Objects that Sound

CAP6412 Advanced Computer Vision

Furkan Çimen

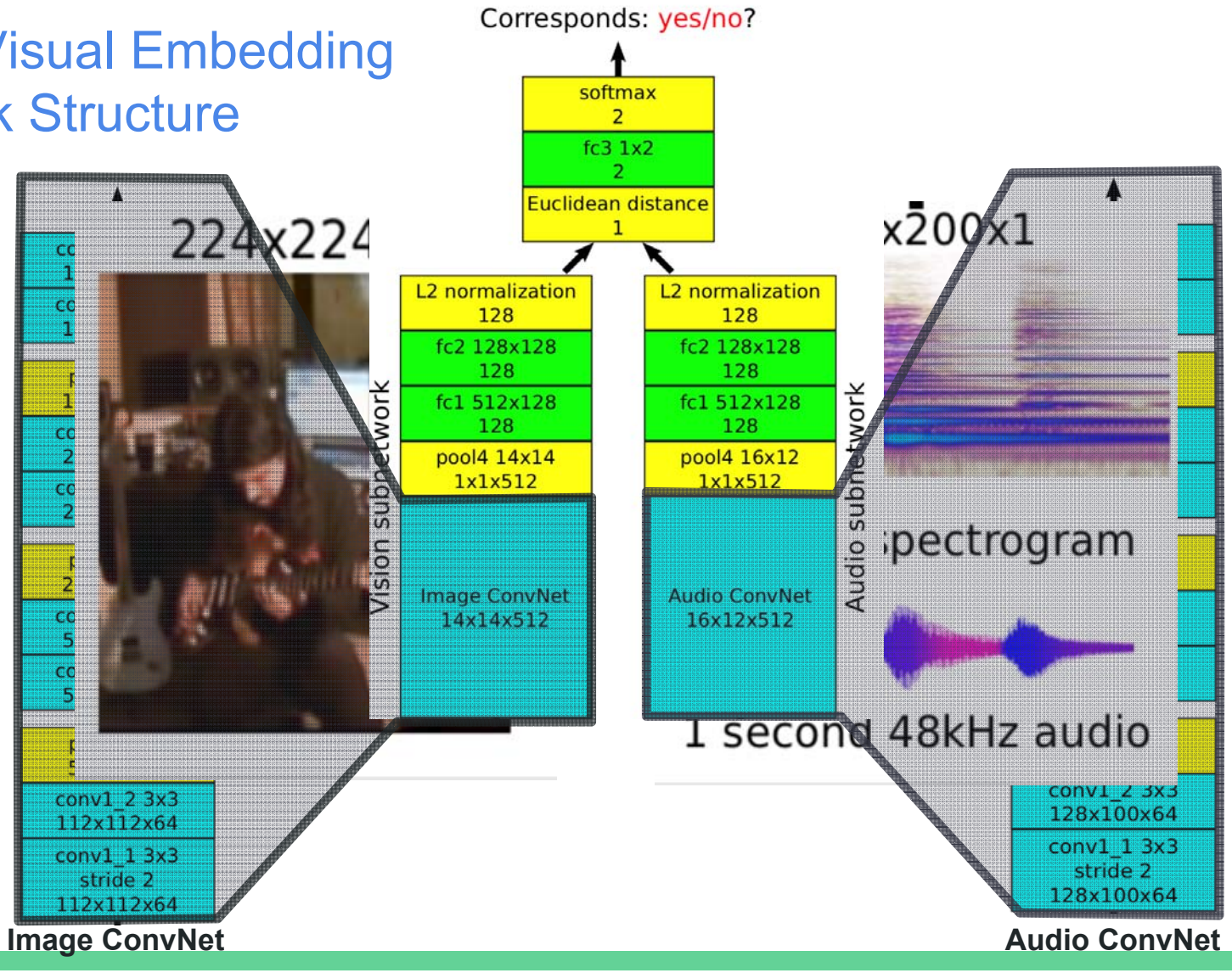
Motivation

- Show audio and visual embedding
  - Can be mapped into a common space
  - Can be learned only using audio-visual correspondence (AVC)
  - Suitable for within single modality and cross-modal retrieval
- Semantic object that sounds can be localized using only sound
- Explore various network architectures for AVC task

