中國人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Self-supervised Audiovisual Learning

Di Hu
Gaoling School of Artificial Intelligence
Renmin University of China

Email: dihu@ruc.edu.cn
19-06-2021

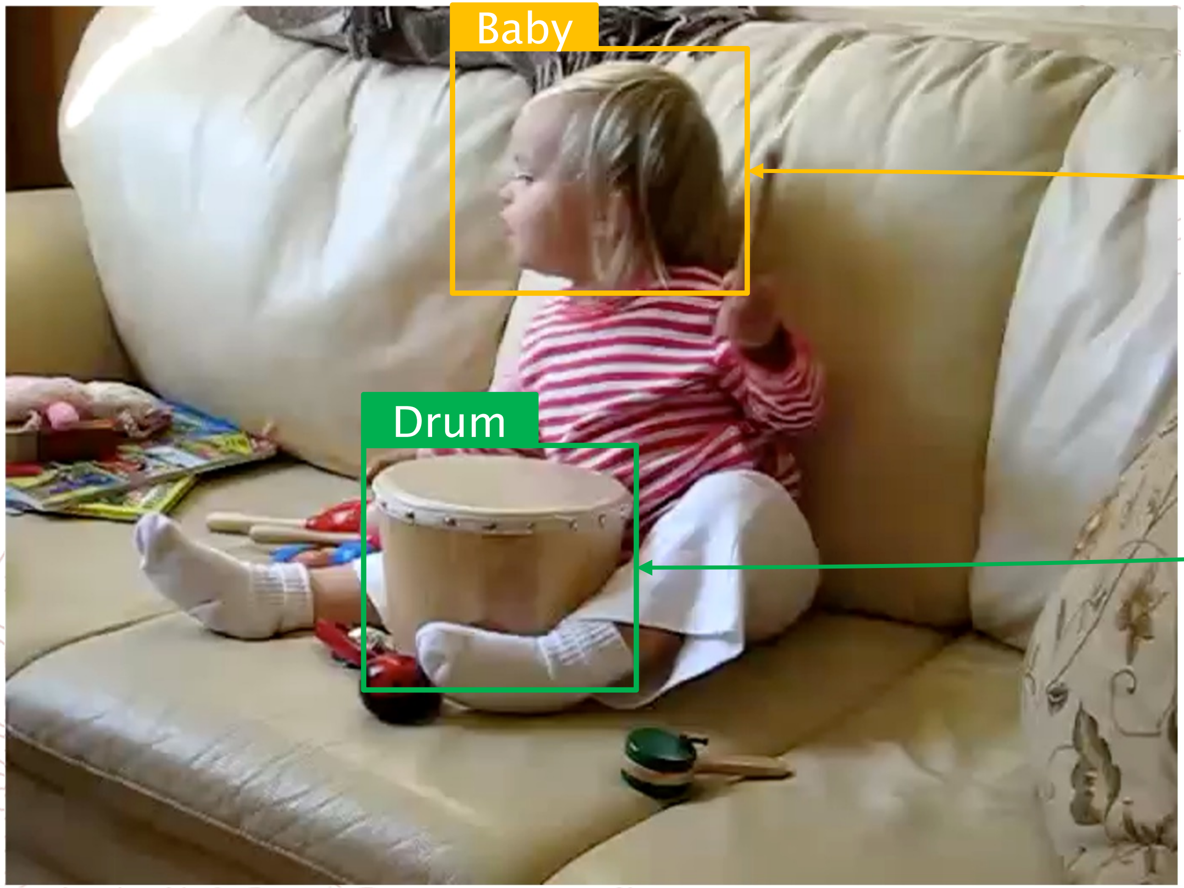
Outline

- Topic Overview
  - What is and why using self-supervised audiovisual learning
- Approaches Overview
  - What are the state of the art approaches and what is the inherent self-supervision
- Summary
  - What are the core challenges and future directions

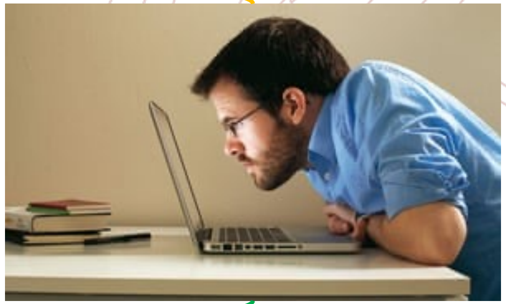
Outline

- Topic Overview
  - What is and why using self-supervised audiovisual learning
- Approaches Overview
  - What are the state of the art approaches and what is the inherent self-supervision
- Summary
  - What are the core challenges and future directions

What is self-supervised audiovisual learning



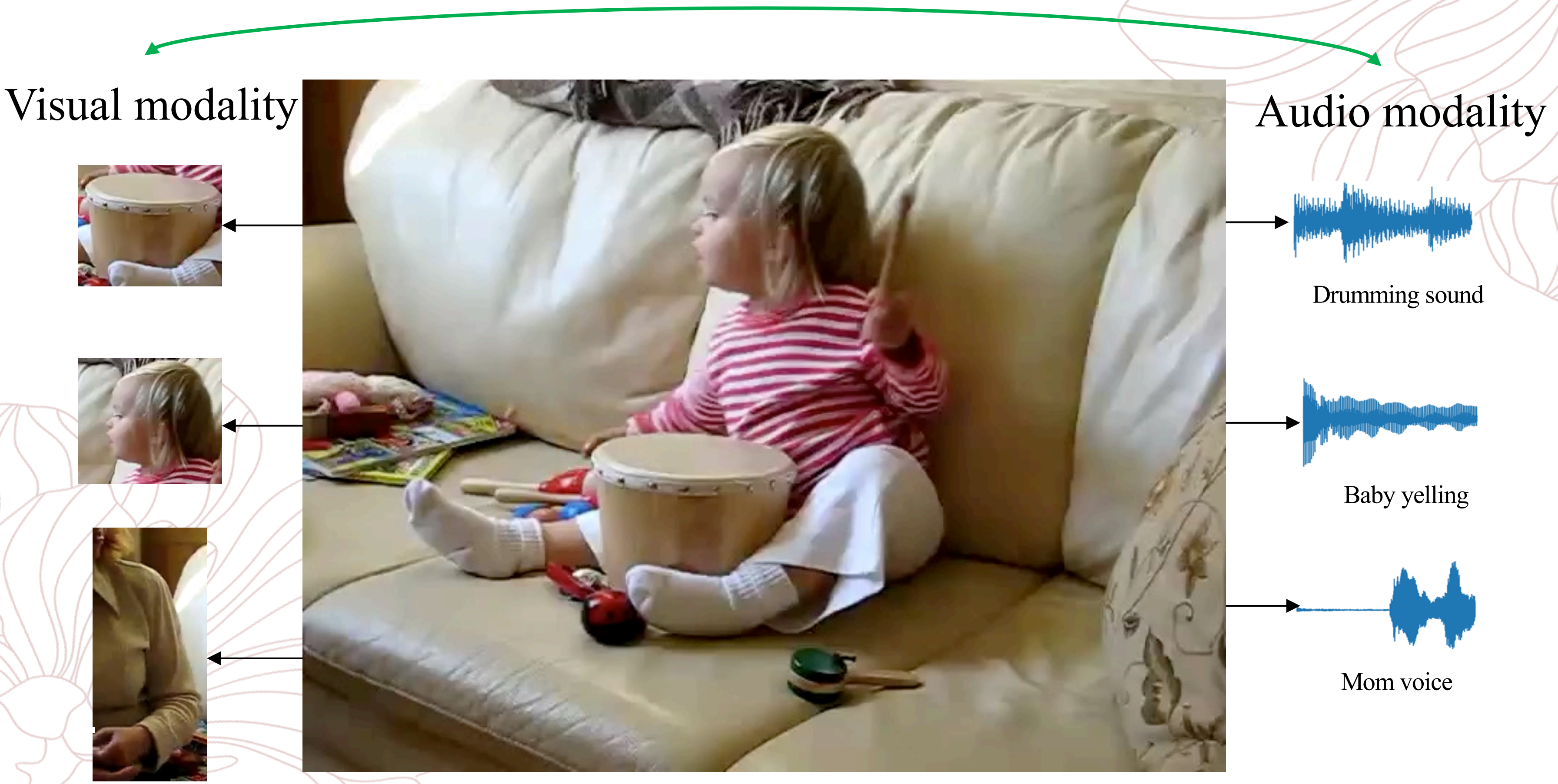
Learning from annotations



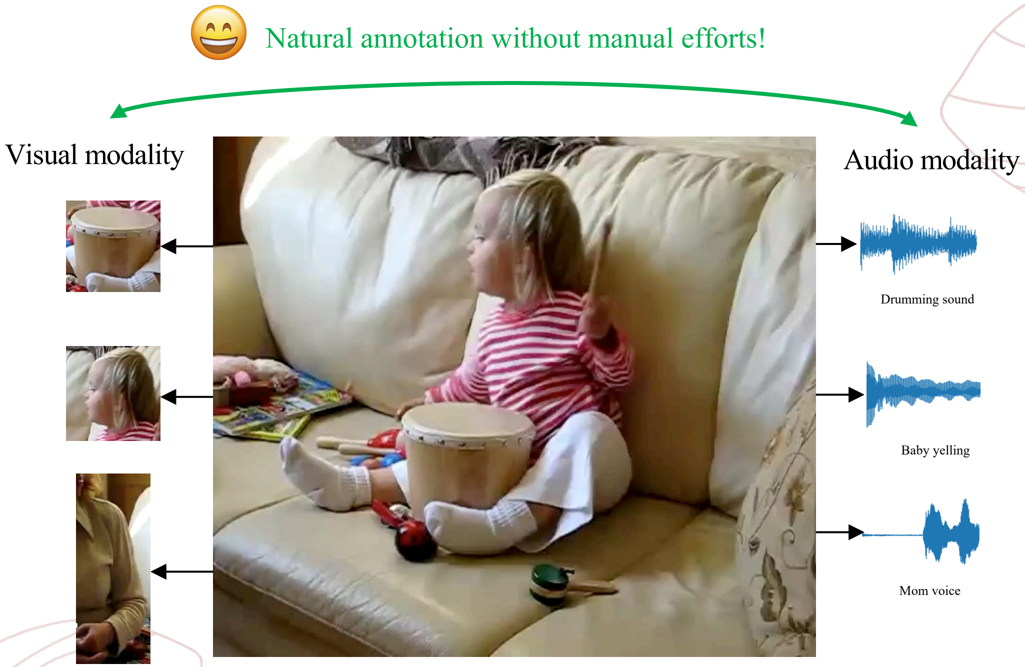
But this is not the real-world!

What is self-supervised audiovisual learning

Natural annotation without manual efforts!



What is self-supervised audiovisual learning



Unlike the conventional unimodal case, video is a rich source of audio and visual modalities, where the correlation between modalities can be used as a supervisory signal for self-supervised learning.

Why using self-supervised audiovisual learning

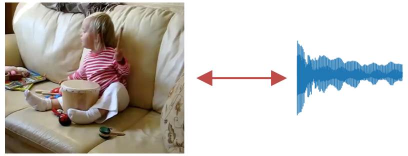
Sound is produced by the oscillation of object！

Free！ **Reliable！** **Pervasive！**

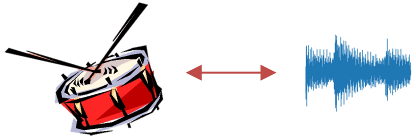
Approaches Overview

<2017 | 2017-2020 | 2020-Present

Cross-modal knowledge transfer



Self-supervised audiovisual scene learning



Self-supervised audiovisual object perception

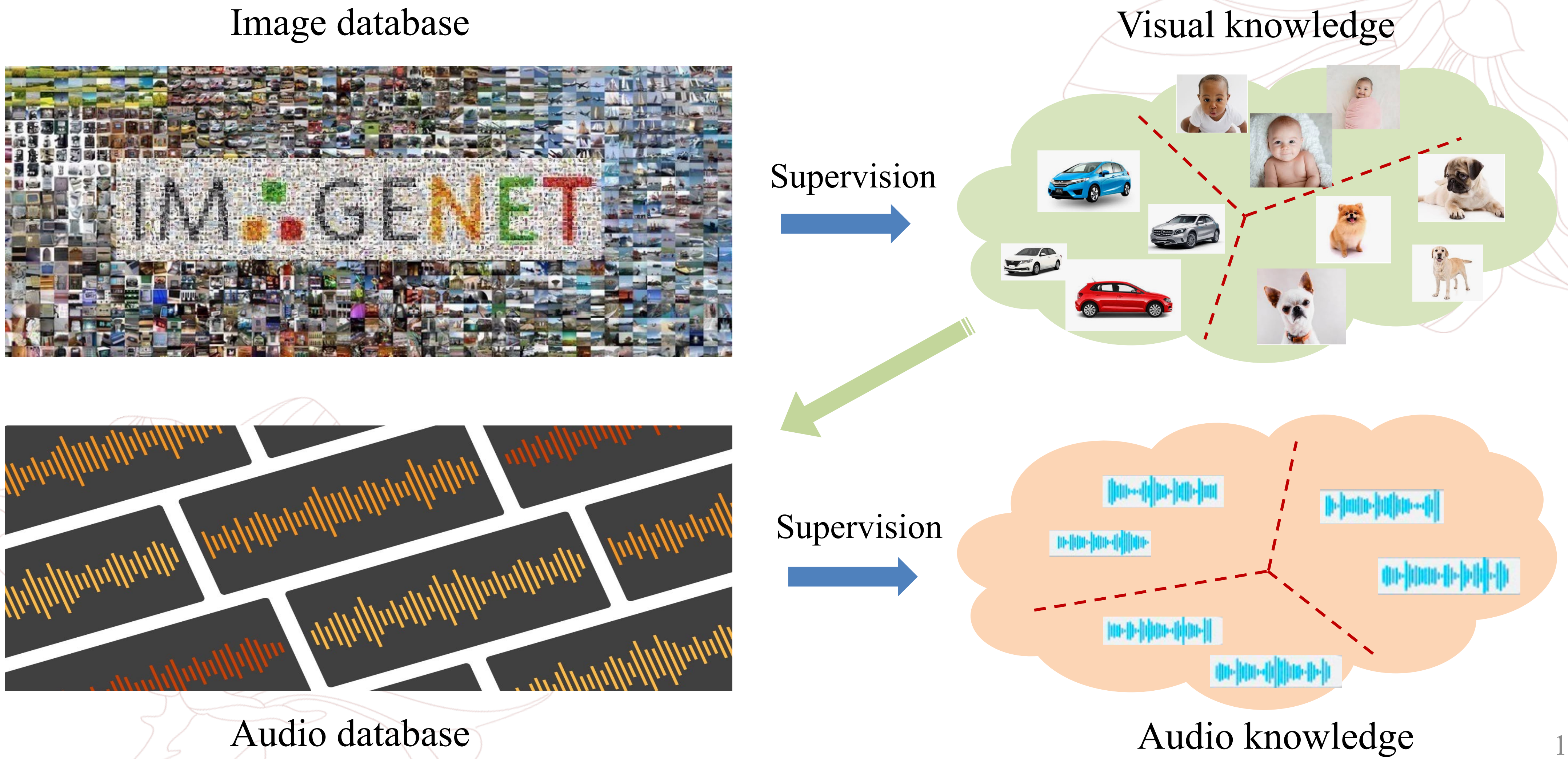
Outline

- Topic Overview
  - What is and why using self-supervised audiovisual learning
- Approaches Overview
  - What are the state of the art approaches and what is the inherent self-supervision
- Summary
  - What are the core challenges and future directions

