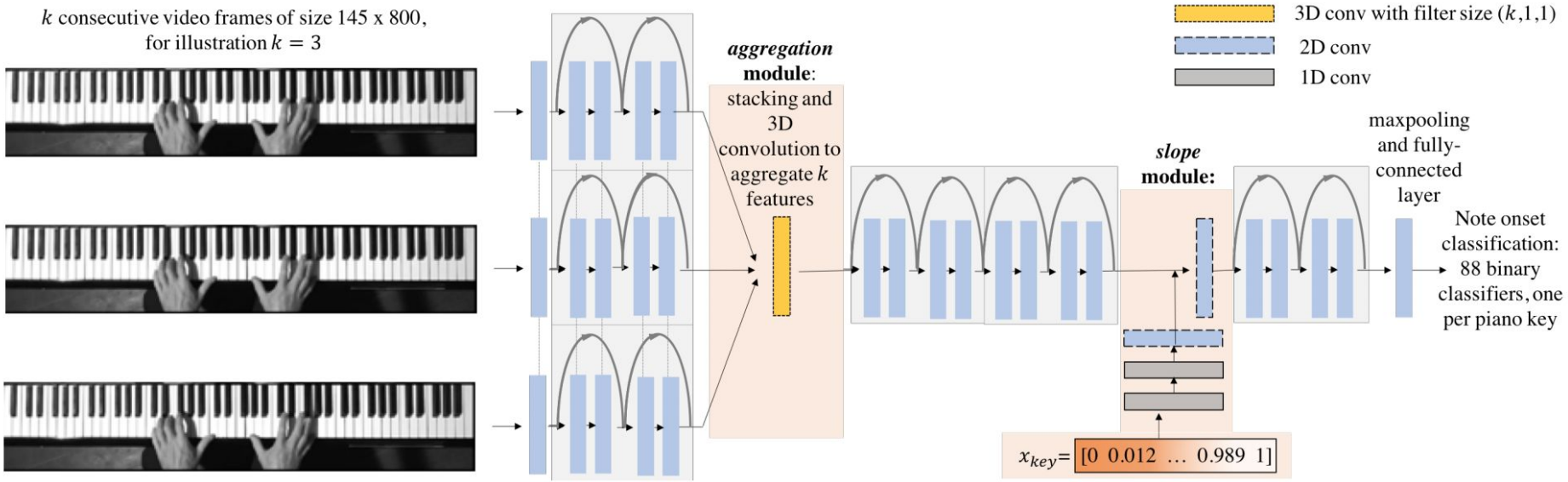
Gao, R., Chen, C., Al-Halah, Z., Schissler, C., & Grauman, K. VisualEchoes: Spatial Image Representation Learning through Echolocation. ECCV 2020.

Outline

1. Motivation
2. Feature Learning
3. **Cross-modal Translation**
    a. Sound to Vision
    b. Vision to Sound
4. Embodied AI