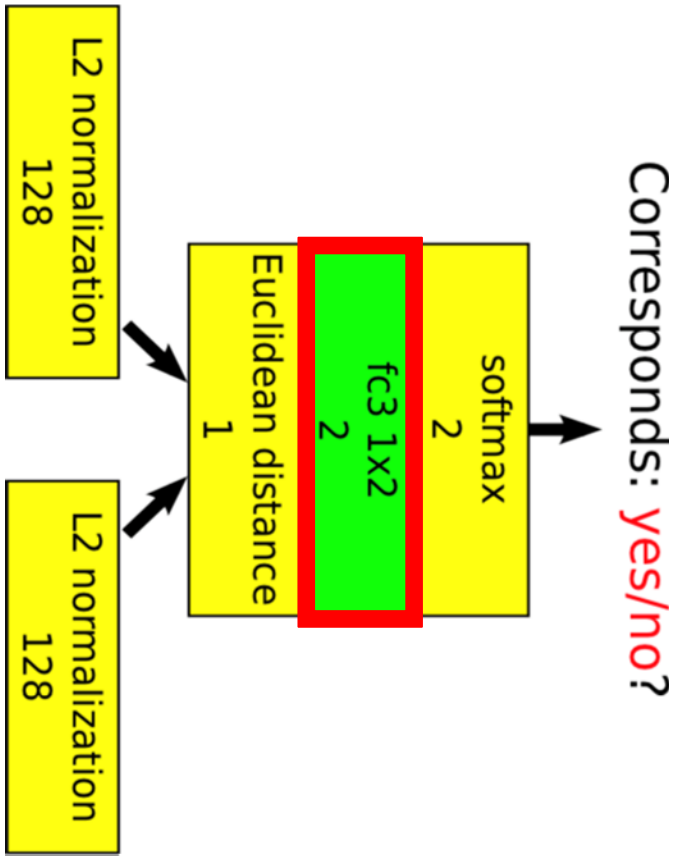
Dataset from AudioSet

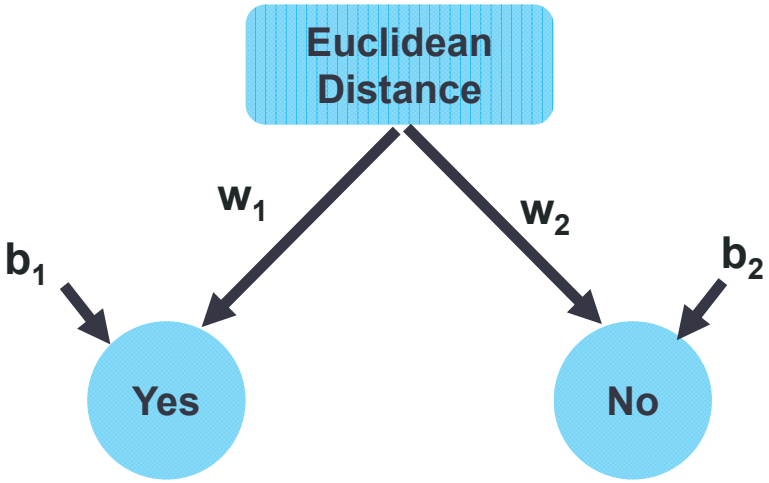
- Publicly available AudioSet dataset
- 10 second clips from YouTube
- Filtered to include only musical instruments and singing
  - 110 audio classes, i.e. Accordion, Flute, Piano …
  - Train, validation and test splits: 263k, 30k and 4.3k
  - Labels only used for quantitative evaluation