Approaches Overview

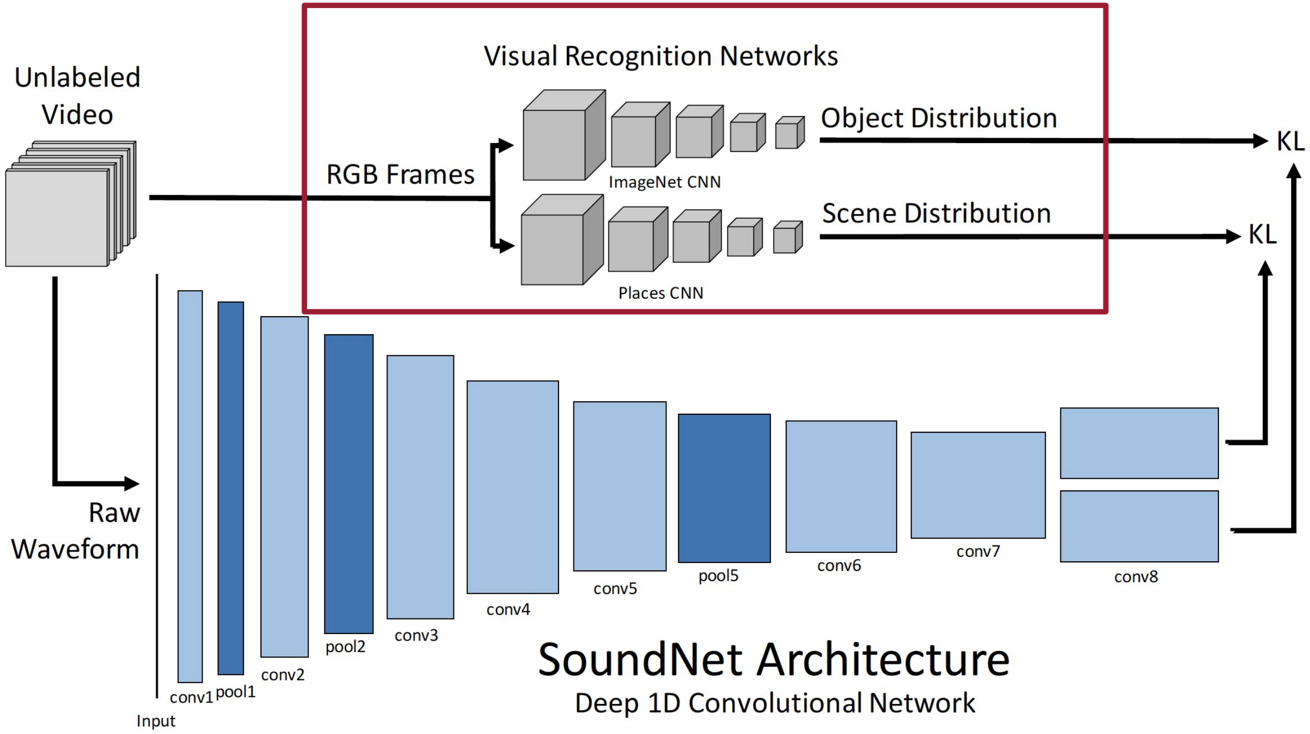
- Cross-modal knowledge transfer
- Self-supervised audiovisual scene learning
- Self-supervised audiovisual object perception

Cross-modal knowledge transfer

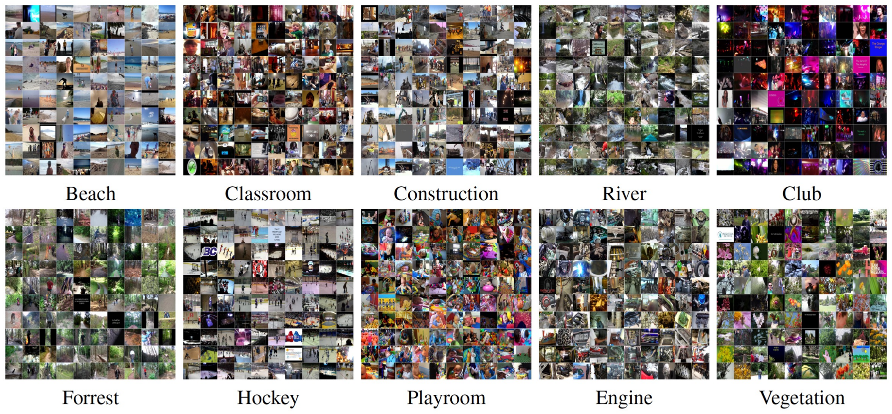


Cross-modal knowledge transfer

Visual knowledge



Unlabeled Video Dataset



Feature Evaluation

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

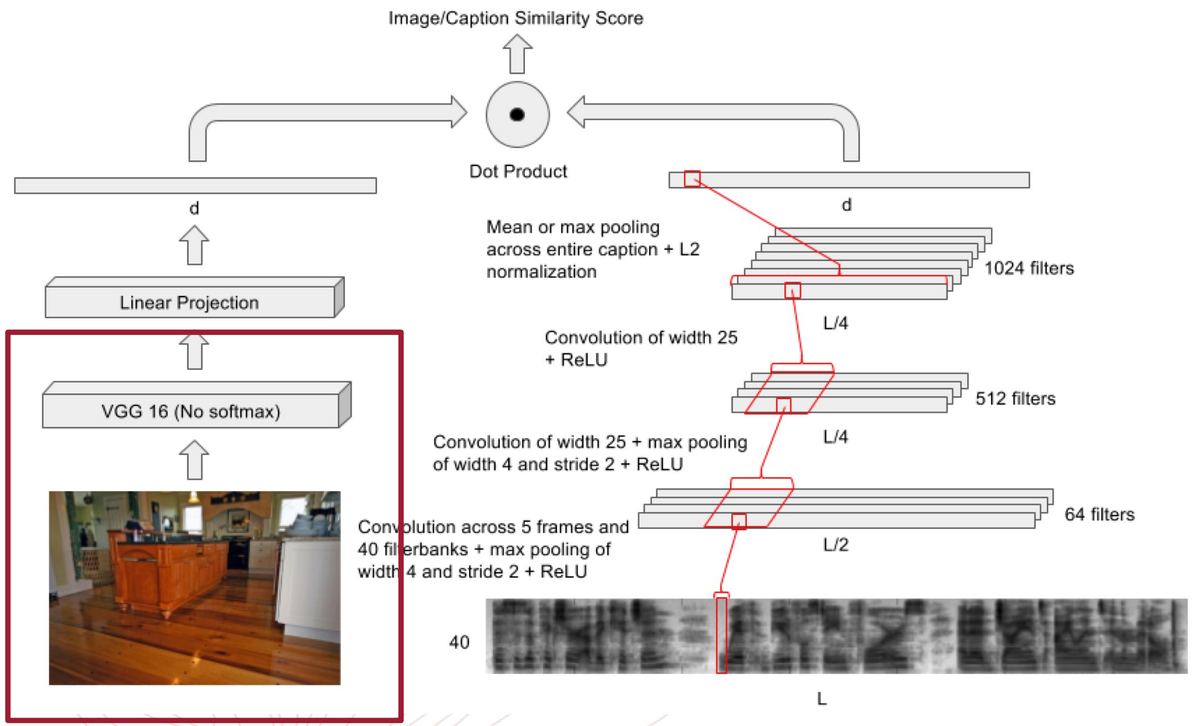
Table 3: **Acoustic Scene Classification on DCASE:** We evaluate classification accuracy on the DCASE dataset. By leveraging large amounts of unlabeled video, SoundNet generally outperforms hand-crafted features by 10%.

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.

[3] Y. Aytar, C. Vondrick, and A. Torralba, “SoundNet: Learning Sound Representations from Unlabeled Video,” *Advances in Neural Information Processing Systems*, 2016.

Cross-modal knowledge transfer



Visual knowledge

Image-spectrogram similarity

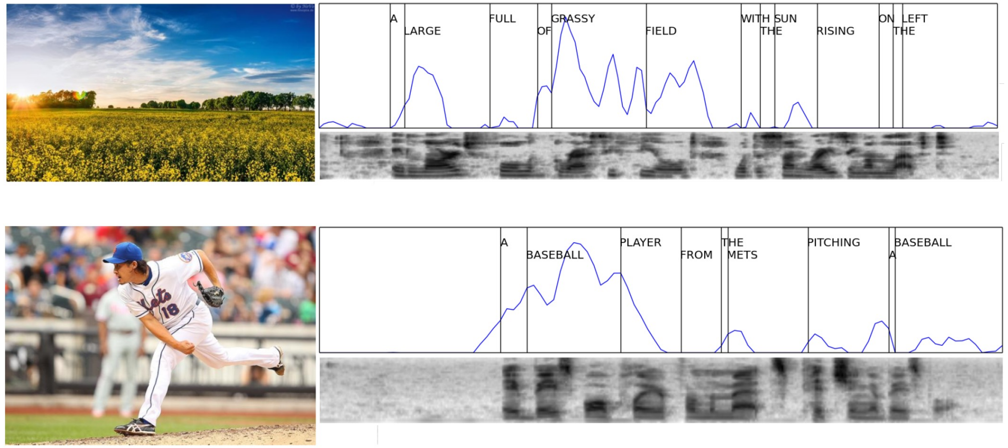


Image2sound retrieval

many cars are parked in the large parking lot there a large residential neighborhood with many apartment buildings

a sidewalk in front of the building there are bushes and a car parked

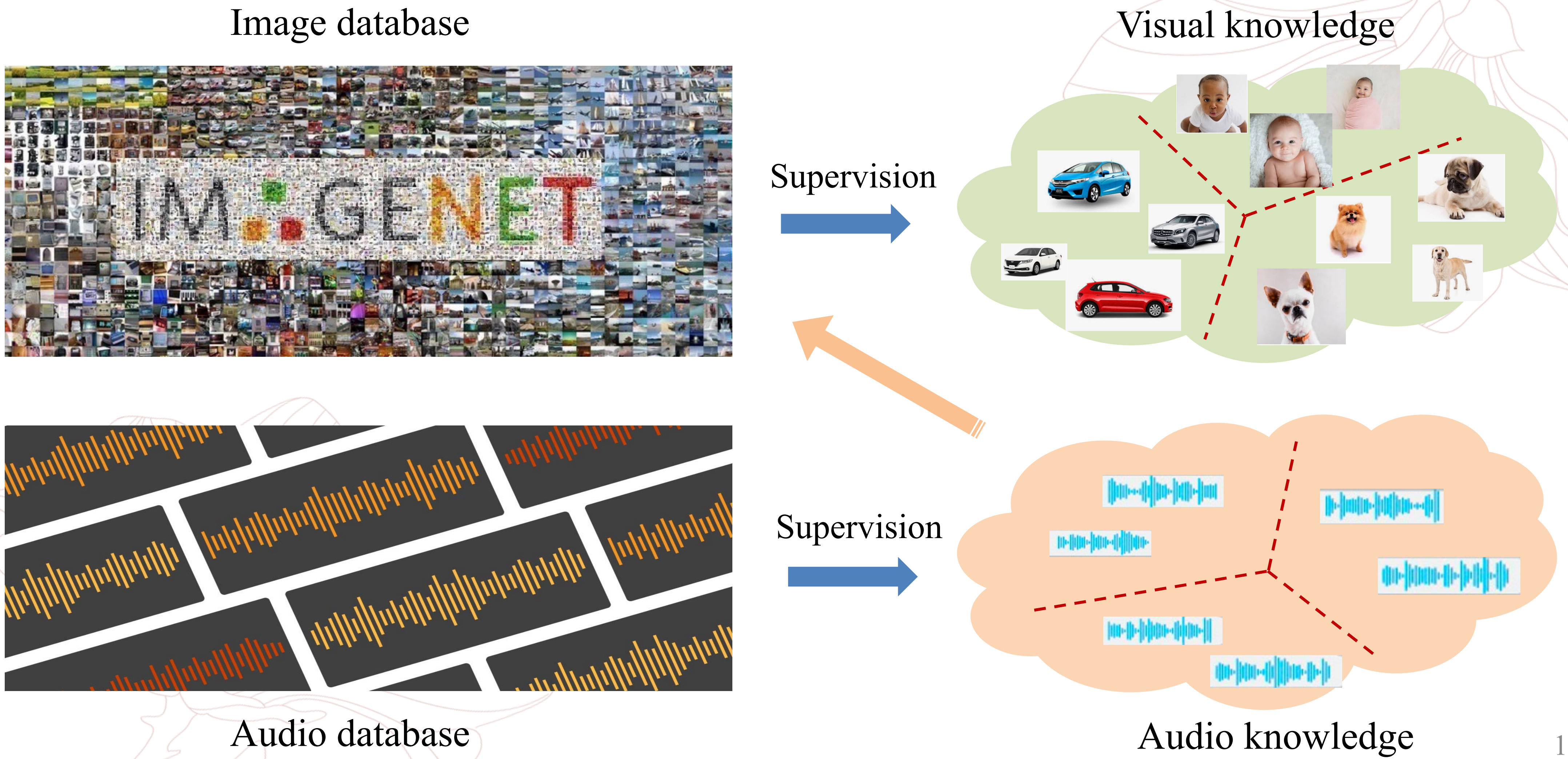
several green trees along a street with many parked cars

three cars are parked next to each other there's tar everywhere

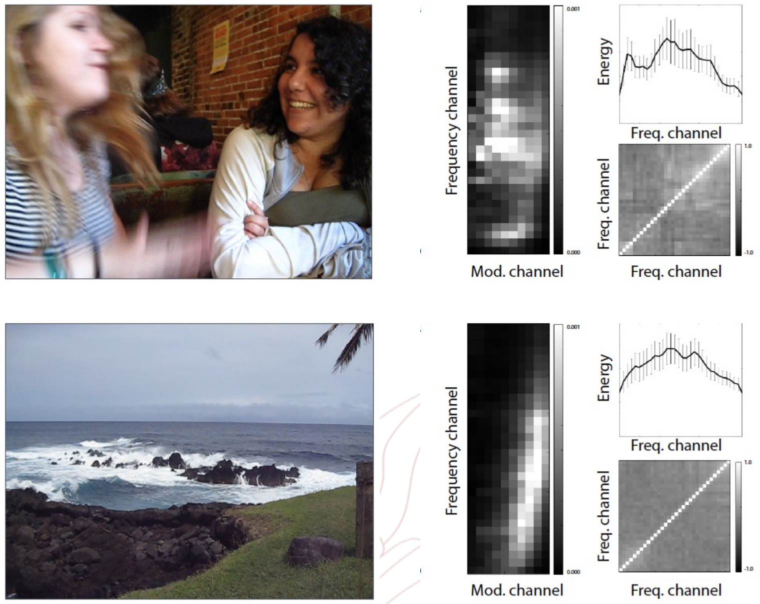
car one on down the line in a factory assign sale and stop the first <spoken_noise> is

[4] D. Harwath, A. Torralba, and J. Glass, “Unsupervised Learning of Spoken Language with Visual Context,” *Advances in Neural Information Processing Systems*, 2016.
[5] D. Harwath and J. Glass, “Learning word-like units from joint audio-visual analysis,” *Proc. Annual Meeting of the Association for Computational Linguistics,* 2017.