Encoder
Representation
Decoder

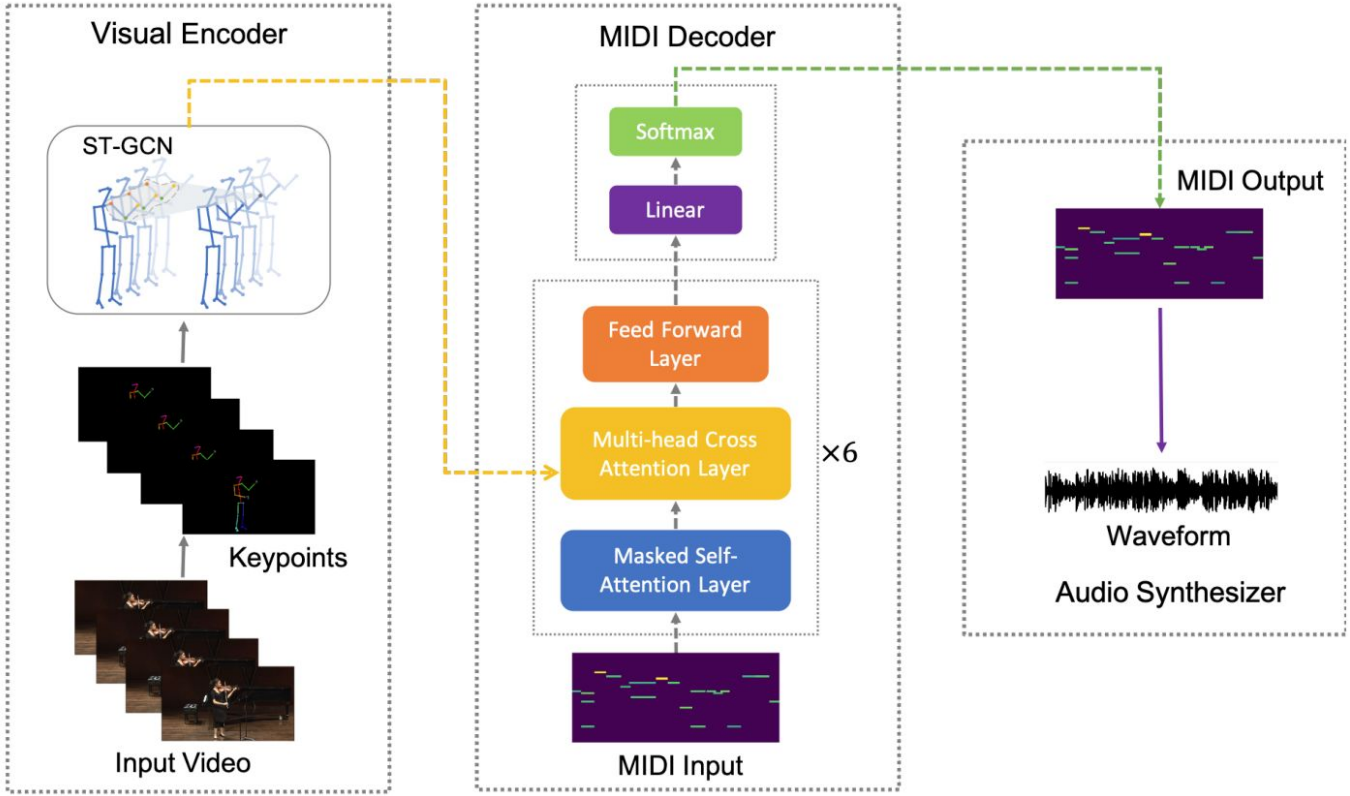
Piano Transcription (MIDI)



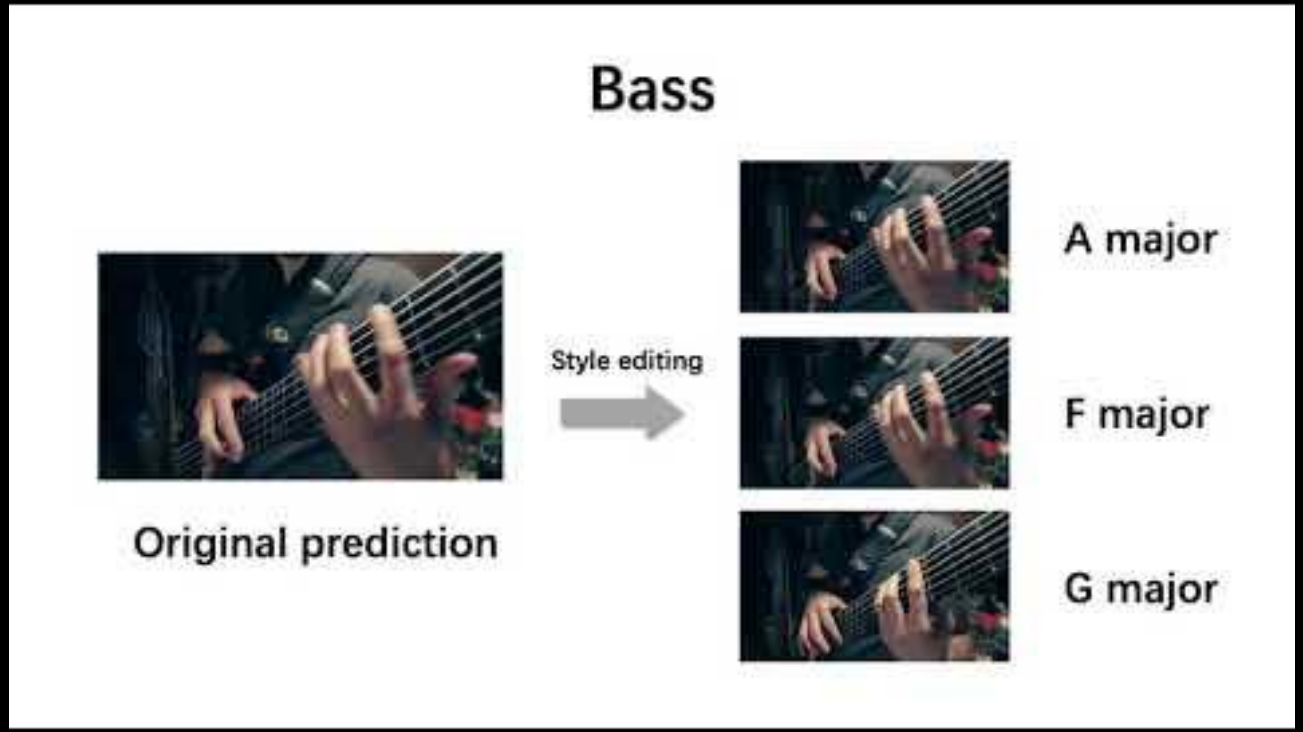
AS Koepke, O Wiles, Y Moses, A Zisserman, "SIGHT TO SOUND: AN END-TO-END APPROACH FOR VISUAL PIANO TRANSCRIPTION". ICASSP 2020.