Audio-Visual Embedding Network Structure



Instead of Hyper Parameters Tiny FC Layer





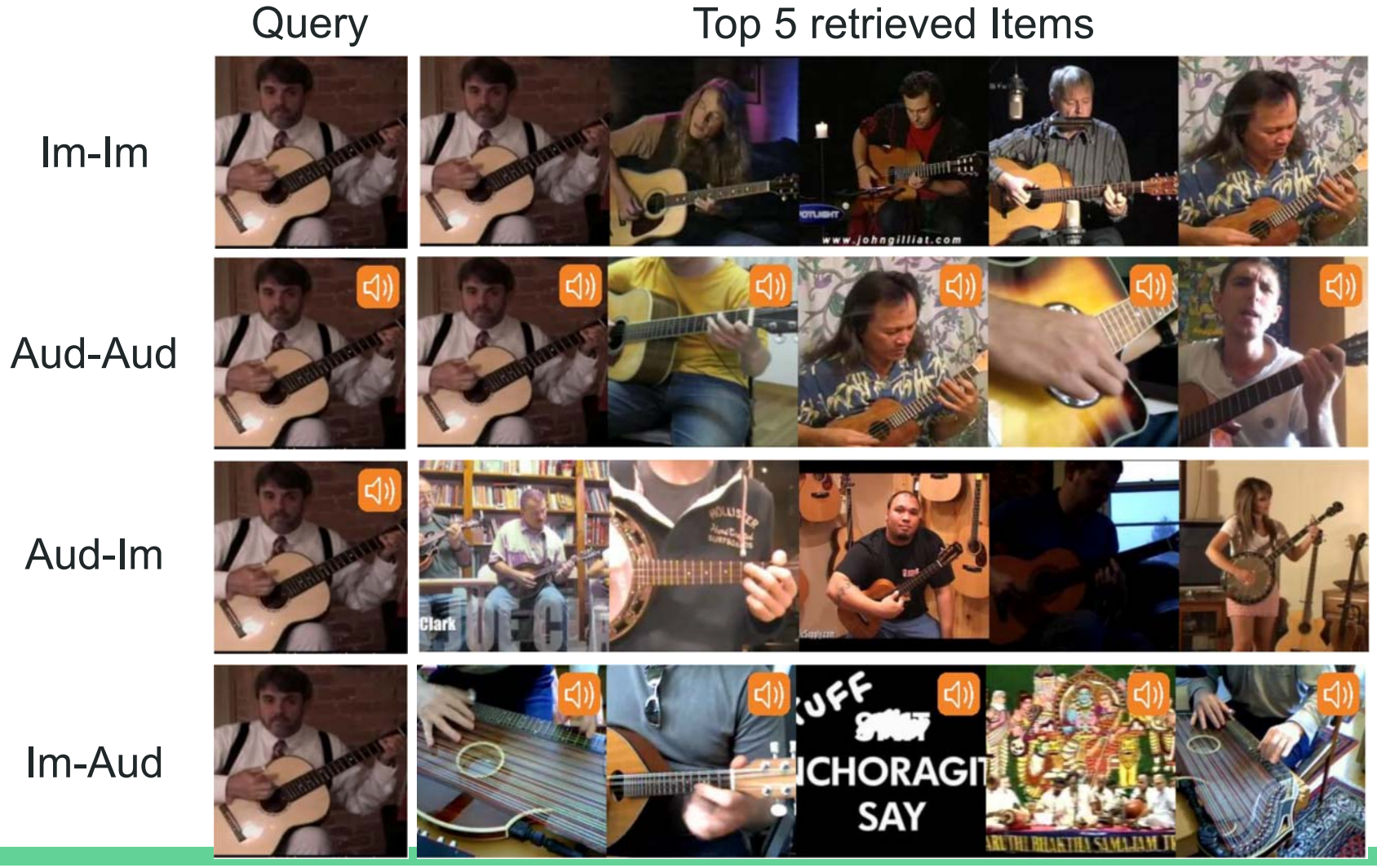
w_1=-1 b_1= 5 w_2= 1 b_2=-5

E.D.	Yes	Y/Softmax	No	N/Softmax
10	-5	~0	5	~1
1	4	~1	-4	~0

nDCG@30 Cross-modal and Intra-model Retrieval Comparison