Cross-modal knowledge transfer

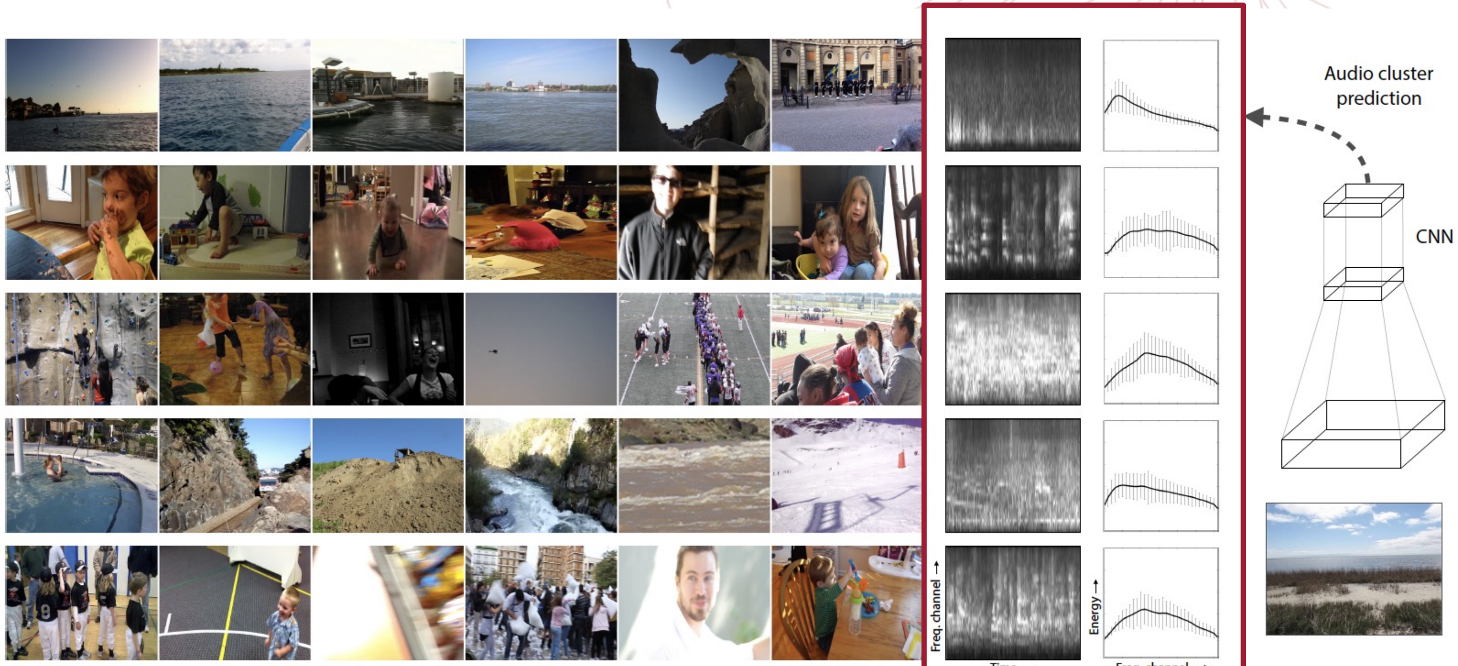


Cross-modal knowledge transfer



(a) Video frame (c) Summary statistics

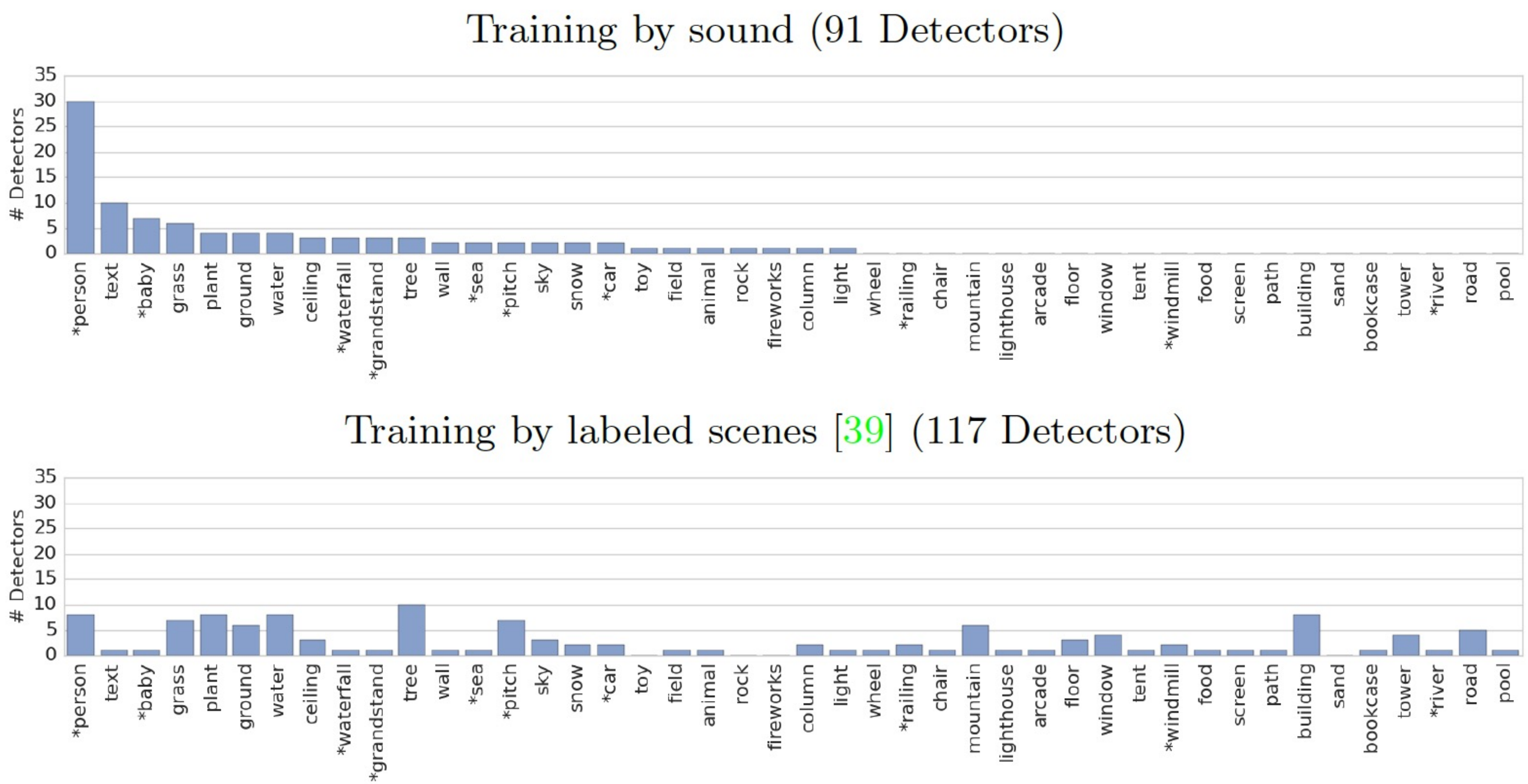
Audio knowledge



(a) Images grouped by audio cluster (b) Clustered audio stats. (c) CNN model

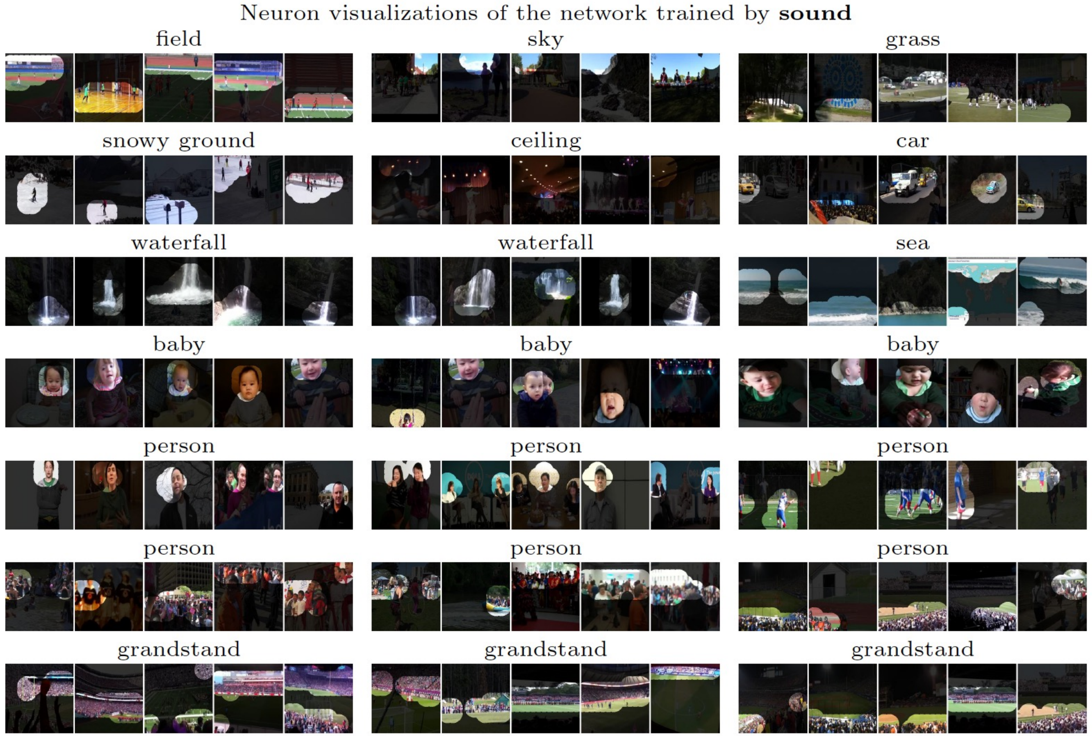
[6] A. Owens, J. Wu, J. McDermott, W. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," *Proc. European Conf. Computer Vision*, 2016.

Cross-modal knowledge transfer



[6] A. Owens, J. Wu, J. McDermott, W. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," *Proc. European Conf. Computer Vision*, 2016.

Cross-modal knowledge transfer



Method	VOC Cls. (%mAP)				SUN397 (%acc.)			
	max5	pool5	fc6	fc7	max5	pool5	fc6	fc7
Sound (cluster)	36.7	45.8	44.8	44.3	**17.3**	**22.9**	20.7	14.9
Sound (binary)	**39.4**	**46.7**	**47.1**	**47.4**	17.1	22.5	**21.3**	**21.4**
Sound (spect.)	35.8	44.0	44.4	44.4	14.6	19.5	18.6	17.7
Texton-CNN	28.9	37.5	35.3	32.5	10.7	15.2	11.4	7.6
K-means [20]	27.5	34.8	33.9	32.1	11.6	14.9	12.8	12.4
Tracking [35]	33.5	42.2	42.4	40.2	14.1	18.7	16.2	15.1
Patch pos. [4]	27.7	**46.7**	-	-	10.0	22.4	-	-
Egomotion [1]	22.7	31.1	-	-	9.1	11.3	-	-
ImageNet [21]	**63.6**	**65.6**	**69.6**	**73.6**	29.8	34.0	37.8	37.8
Places [39]	59.0	63.2	65.3	66.2	**39.4**	**42.1**	**46.1**	**48.8**

(a) Image classification with linear SVM

Method	(%mAP)
Random init. [20]	41.3
Sound (cluster)	44.1
Sound (binary)	43.3
Motion [35,20]	47.4
Egomotion [1,20]	41.8
Patch pos. [4,20]	46.6
Calib. + Patch [4,20]	**51.1**
ImageNet [21]	**57.1**
Places [39]	52.8

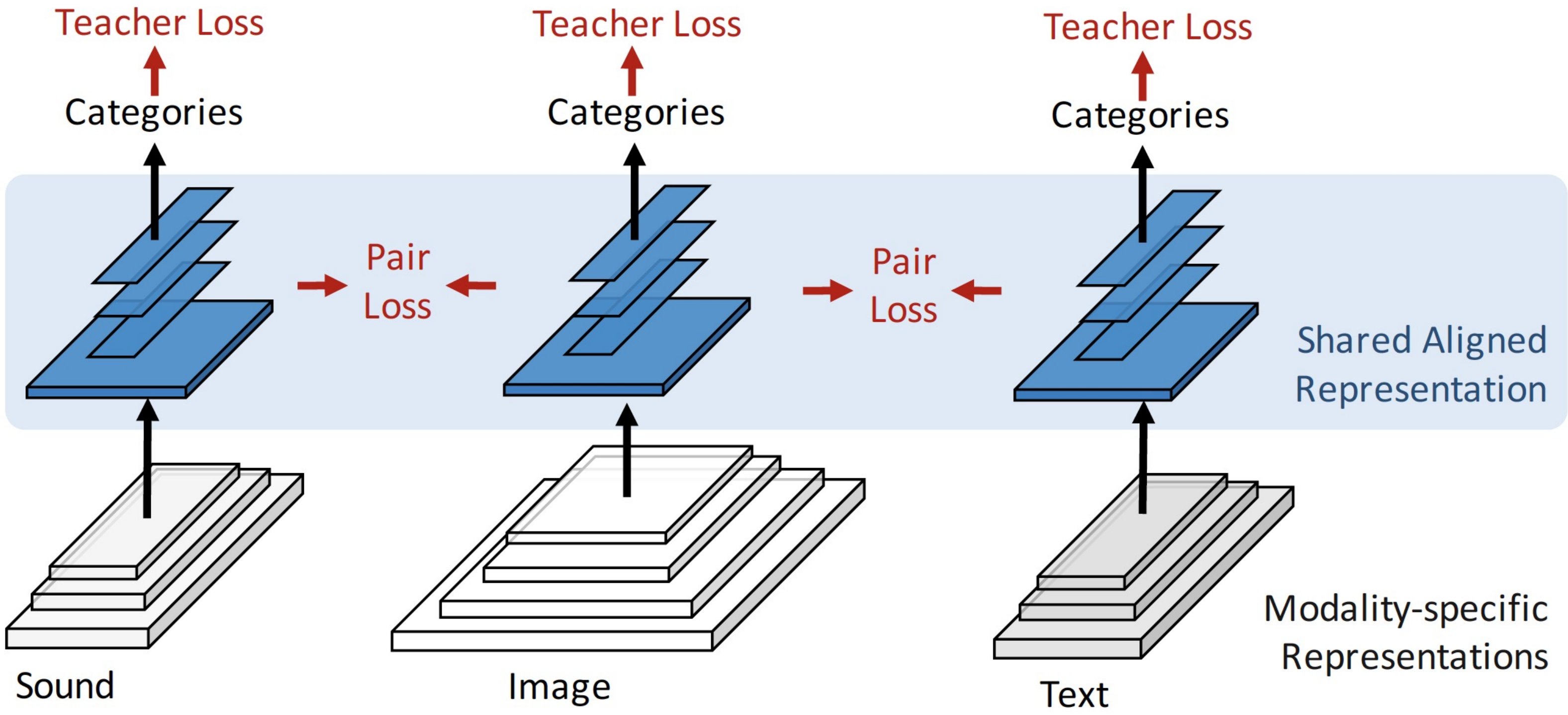
(b) Finetuning detection

Method	aer	bk	brd	bt	btl	bus	car	cat	chr	cow	din	dog	hrs	mbk	prs	pot	shp	sfa	trn	tv
Sound (cluster)	68	**47**	38	54	15	45	**66**	45	42	23	37	28	73	58	**85**	25	26	32	67	42
Sound (binary)	69	45	38	56	**16**	**47**	65	45	41	**25**	37	28	**74**	**61**	**85**	26	**39**	32	**69**	38
Sound (spect.)	65	40	35	54	14	42	63	41	39	24	32	25	72	56	81	**27**	33	28	65	40
Texton-CNN	65	35	28	46	11	31	63	30	41	17	28	23	64	51	74	9	19	33	54	30
K-means	61	31	27	49	9	27	58	34	36	12	25	21	64	38	70	18	14	25	51	25
Motion [35]	67	35	41	54	11	35	62	35	39	21	30	26	70	53	78	22	32	37	61	34
Patches [4]	**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
ImageNet [21]	79	**71**	**73**	75	**25**	60	80	**75**	51	**45**	60	**70**	**80**	**72**	**91**	42	**62**	56	82	62
Places [39]	**83**	60	56	**80**	23	**66**	**84**	54	**57**	40	**74**	41	**80**	68	90	**50**	45	**61**	**88**	**63**

(c) Per class mAP for image classification on PASCAL VOC 2007

[6] A. Owens, J. Wu, J. McDermott, W. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," *Proc. European Conf. Computer Vision*, 2016.

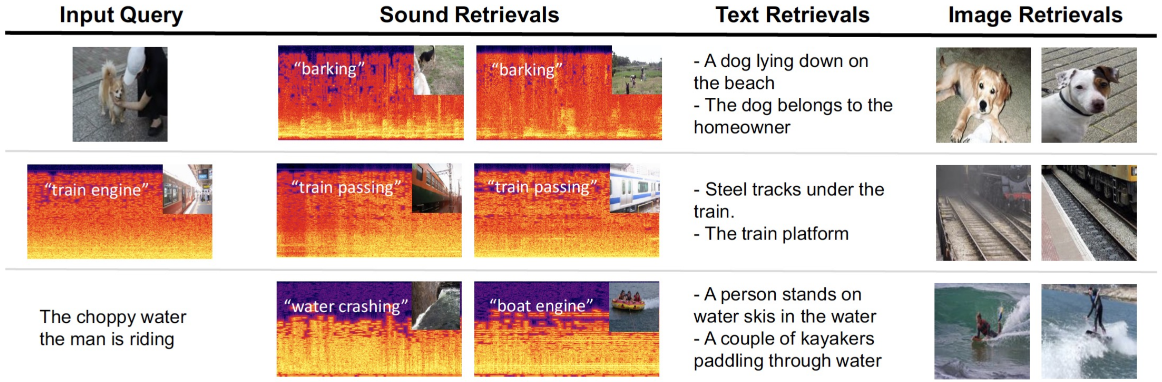
Cross-modal knowledge transfer



More general cross-modal transfer framework

[7] Y. Aytar, C. Vondrick, and A. Torralba, “See, hear, and read: Deep aligned representations,” arXiv:1706.00932, 2017.