AS Koepke, O Wiles, Y Moses, A Zisserman, "SIGHT TO SOUND: AN END-TO-END APPROACH FOR VISUAL PIANO TRANSCRIPTION". ICASSP 2020.

Silent Video Sonorization (MIDI)



#FoleyMusic Gan, Chuang, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. "Foley Music: Learning to Generate Music from Videos." ECCV 2020.



#FoleyMusic Gan, Chuang, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. "Foley Music: Learning to Generate Music from Videos." ECCV 2020.

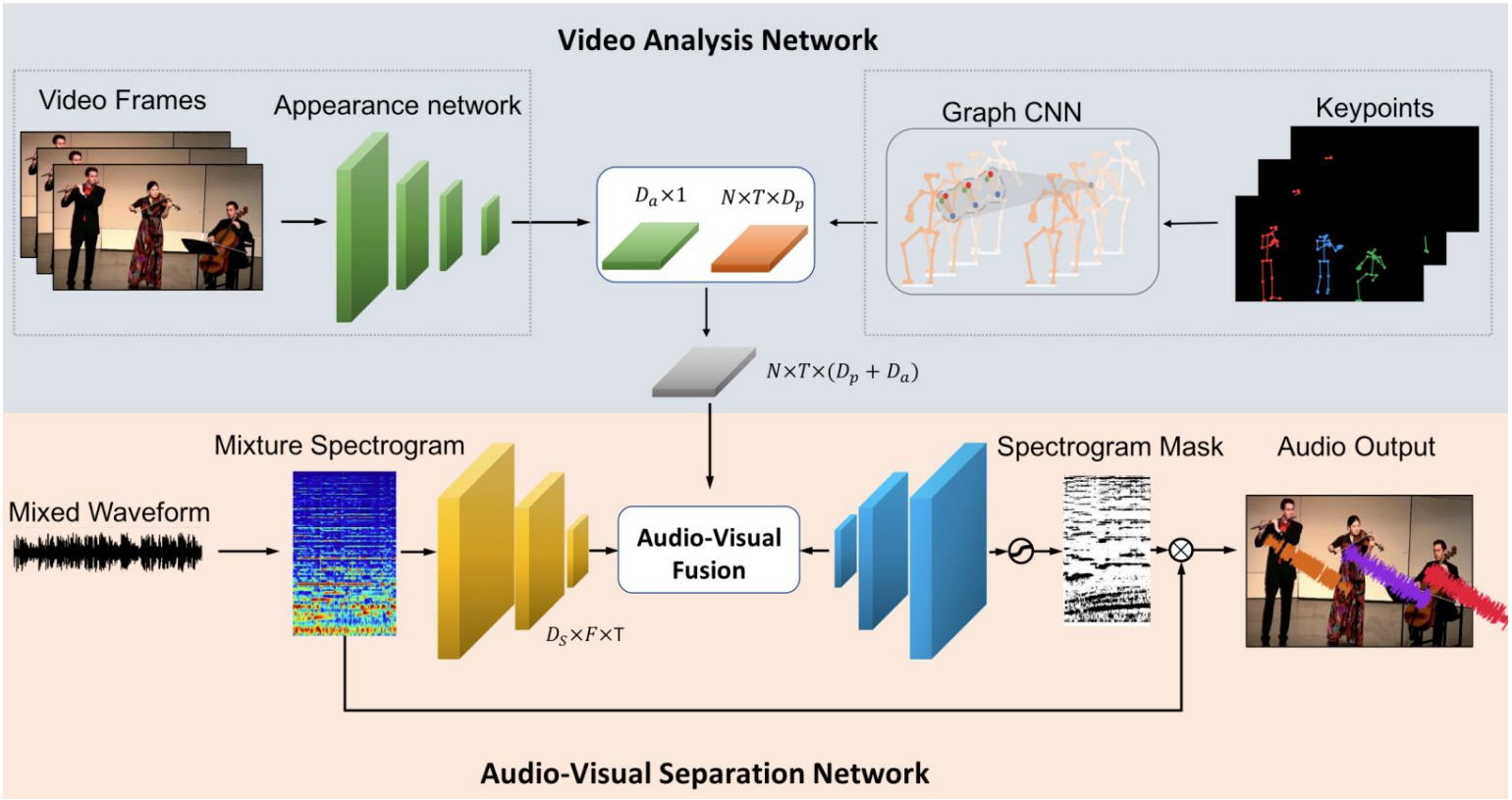
Encoder

Encoder

Representation

Decoder

Sound Separation



Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, Antonio Torralba, "Music Gesture for Visual Sound Separation" CVPR 2020.

Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, Antonio Torralba, "Music Gesture for Visual Sound Separation" CVPR 2020.