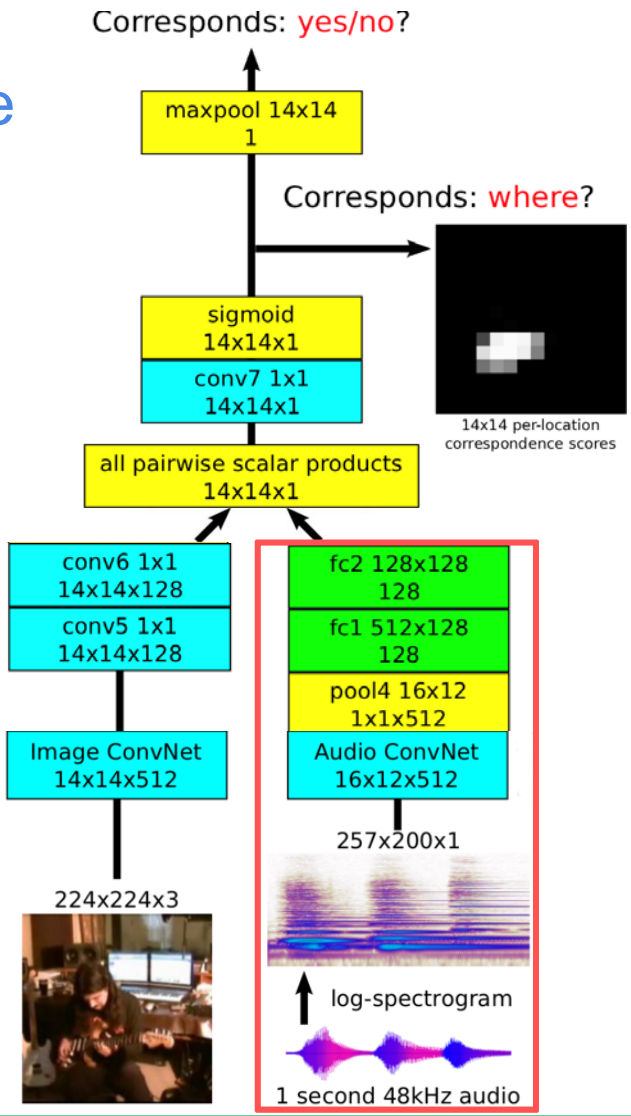
Method	im-im	im-aud	aud-im	aud-aud
Random	.407	.407	.407	.407
L^3-Net	.567	.418	.385	.653
L^3-Net with CCA	.578	.531	.560	.649
VGG16-ImageNet	.600	-	-	-
VGG16-ImageNet + L^3-Audio CCA	.493	.458	.464	.618
AVE-Net	**.604**	**.561**	**.587**	**.665**