Cross-modal knowledge transfer



Cross-modal retrieval

Semantic Units across modalities

[7] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," arXiv:1706.00932, 2017.

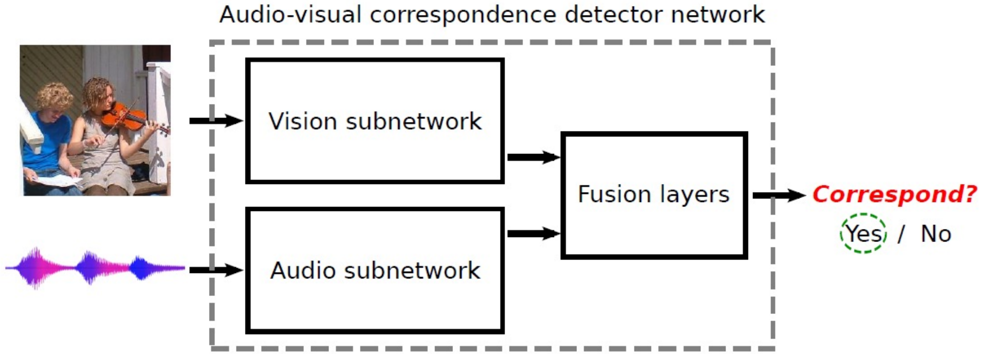
Approaches Overview

- Cross-modal knowledge transfer
- Self-supervised audiovisual scene learning
- Self-supervised audiovisual object perception

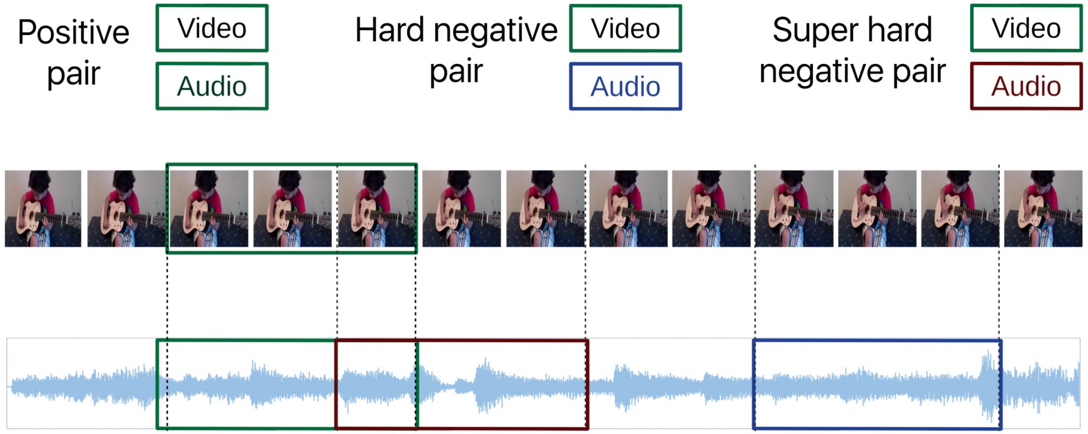
Self-supervised audiovisual scene learning

Without supervised pretraining

Visible scenes ⟷ Soundscape



Video-level Correspondence[1]

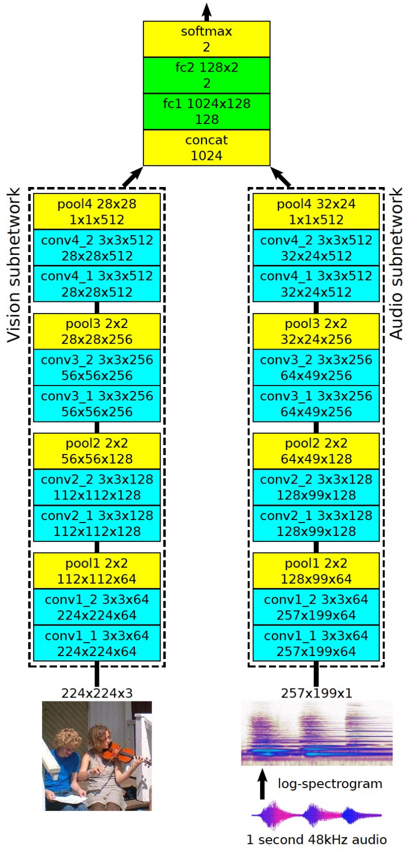


Temporal-level synchronization[2]

[1] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," *Proc. IEEE Conf. Computer Vision*, 2017.

[2] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems*, 2018.

Self-supervised audiovisual scene learning



Video-level Correspondence

Visual feature evaluation

Method	Top 1 accuracy
Random	18.3%
Pathak *et al.* [24]	22.3%
Krähenbühl *et al.* [16]	24.5%
Donahue *et al.* [7]	31.0%
Doersch *et al.* [6]	31.7%
Zhang *et al.* [36] (init: [16])	32.6%
Noroozi and Favaro [21]	34.7%
Ours random	12.9%
Ours	32.3%

Audio feature evaluation

(a) ESC-50

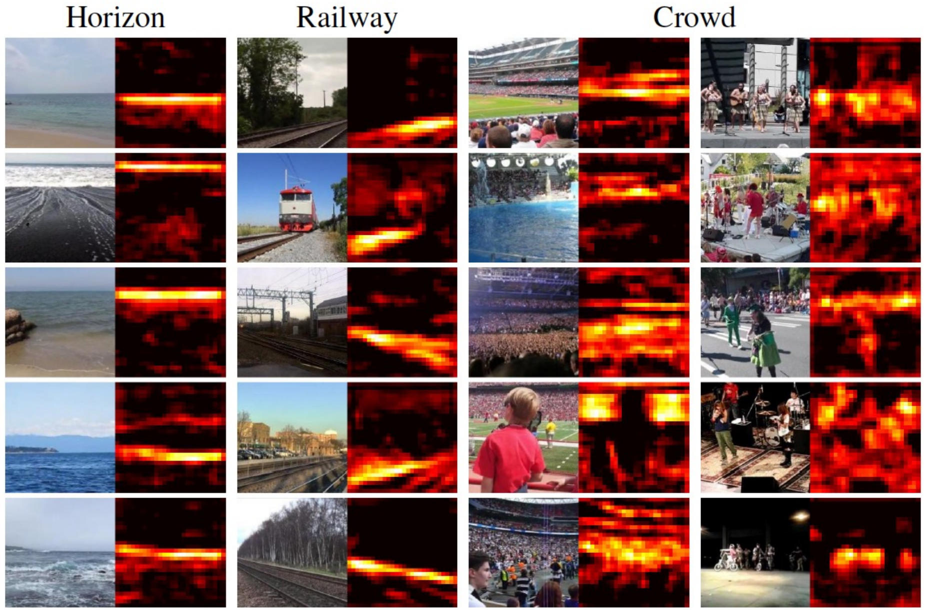
Method	Accuracy
SVM-MFCC [26]	39.6%
Autoencoder [2]	39.9%
Random Forest [26]	44.3%
Piczak ConvNet [25]	64.5%
SoundNet [2]	74.2%
Ours random	62.5%
Ours	**79.3%**
Human perf. [26]	*81.3%*

(b) DCASE

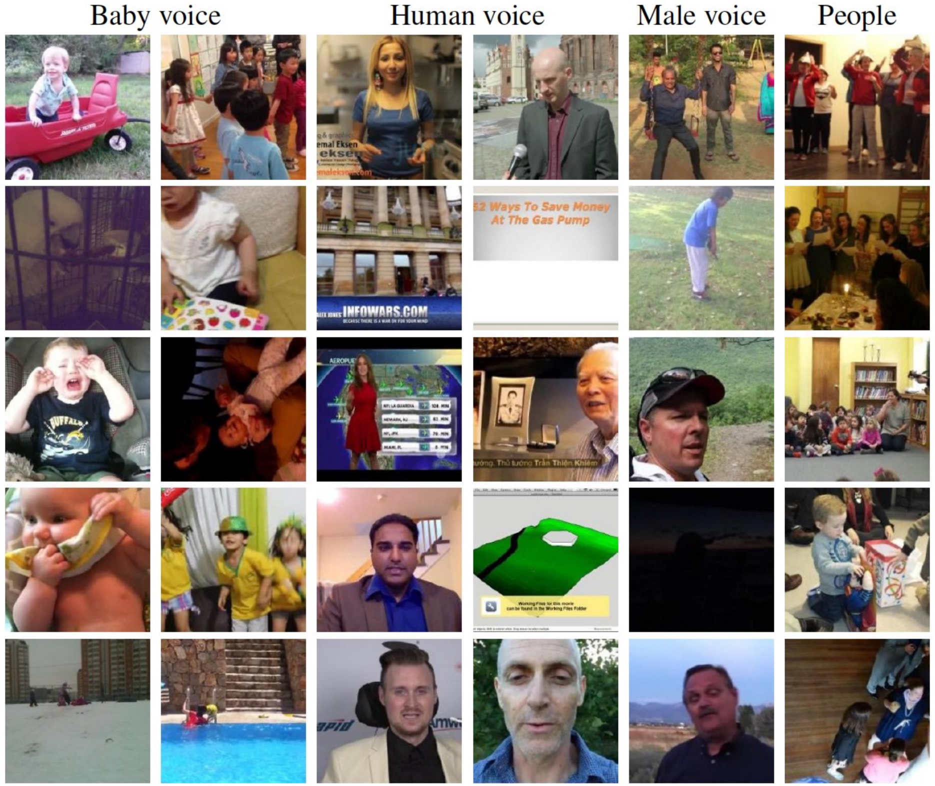
Method	Accuracy
RG [27]	69%
LTT [19]	72%
RNH [28]	77%
Ensemble [32]	78%
SoundNet [2]	88%
Ours random	85%
Ours	**93%**

[1] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," *Proc. IEEE Conf. Computer Vision*, 2017.

Self-supervised audiovisual scene learning



Semantic heatmaps of visual concepts



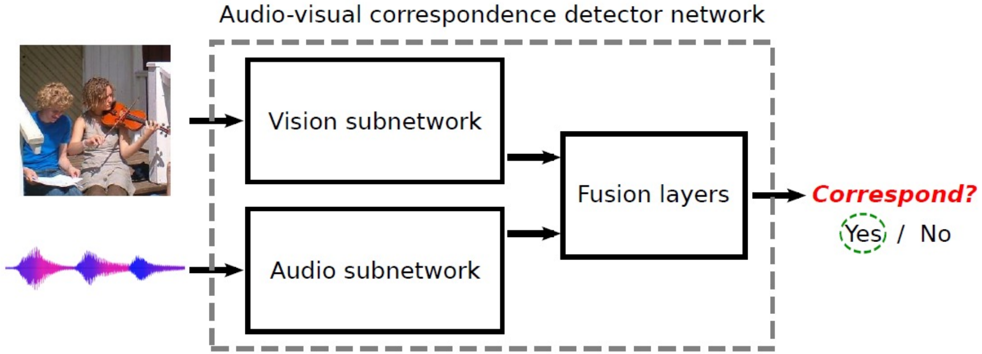
Audio concepts

[1] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," *Proc. IEEE Conf. Computer Vision*, 2017.

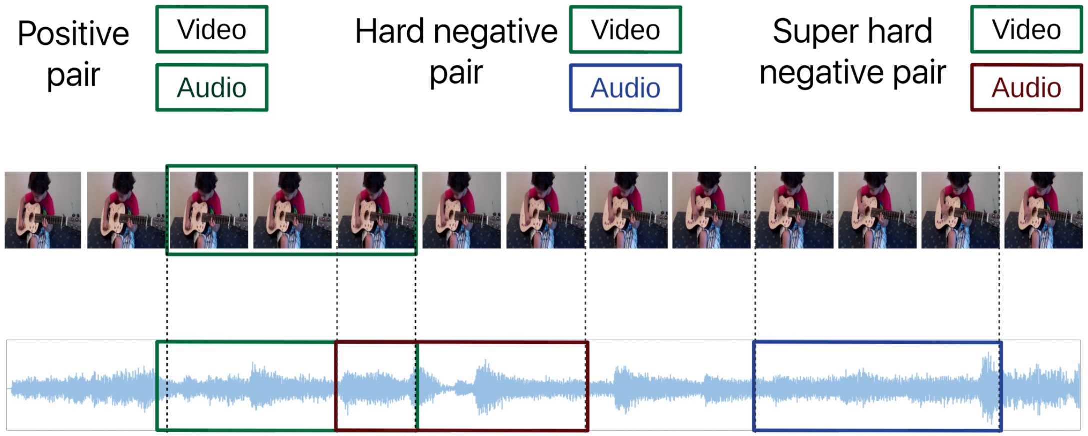
Self-supervised audiovisual scene learning

Visible scenes ⟷ Soundscape

Without supervised pretraining



Video-level Correspondence[1]

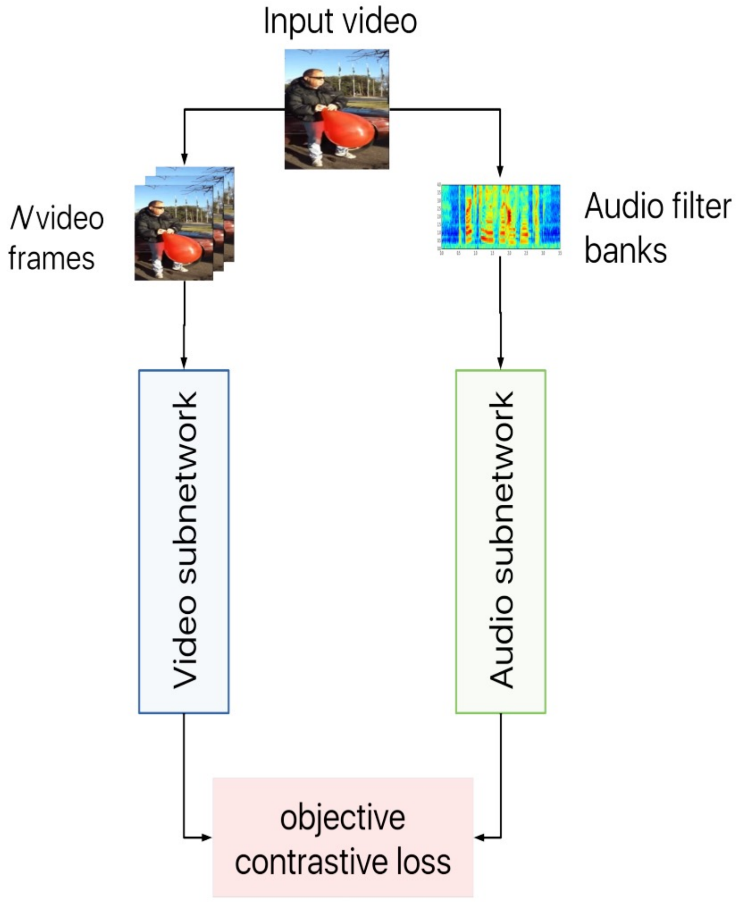


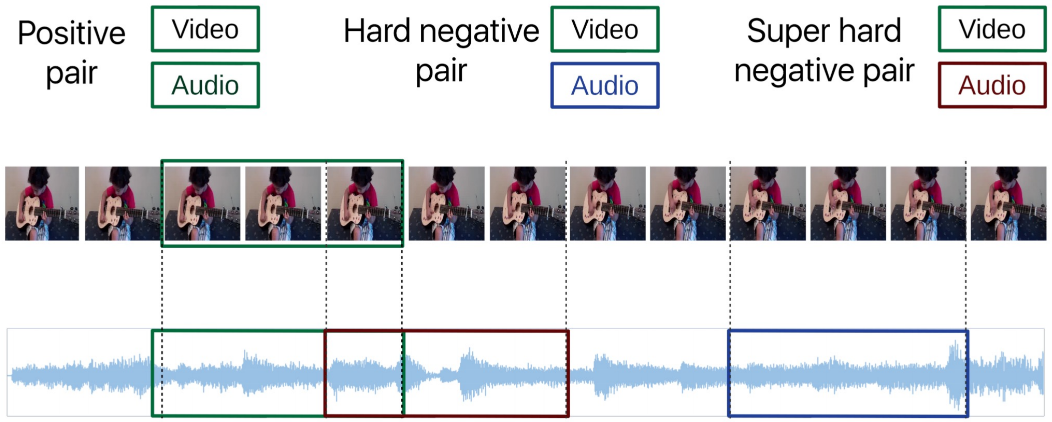
Temporal-level synchronization[2]

[1] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," *Proc. IEEE Conf. Computer Vision*, 2017.