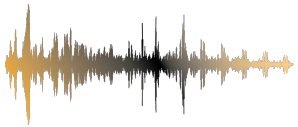
Encoder

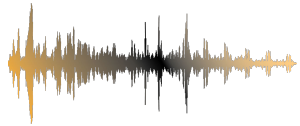
Encoder

Representation

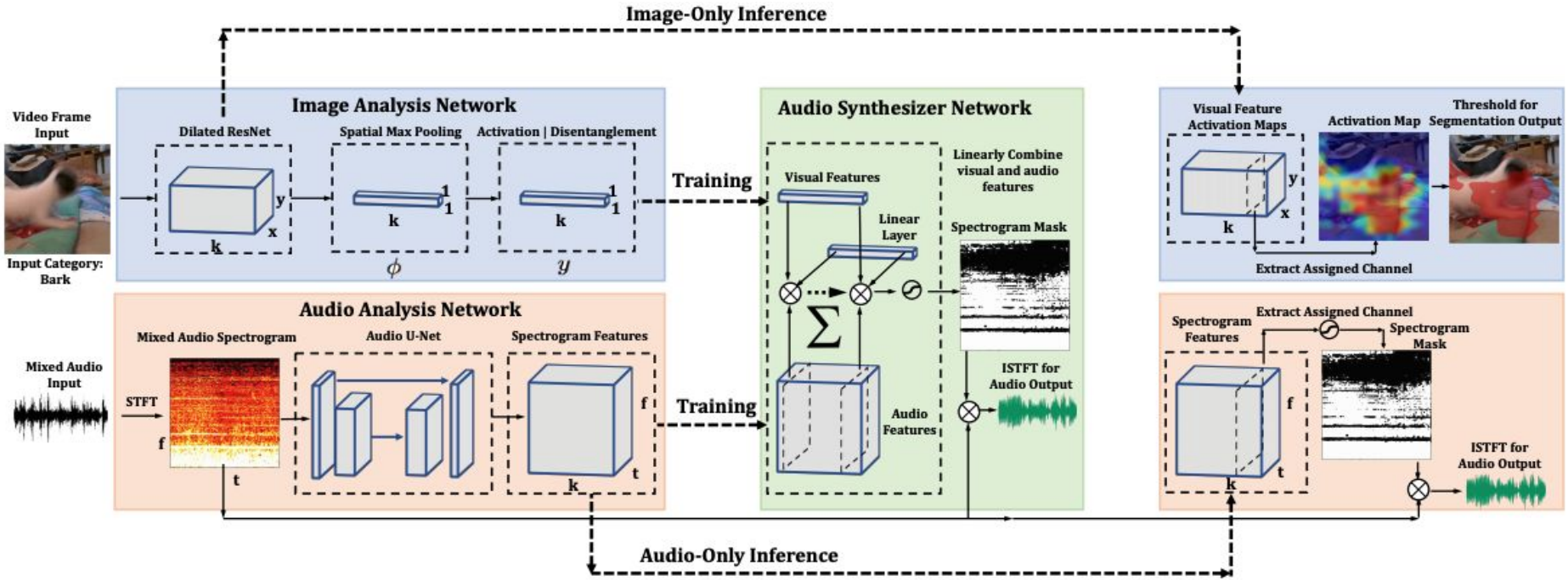
Decoder

Decoder



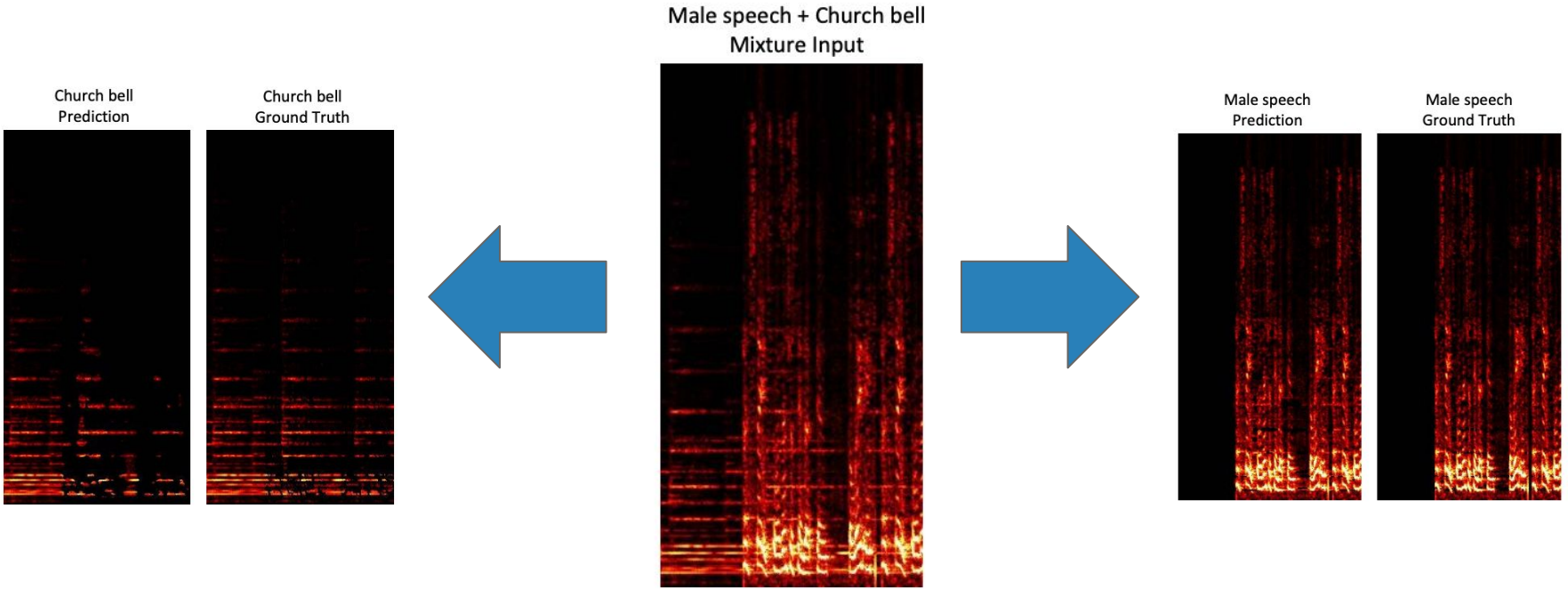


Source Separation + Segmentation



Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., & Torralba, A.. Self-supervised audio-visual co-segmentation. ICASSP 2019..