Qualitative Cross-modal and Intra-model Retrieval Results

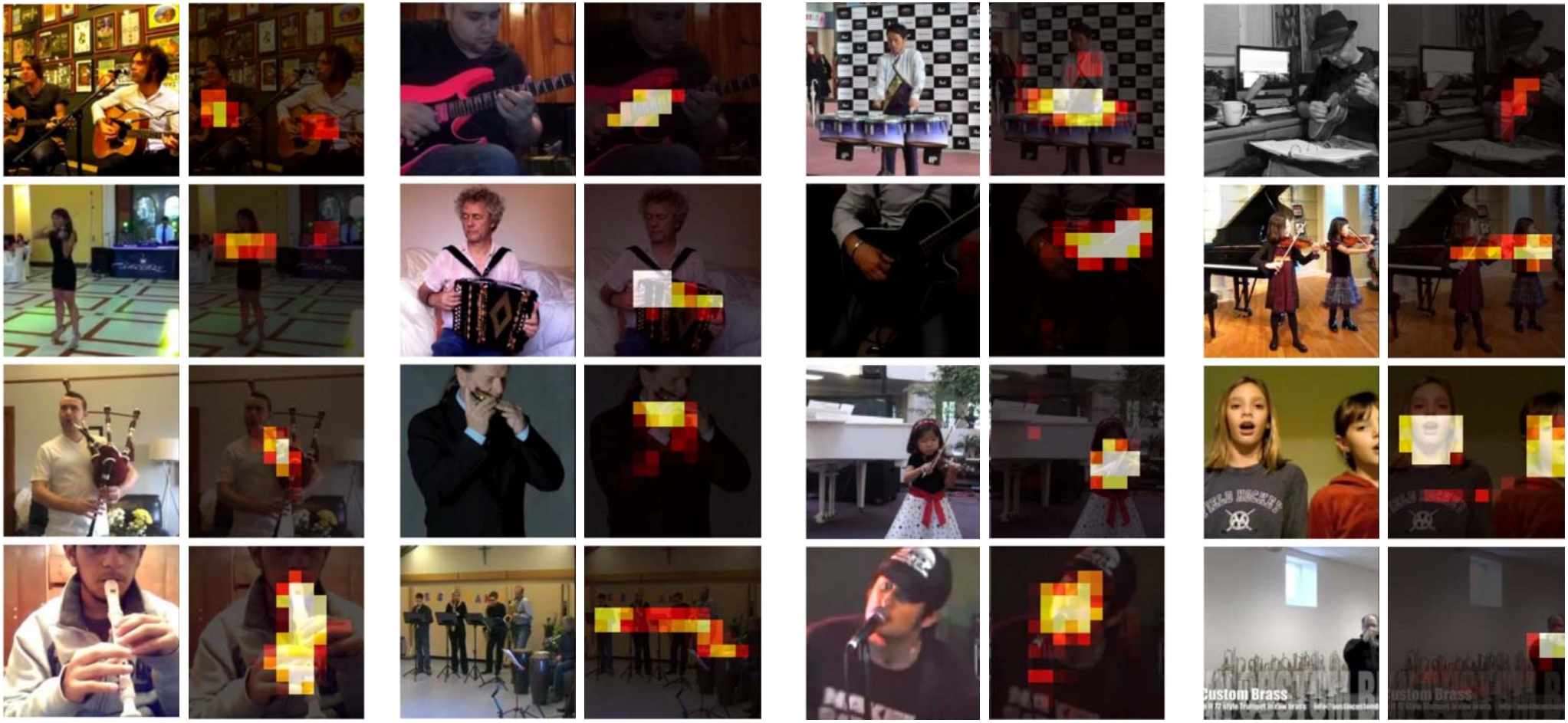


Audio-Visual Object Localization Network Structure

- Feature normalization is removed
- Audio subnetwork is the same
- Changes in visual subnetwork:
  - Pooling over conv4_2 removed to keep 14×14 resolution
  - To keep this resolution fc layers turned into 1×1 convolutions
- Scalar multiplication between visual and audio descriptors
- Network trained similar to AVC task:
  whether video and audio correspond

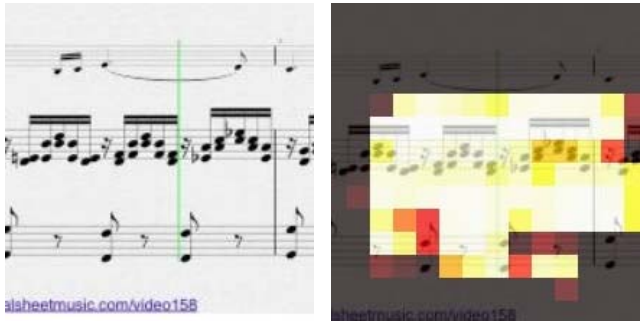


What is making the sound?