[2] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems*, 2018.

Self-supervised audiovisual scene learning





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

[2] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems*, 2018.

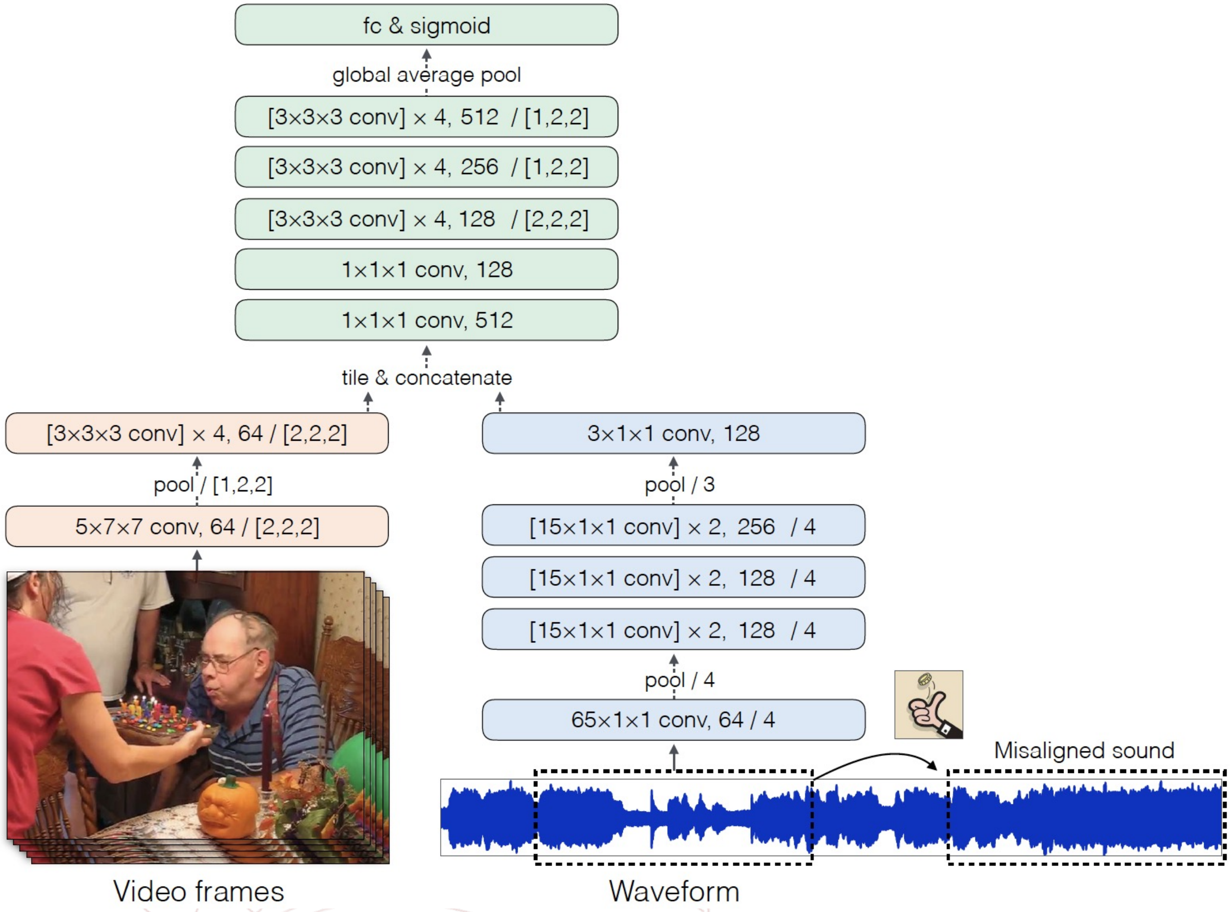
Self-supervised audiovisual scene learning

Sound is produced by the $\boxed{\text{oscillation}}$ of object！

Static image

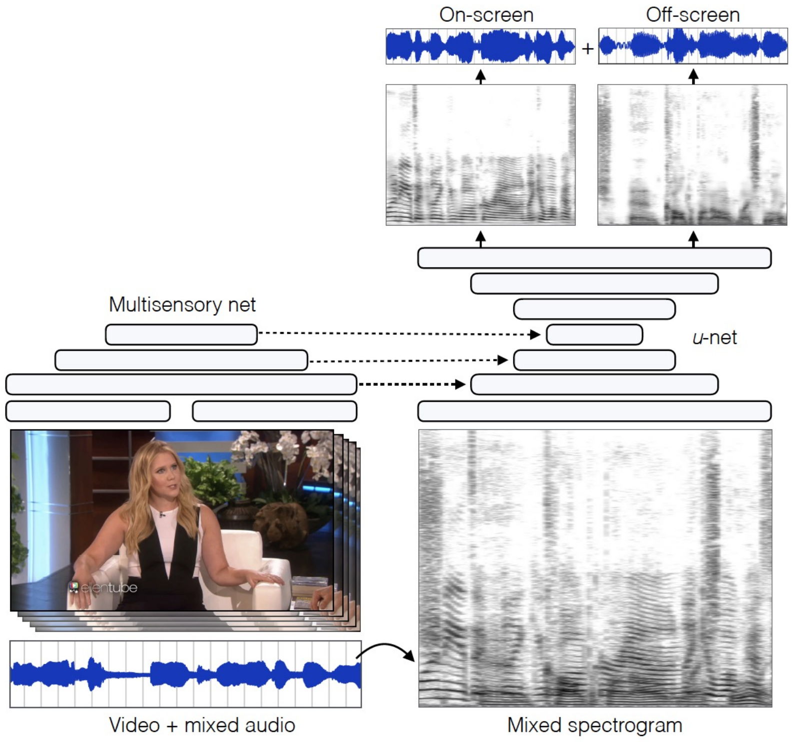
Temporal sequence

Self-supervised audiovisual scene learning



Temporal sequence

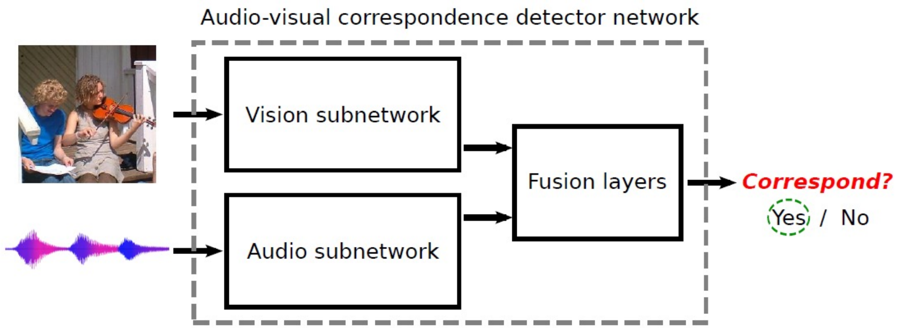
Temporal-level synchronization



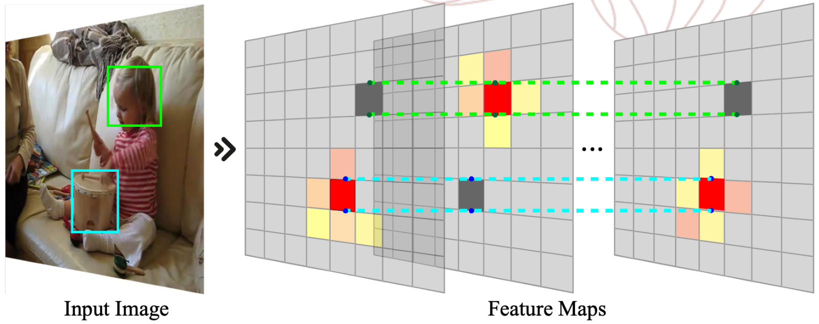
On-off screen speech separation

[8] A. Owens, and A.Efros, "Audio-visual scene analysis with self-supervised multisensory features," *Proc. European Conf. Computer Vision*, 2018.

Self-supervised audiovisual scene learning



- The unconstrained visual scene contains multiple sound-makers.
- The sound-maker does not always produce distinctive sound.
- The sound-maker may be even out of the screen



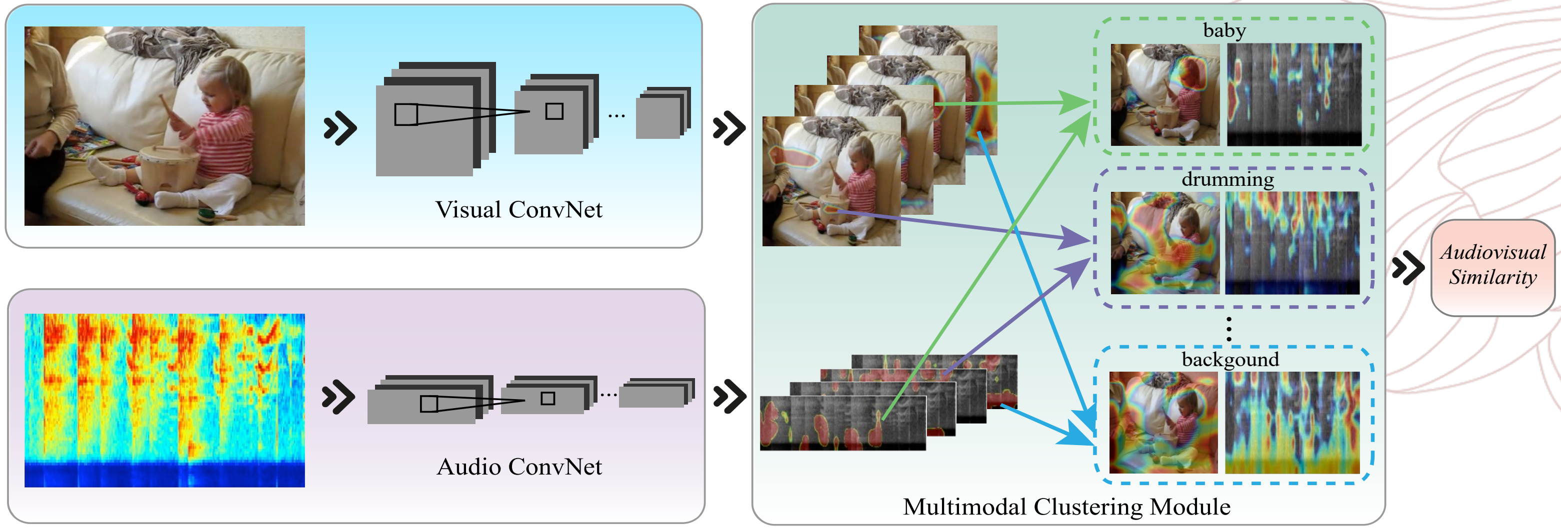
Modality Features: $\{u_1, u_2, ..., u_p | u_i \in R^n\}$

Representation Centers: $C = \{c_1, c_2, ..., c_k | c_i \in R^m\}$

Clustering Objective: $\mathcal{F}(C) = \sum_{i=1}^{p} \min_{j=1}^{k} d(u_i, c_j)$

[9] D. Hu, F. Nie, and X. Li, "Deep Multimodal Learning for Unsupervised Audiovisual Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

Self-supervised audiovisual scene learning



$$\mathcal{F}(C)=\sum_{i=1}^{p}\min_{j=1}^{k} d\left(u_i, c_j\right) \quad + \quad \min\left\{d_{i1}, d_{i2}, \ldots, d_{ik}\right\} \approx -\frac{1}{z}\log\left(\sum_{j=1}^{k} e^{-d_{ij}z}\right)$$

Shared Across Modalities

$$c_j^{(r+1)}=\sum_{i=1}^{n} s_{ij}^{(r)} u_i \quad \Rightarrow \quad c_j^{(r+1)}=\sum_{i=1}^{n} s_{ij}^{(r)} \boxed{W_j} u_i$$

Multimodal Clustering

Max-Margin Loss Function: $$loss=\sum_{i=1, i\neq j}^{p}\max\left(0, s\left(c_j^a, c_i^v\right)-s\left(c_i^a, c_i^v\right)+\Delta\right)$$

[9] D. Hu, F. Nie, and X. Li, "Deep Multimodal Learning for Unsupervised Audiovisual Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

Self-supervised audiovisual scene learning

Feature Evaluation

(a) ESC-50

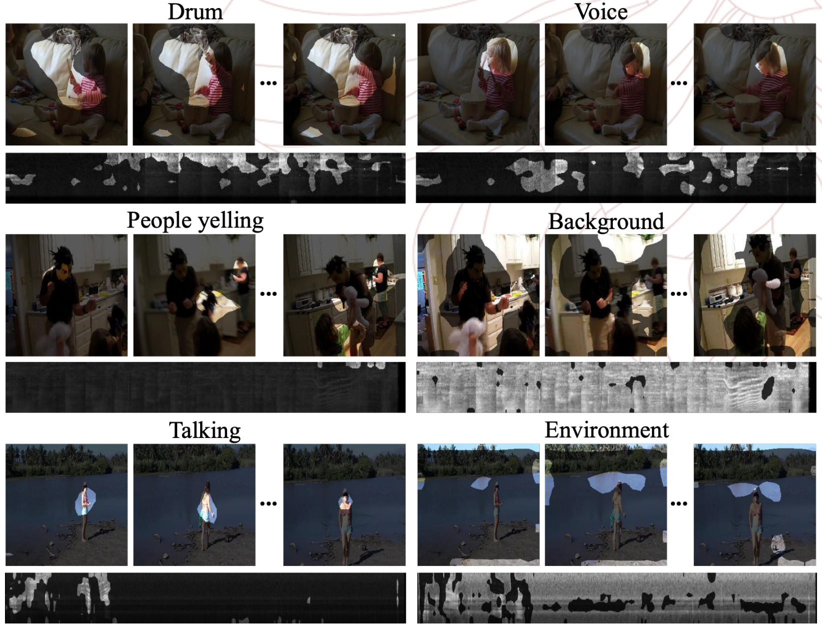
Methods	Accuracy
Autoencoder	0.399
Rand. Forest	0.443
ConvNet	0.645
SoundNet	0.742
L^3 [1]	0.761
†L^3[1]	0.793
†AVTS [2]	0.823
DMC	0.798
‡DMC	**0.826**
Human Perfor.	0.813

(b) Pascal VOC 2007

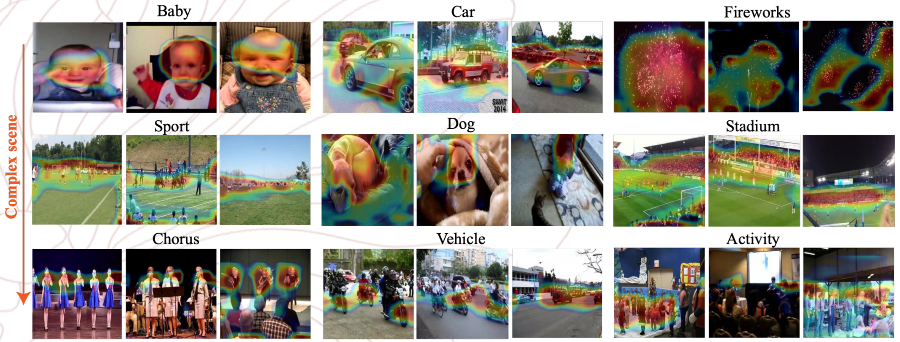
Methods	Accuracy
Taxton.	0.375
Kmeans	0.348
Tracking	0.422
Patch.	0.467
Egomotion	0.311
Sound(spe.)[3]	0.440
Sound(clu.) [3]	0.458
Sound(bia.) [3]	0.467
DMC	**0.514**
ImageNet	0.672