Source Separation + Segmentation

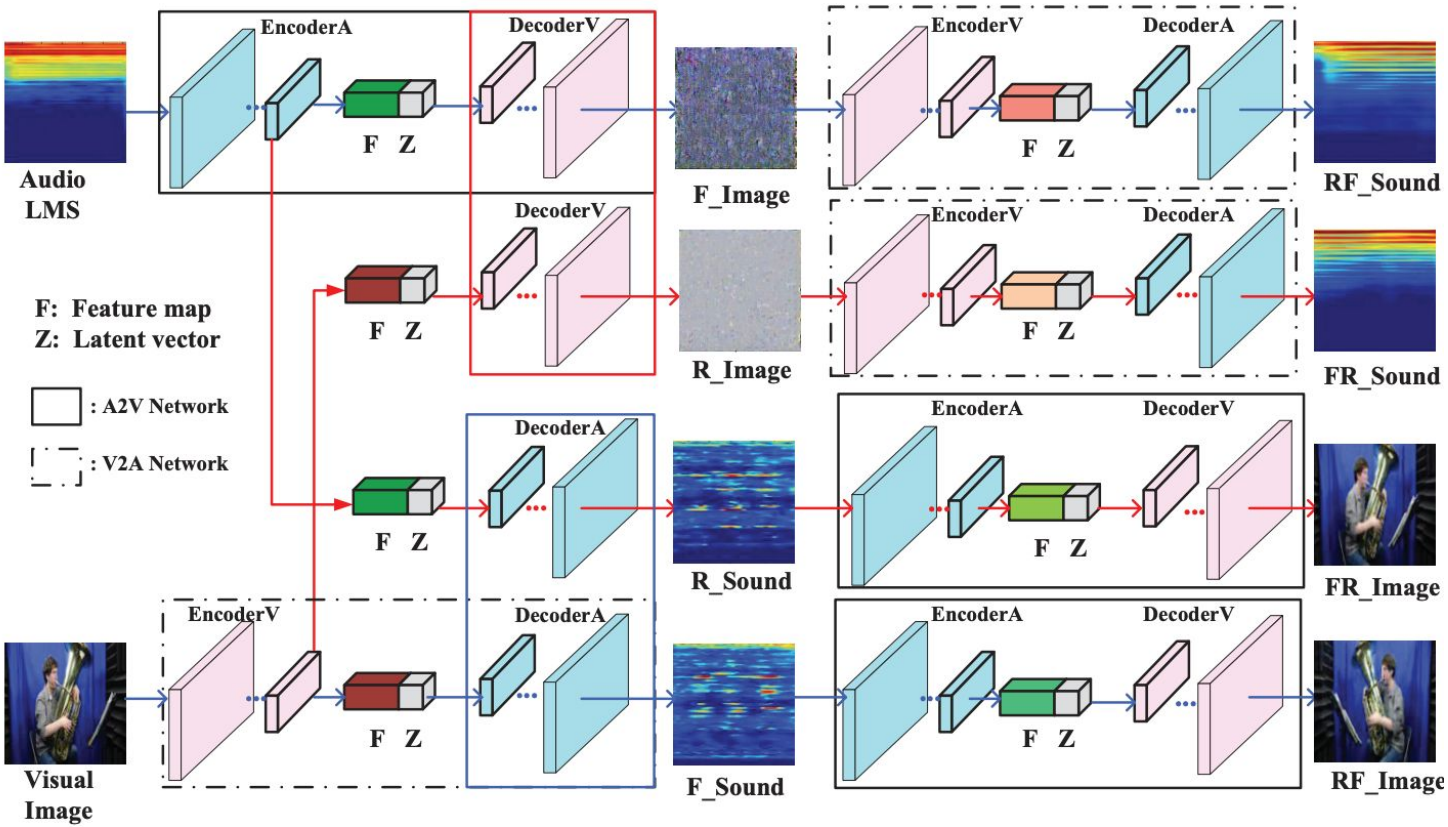


Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., & Torralba, A.. Self-supervised audio-visual co-segmentation. ICASSP 2019.