What is making the sound? - Failures

Likely from the noise of data



Detection of salient parts



Fails to detect the objects

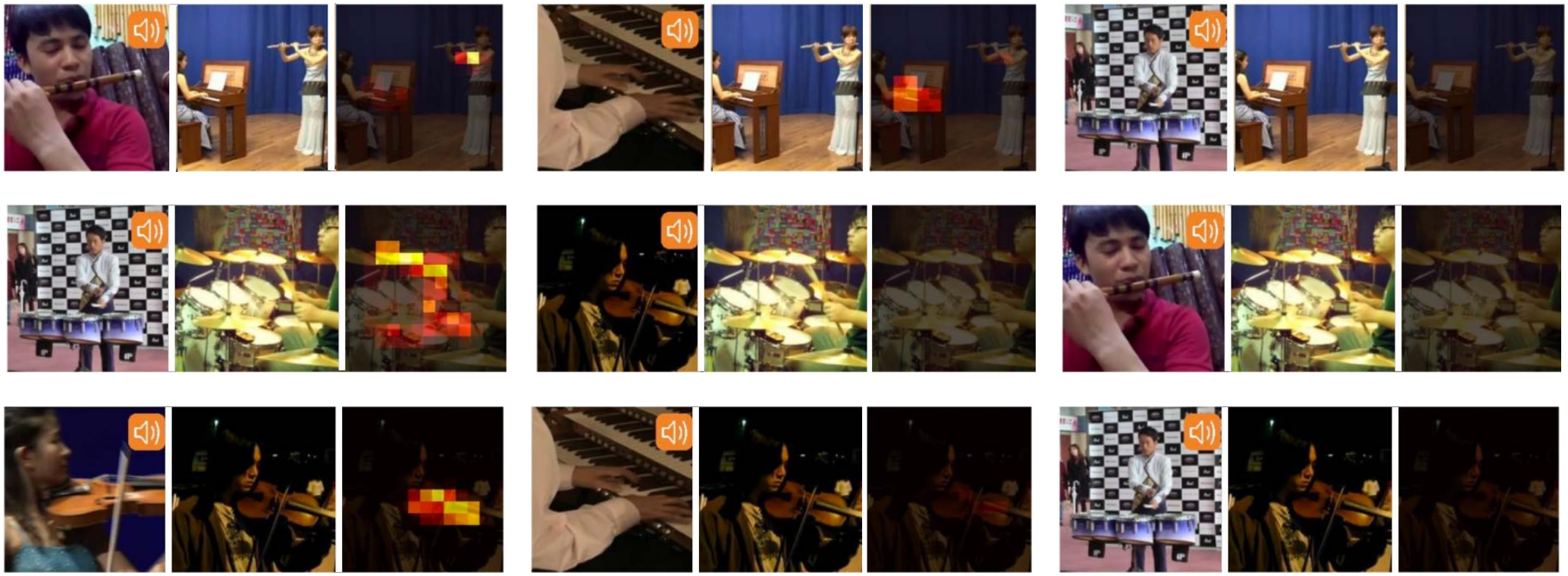








What would make this the sound?



What is making the sound?



What is making the sound?

Objects that Sound

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

This video shows that if the audio does not match the image, then the heatmap has a very low score, i.e. the heatmap is not simply a saliency map.