Evaluation of the Extracted A/V Features

Audiovisual Understanding



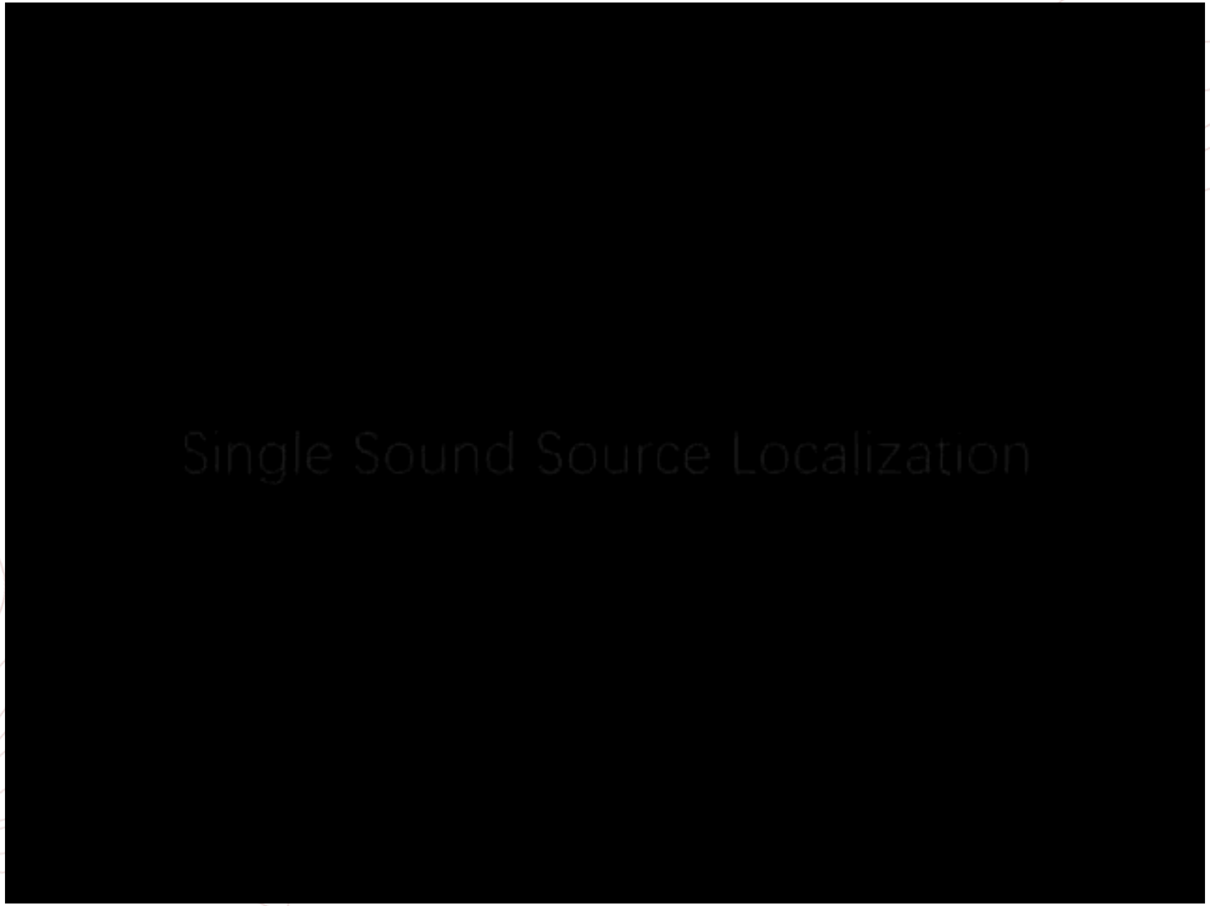
Sound Source Localization



Methods	cIoU(0.7)	AUC
Random	-	32.3
Unsupervised† [4]	-	51.2
Unsupervised [4]	∼18.8	55.8
Supervised [4]	∼25.5	60.3
Sup.+Unsup.[4]	∼28.8	62.0
DMC (unrelated)	5.2	21.1
DMC (related)	26.2	56.8

[9] D. Hu, F. Nie, and X. Li, “Deep Multimodal Learning for Unsupervised Audiovisual Learning,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

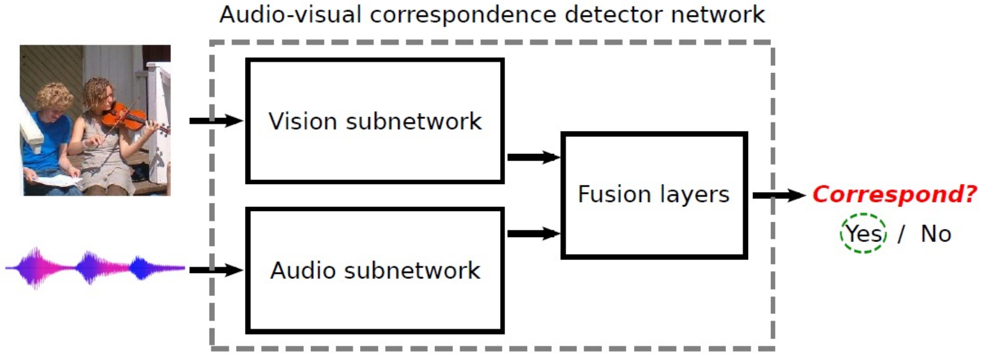
Self-supervised audiovisual scene learning



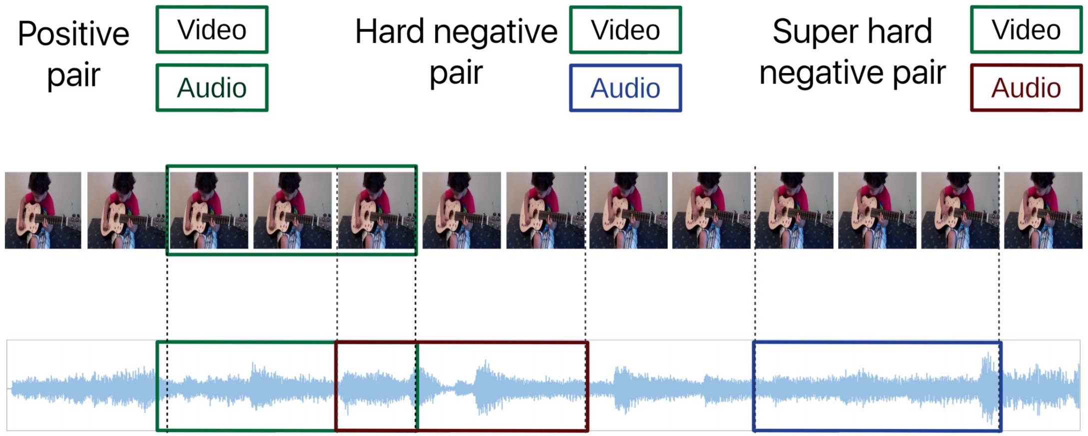
[9] D. Hu, F. Nie, and X. Li, “Deep Multimodal Learning for Unsupervised Audiovisual Learning,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

Self-supervised audiovisual scene learning

Visible scenes ⟷ Soundscape



Video-level Correspondence[1]



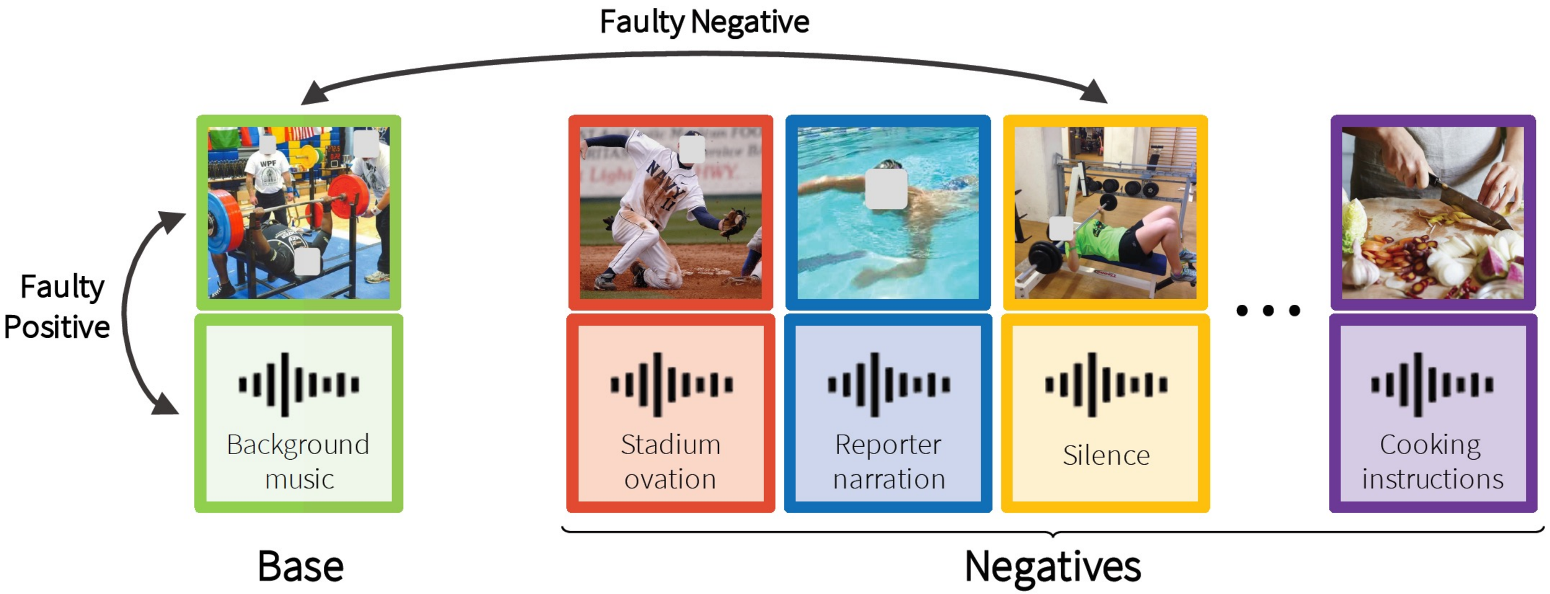
Temporal-level synchronization[2]

What if the audio and visual modality are not well-corresponding?

[1] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," *Proc. IEEE Conf. Computer Vision*, 2017.

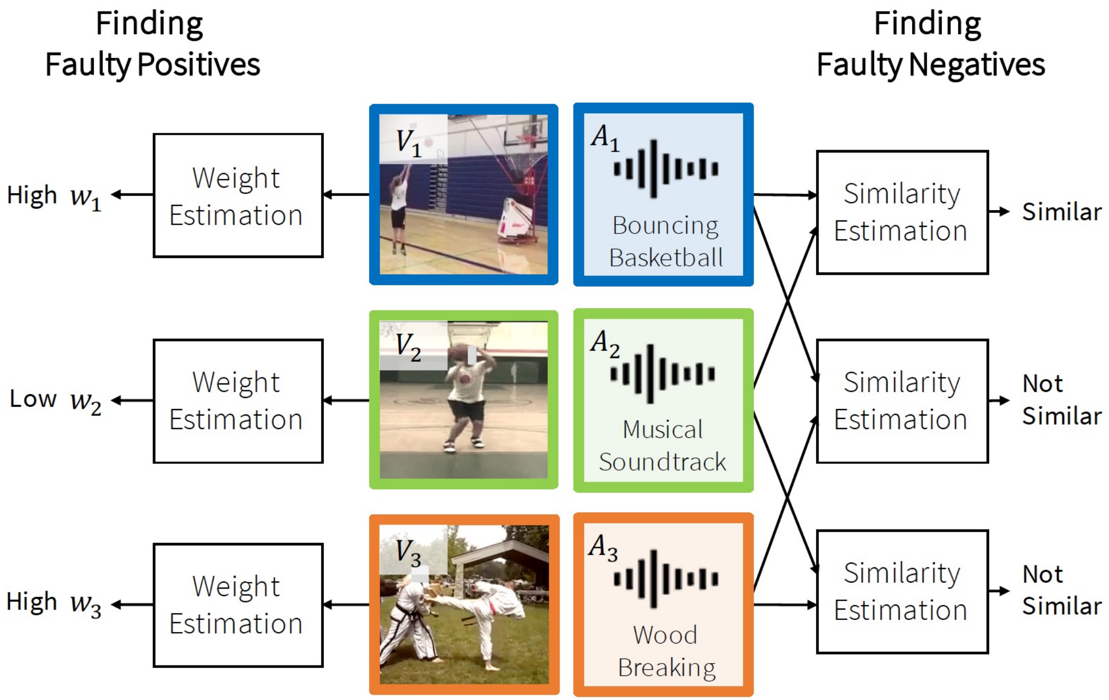
[2] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems*, 2018.

Self-supervised audiovisual scene learning



[10] P. Morgado, I. Misra, and N. Vasconcelos, "Robust Audio-Visual Instance Discrimination," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2021.

Self-supervised audiovisual scene learning



- Tackling Faulty Positives

$$\mathcal{L}_{\text{RxID}} = \frac{\sum_i w_i \mathcal{L}_{\text{xID}}(\mathbf{v}_i, \mathbf{a}_i)}{\sum_i w_i}$$

- Tackling Faulty Negatives

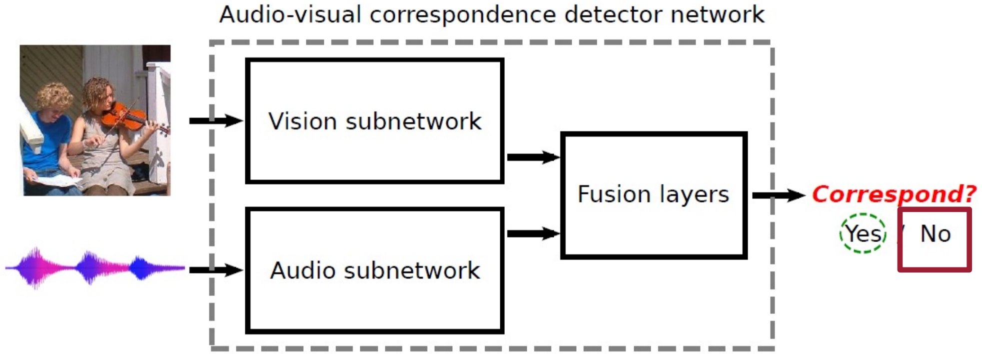
$$\begin{aligned}\mathcal{L}_{\text{Soft-xID}}(\mathbf{v}_i, \mathbf{a}_i) = &- \sum_j T_v(j|i) \log P(\bar{\mathbf{a}}_j|\mathbf{v}_i;\tau) \\ &- \sum_j T_a(j|i) \log P(\bar{\mathbf{v}}_j|\mathbf{a}_i;\tau)\end{aligned}$$

$$T_v(j|i) = (1-\lambda)\mathbf{1}_{i=j} + \lambda S_v(j|i)$$
$$T_a(j|i) = (1-\lambda)\mathbf{1}_{i=j} + \lambda S_a(j|i)$$

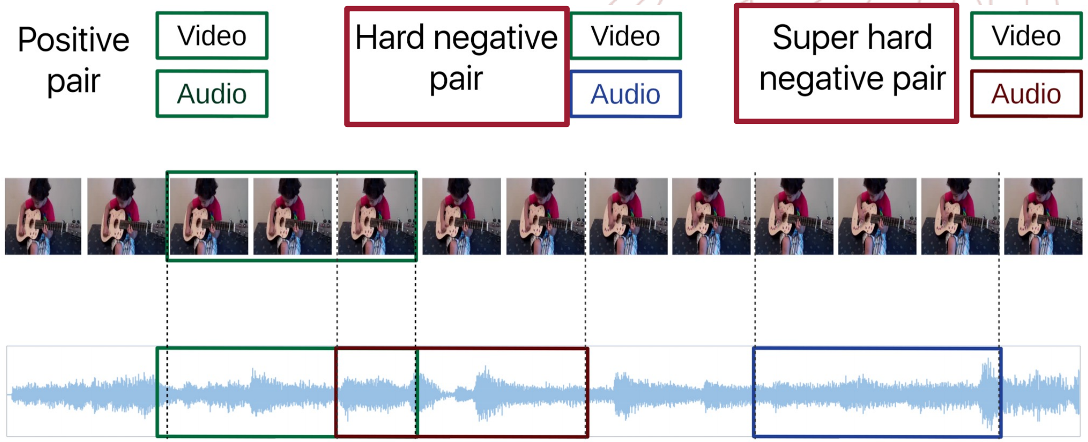
[10] P. Morgado, I. Misra, and N. Vasconcelos, “Robust Audio-Visual Instance Discrimination,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2021.