Source Separation + Segmentation

Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., & Torralba, A.. Self-supervised audio-visual co-segmentation. ICASSP 2019..

Visual & Audio Generation (cycle)

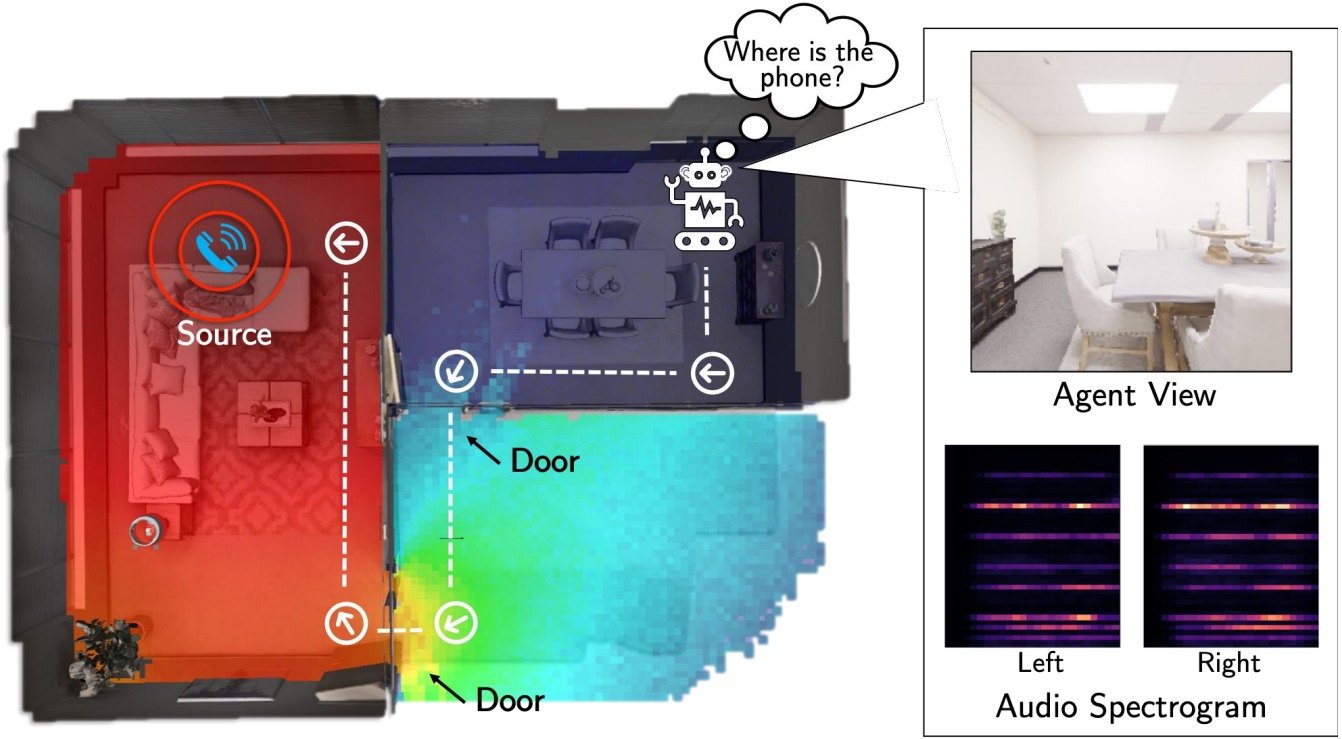


#CMCGAN Hao, W., Zhang, Z., & Guan, H. Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. AAAI 2018.