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

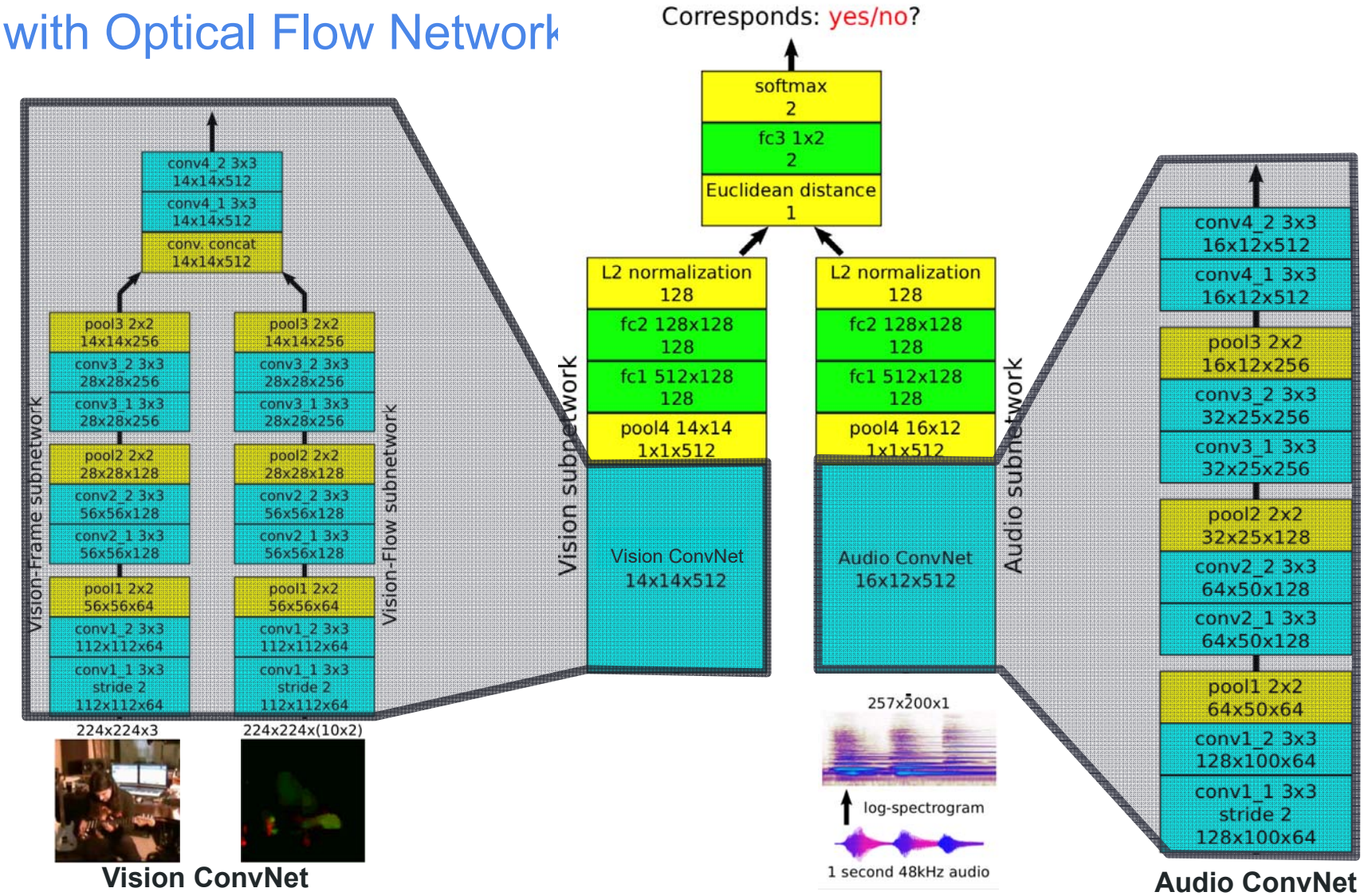
Input single frame

Frame/ Localization overlaid

Localization



AVE with Optical Flow Network



nDCG@30 Cross-modal and Intra-model Retrieval Comparison

- The performance of the AVE+MF and AVE+OF networks on the AVC task increased from 81.9% to 84.7% and 84.9%, respectively.
- No effect on retrieval performance

Method	im-im	im-aud	aud-im	aud-aud
Random	.407	.407	.407	.407
L^3-Net	.567	.418	.385	~.653
L^3-Net with CCA	.578	.531	.560	.649
VGG16-ImageNet	.600	-	-	-
VGG16-ImageNet + L^3-Audio CCA	.493	.458	.464	.618
AVE-Net	.604	**.561**	.587	**.665**
AVE+OF-Net	**.608**	.558	**.588**	**.665**

Conclusion

- Audio visual correspondence task enables:
  - Learning embedding for Cross-modal retrieval
  - Semantic based localization of objects that sound
- Same task with appropriate network design enables new functionalities to be learnt