Self-supervised audiovisual scene learning

Visible scenes ⟷ Soundscape



Video-level Correspondence[1]



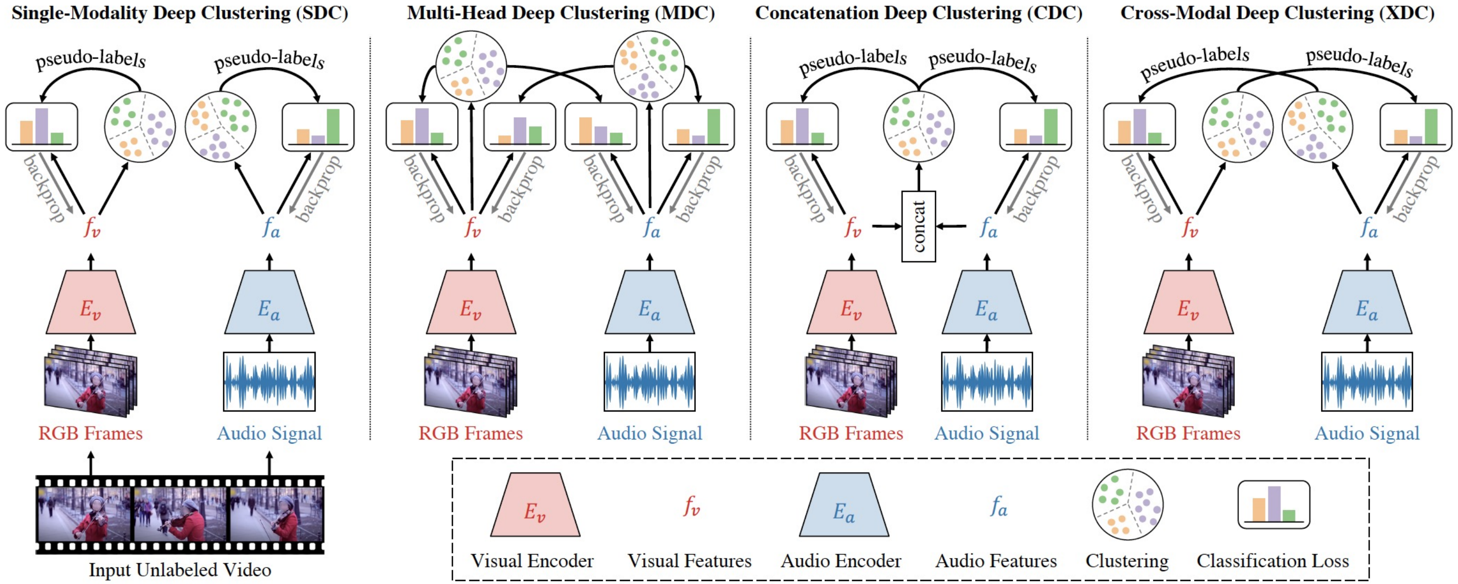
Temporal-level synchronization[2]

What if we do not provide negative samples?

[1] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," *Proc. IEEE Conf. Computer Vision*, 2017.

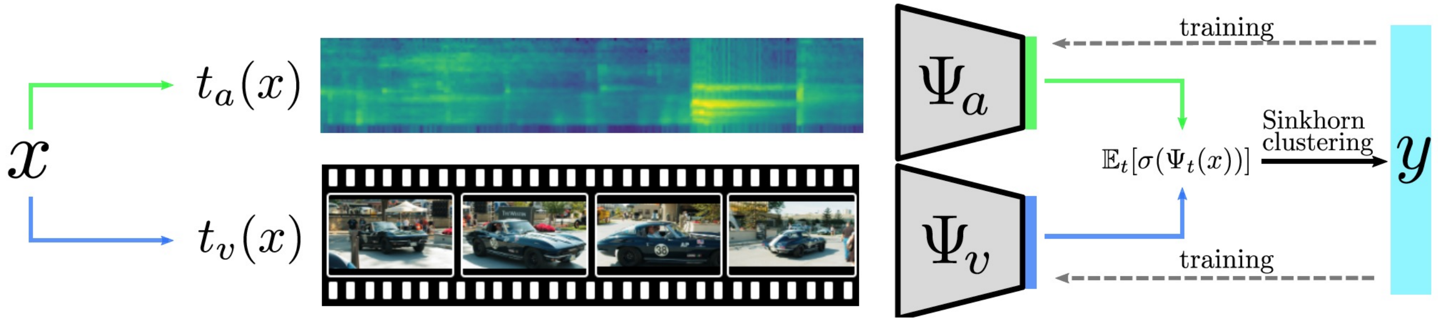
[2] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems*, 2018.

Self-supervised audiovisual scene learning



Multimodal clustering with different heads[9]

Pseudo labelling from Audio-visual modality[10]



[11] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *Advances in Neural Information Processing Systems*, 2020.
[12] Y. Asano*, M. Patrick*, C. Rupprecht, and A. Vedaldi, "Labelling unlabelled videos from scratch with multi-modal self-supervision," *Advances in Neural Information Processing Systems*, 2020.

Self-supervised audiovisual scene learning

(a) Video action recognition.

Method	Pretraining Architecture	Pretraining Dataset	Evaluation UCF101	Evaluation HMDB51
ClipOrder [75]	R(2+1)D-18	UCF101	72.4	30.9
MotionPred [68]	C3D	Kinetics	61.2	33.4
ST-Puzzle [27]	3D-ResNet18	Kinetics	65.8	33.7
DPC [17]	3D-ResNet34	Kinetics	75.7	35.7
CBT [61]	S3D	Kinetics	79.5	44.6
SpeedNet [4]	S3D	Kinetics	81.1	48.8
AVTS [28]*	MC3-18	Kinetics	84.1	52.5
AVTS [28]†	R(2+1)D-18	Kinetics	86.2	52.3
XDC (ours)	R(2+1)D-18	Kinetics	86.8	52.6
AVTS [28]*	MC3-18	AudioSet	87.7	57.3
AVTS [28]†	R(2+1)D-18	AudioSet	89.1	58.1
XDC (ours)	R(2+1)D-18	AudioSet	93.0	63.7
MIL-NCE [37]	S3D	HowTo100M	91.3	61.0
ELo [49]	R(2+1)D-50	YouTube-8M	93.8	67.4
XDC (ours)	R(2+1)D-18	IG-Random	94.6	66.5
XDC (ours)	R(2+1)D-18	IG-Kinetics	**95.5**	**68.9**
Fully supervised	R(2+1)D-18	ImageNet	84.0	48.1
Fully supervised	R(2+1)D-18	Kinetics	94.2	65.1

With negative samples

Without negative samples

(b) Audio event classification.

Method	ESC50
Random Forest [48]	44.3
Piczak ConvNet [47]	64.5
SoundNet [2]	74.2
L^3-Net [1]	79.3
AVTS [28]	82.3
ConvRBM [54]	**86.5**
XDC (AudioSet)	84.8
XDC (IG-Random)	85.4

Method	DCASE
RG [50]	69
LTT [34]	72
RNH [52]	77
Ensemble [58]	78
SoundNet [2]	88
L^3-Net [1]	93
AVTS [28]	94
XDC (AudioSet)	**95**
XDC (IG-Random)	**95**

With negative samples

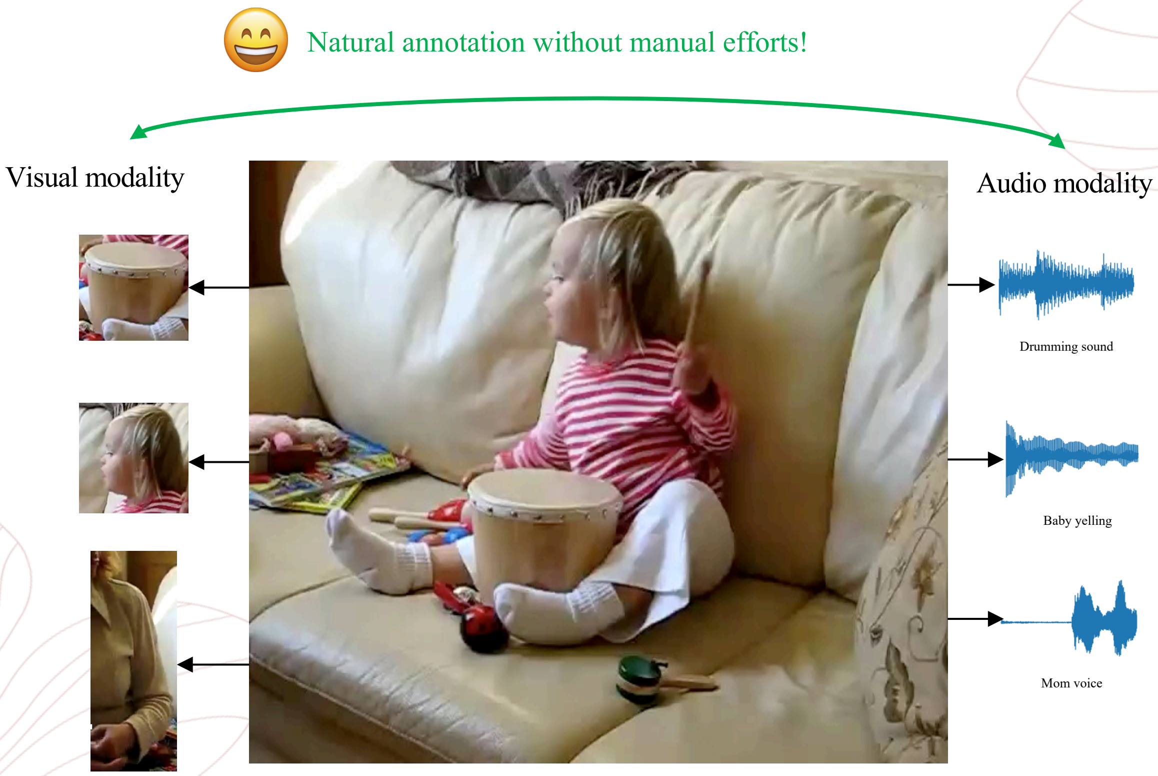
Without negative samples

[11] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *Advances in Neural Information Processing Systems*, 2020.

Approaches Overview

- Cross-modal knowledge transfer
- Self-supervised audiovisual scene learning
- Self-supervised audiovisual object perception

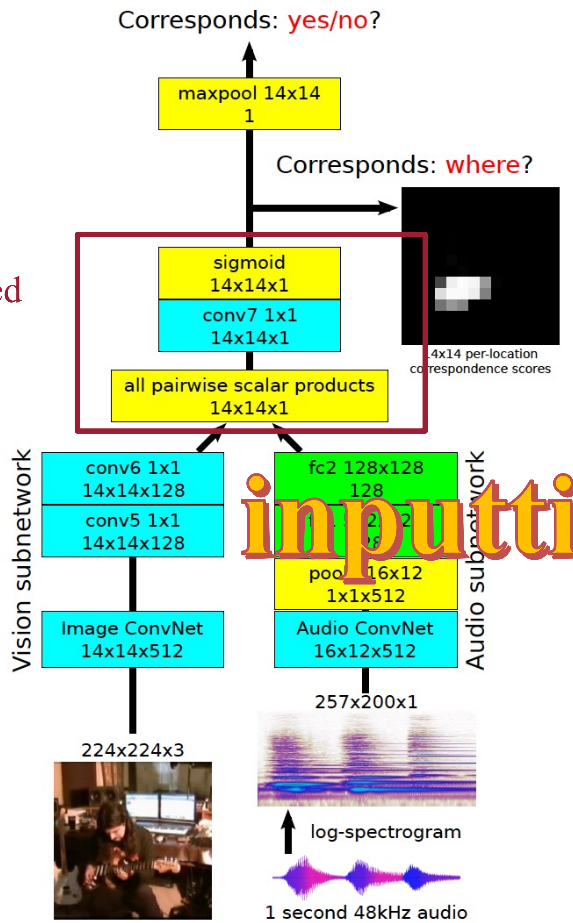
Self-supervised audiovisual object perception



Can we learn to recognize objects via their sounds?

Self-supervised audiovisual object perception

Location-based
Audiovisual
Similarity

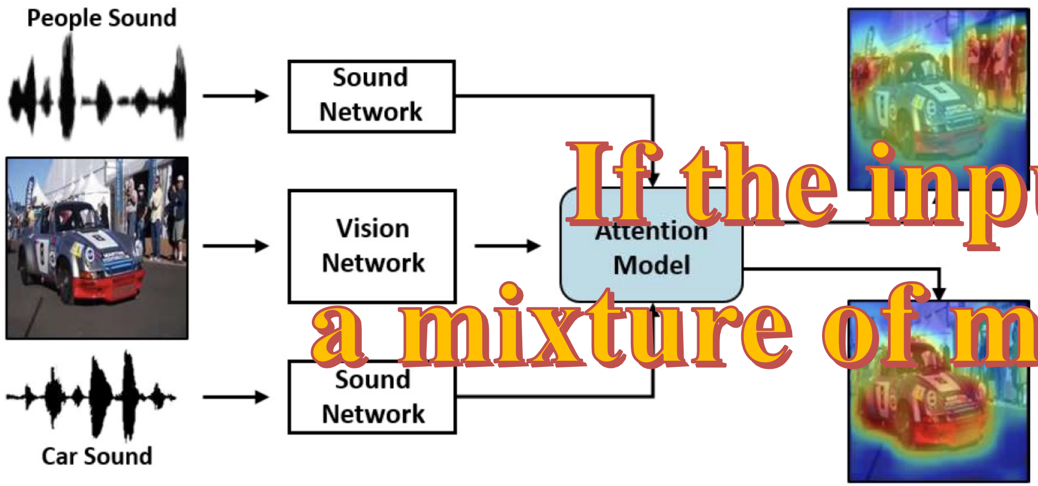


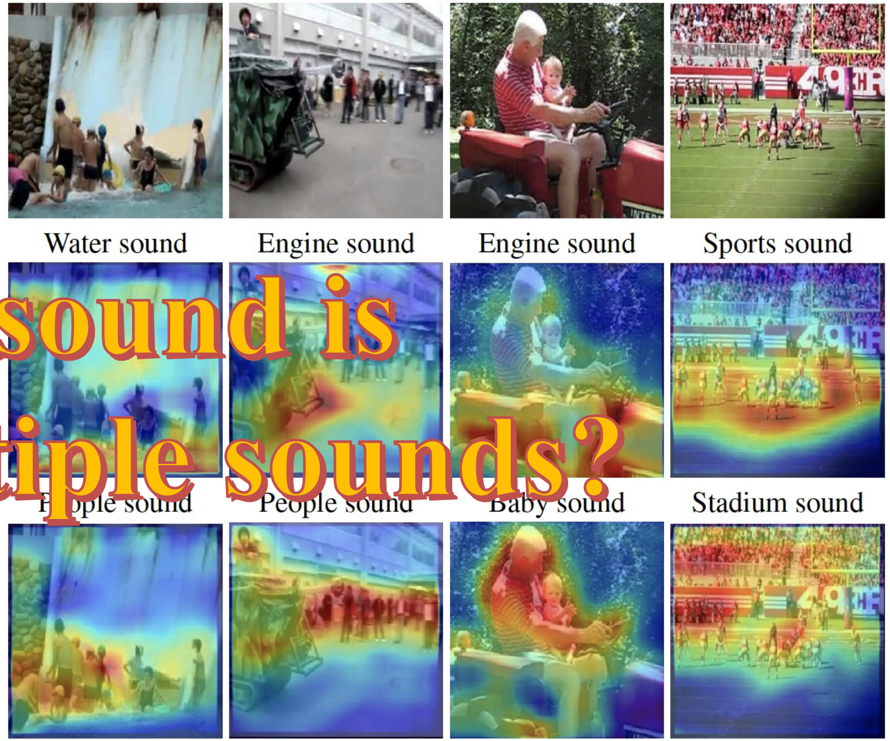
How about inputting different sounds?