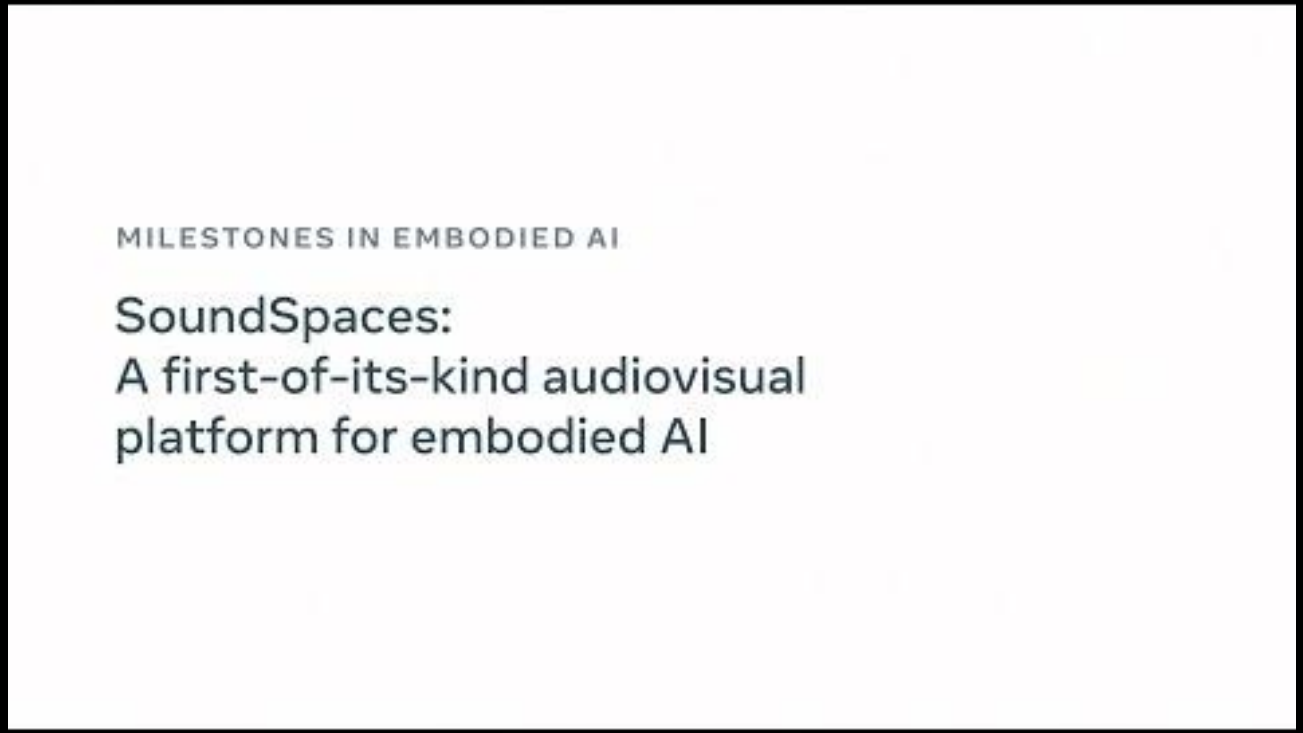
Outline

1. Motivation
2. Feature Learning
3. Cross-modal Translation
    a. Sound to Vision
    b. Vision to Sound
4. **Embodied AI**

Audio-visual Navigation with Deep RL



#SoundSpaces Chen, Changan, Unnat Jain, Carl Schissler, Sebastia Vicenc, Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. "SoundSpaces: Audio-visual navigation in 3d environments." ECCV 2020.



#SoundSpaces Chen, Changan, Unnat Jain, Carl Schissler, Sebastia Vicenc, Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. "SoundSpaces: Audio-visual navigation in 3d environments." ECCV 2020.

Outline

1. Motivation
2. Feature Learning
3. Cross-modal Translation
    a. Sound to Vision
    b. Vision to Sound
4. Embodied AI

Take home message

1. Motivation
2. Feature Learning
3. Cross-modal Translation
    a. Sound to Vision
    b. Vision to Sound
4. **Embodied AI**

Go raibh maith agat / Thank you

Was this tutorial helpful ? Please consider citing:

```
Giro-i-Nieto, X. One Perceptron to Rule Them All: Language,
Vision, Audio and Speech. In Proceedings of the 2020
International Conference on Multimedia Retrieval (pp. 7-8).
```

Xavier Giro-i-Nieto

@DocXavi
xavier.giro@upc.edu

Associate Professor
Universitat Politècnica de Catalunya
Barcelona Supercomputing Center