Localization of sounding objects

[13] R. Arandjelovic and A. Zisserman, "Objects that Sound," *Proc. European Conf. Computer Vision*, 2018.

Self-supervised audiovisual object perception





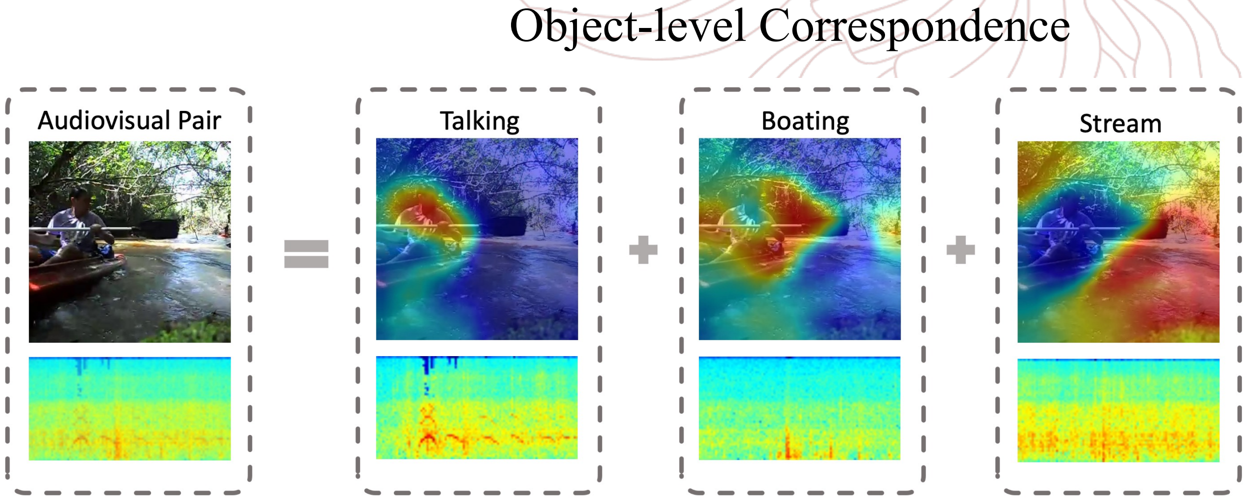
If the input sound is a mixture of multiple sounds?

Attention-based sound source localization

Localization of sounding objects w.r.t. different sounds

[14] A. Senocak, T. Oh, J. Kim, M. Yang, and I. Kweon, "Learning to localize sound source in visual scenes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018.

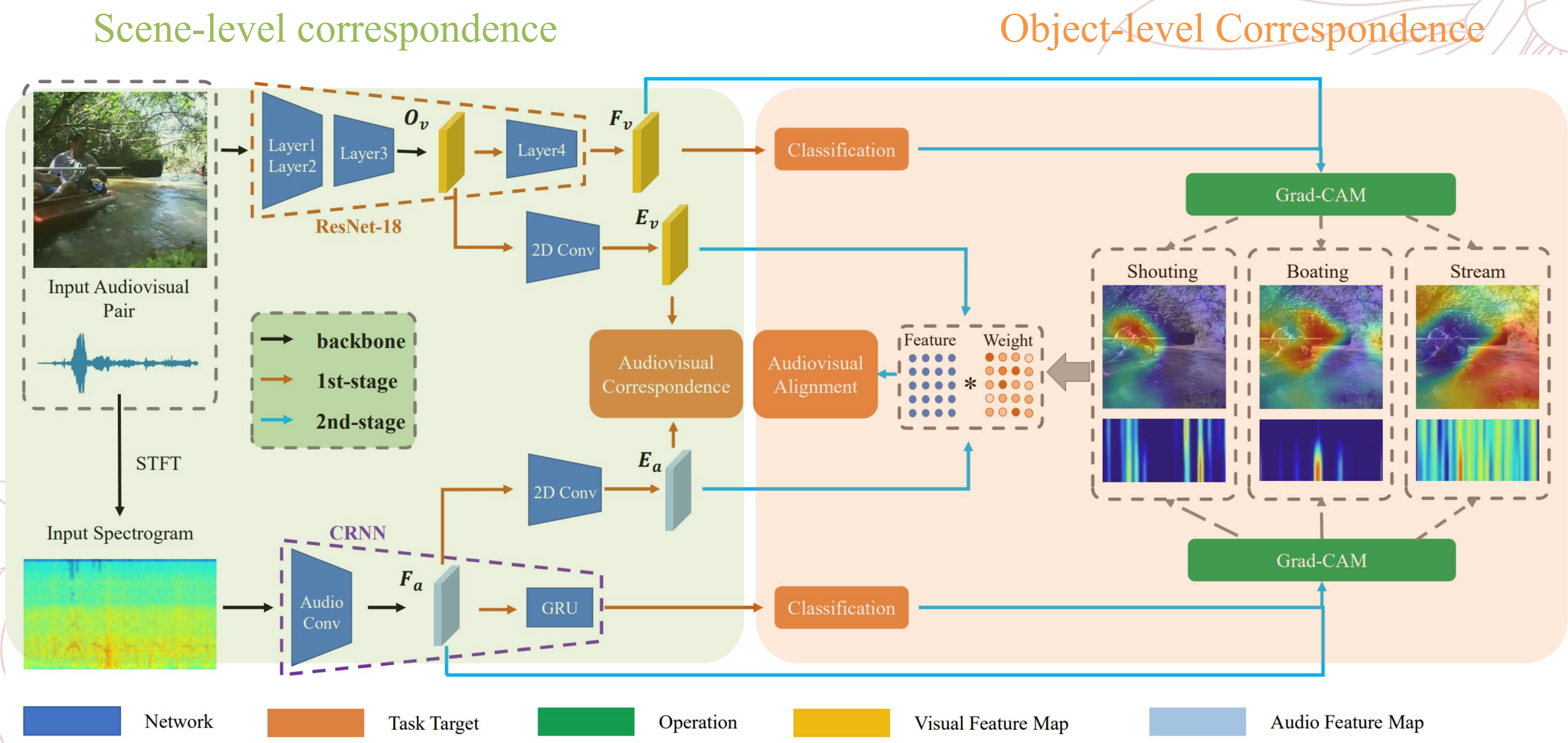
Self-supervised audiovisual object perception



How to explore the fine-grained supervision?

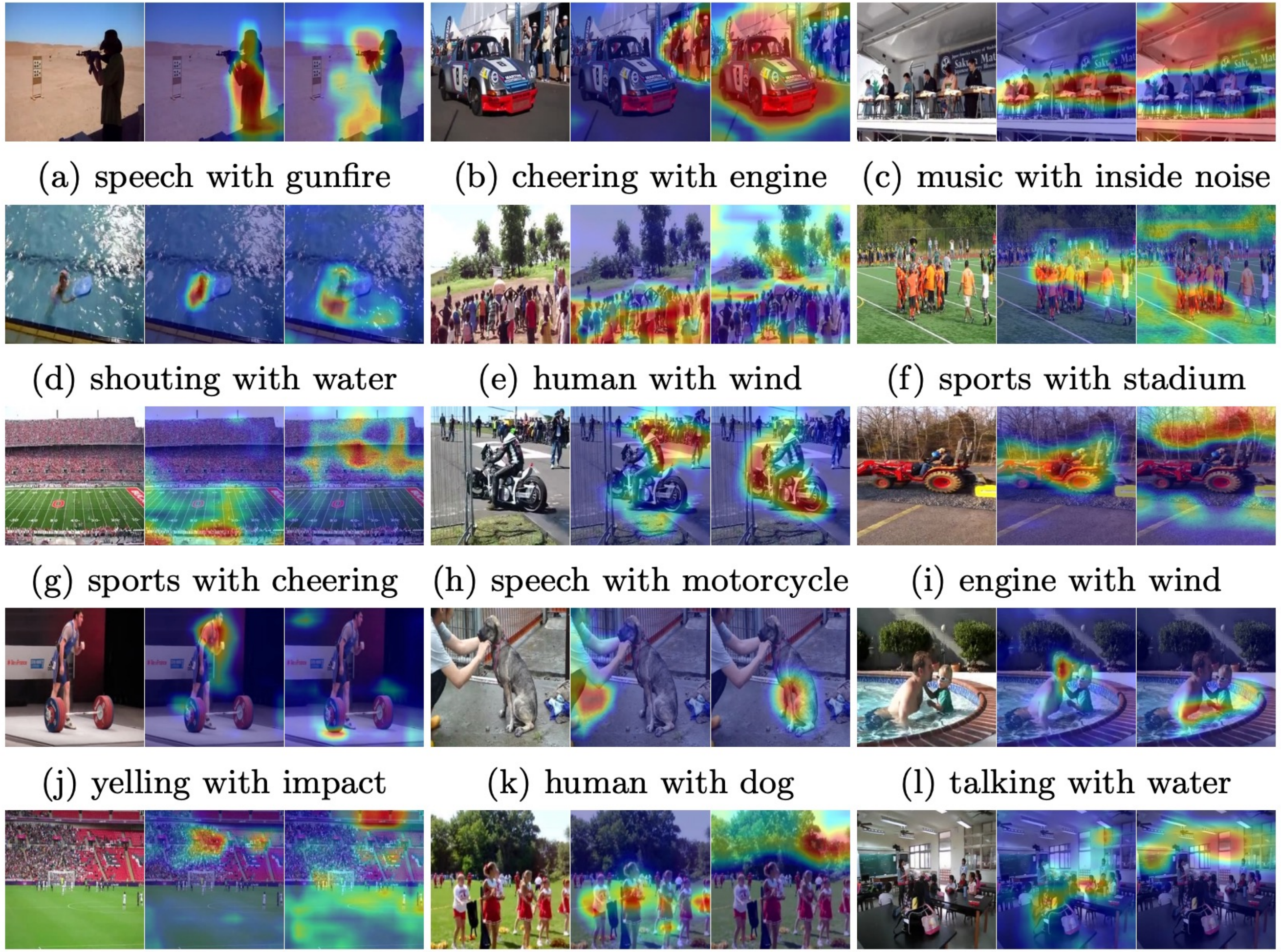
[15] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, “Multiple Sound Sources Localization from Coarse to Fine,” *Proc. European Conf. Computer Vision,* 2020.

Self-supervised audiovisual object perception



[15] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, "Multiple Sound Sources Localization from Coarse to Fine," *Proc. European Conf. Computer Vision,* 2020.

Self-supervised audiovisual object perception

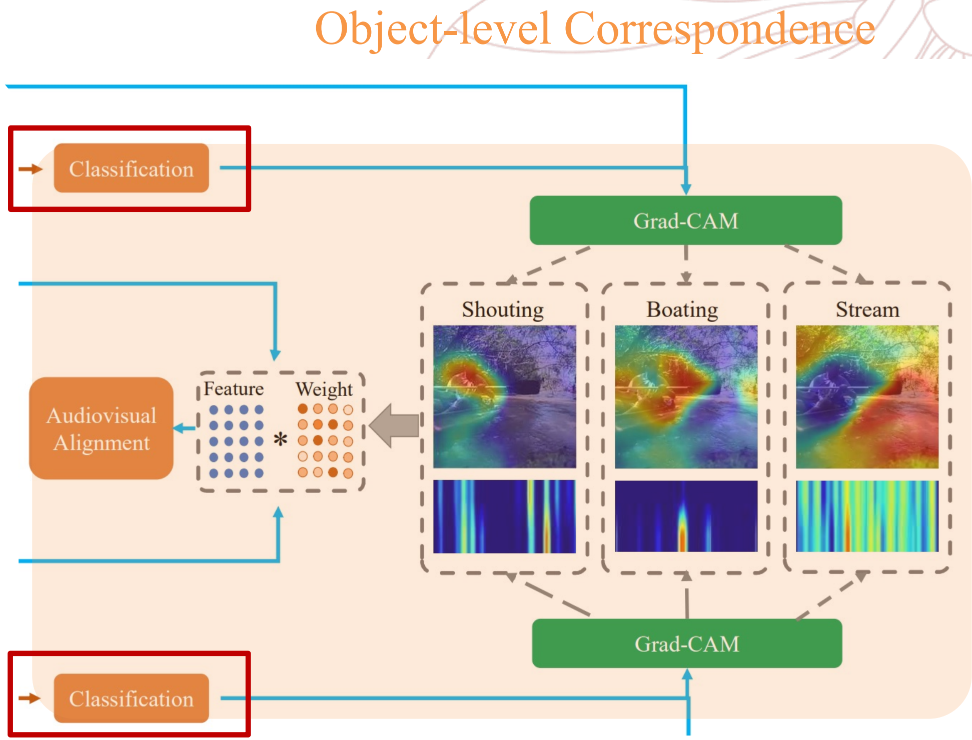


(a) speech with gunfire (b) cheering with engine (c) music with inside noise
(d) shouting with water (e) human with wind (f) sports with stadium
(g) sports with cheering (h) speech with motorcycle (i) engine with wind
(j) yelling with impact (k) human with dog (l) talking with water
(m) screaming with stadium (n) yelling with wind (o) speech with classroom

[15] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, "Multiple Sound Sources Localization from Coarse to Fine," *Proc. European Conf. Computer Vision,* 2020.

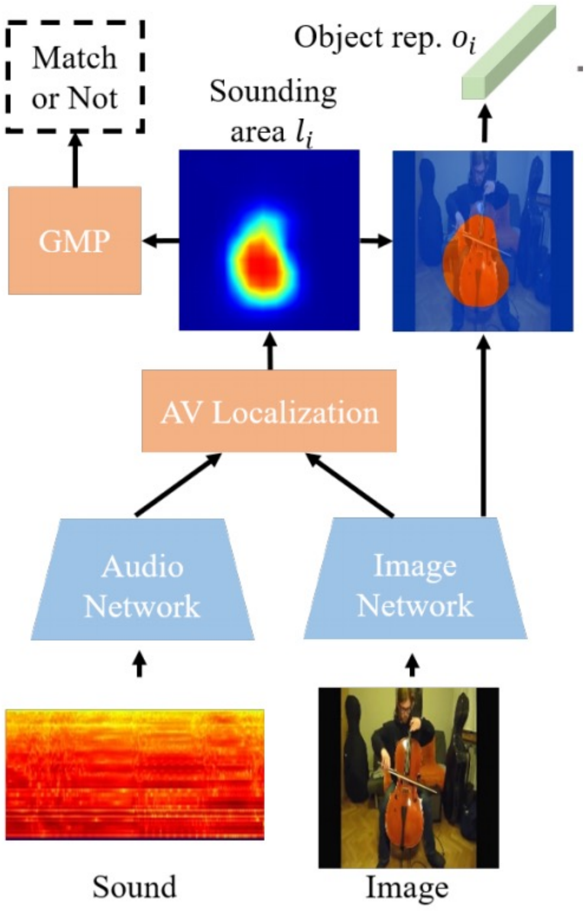
Self-supervised audiovisual object perception

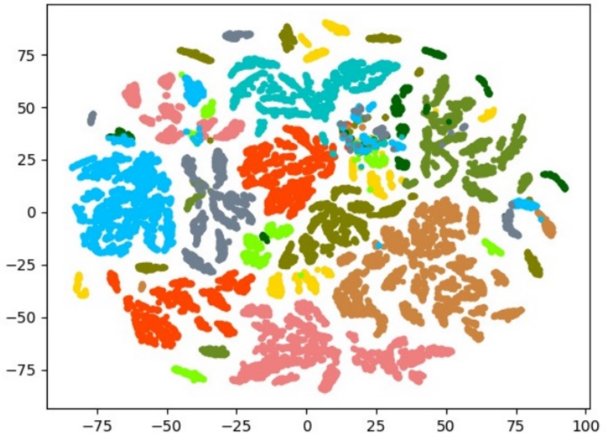
Relying on pre-trained object knowledge



[15] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, "Multiple Sound Sources Localization from Coarse to Fine," *Proc. European Conf. Computer Vision,* 2020.

Self-supervised audiovisual object perception



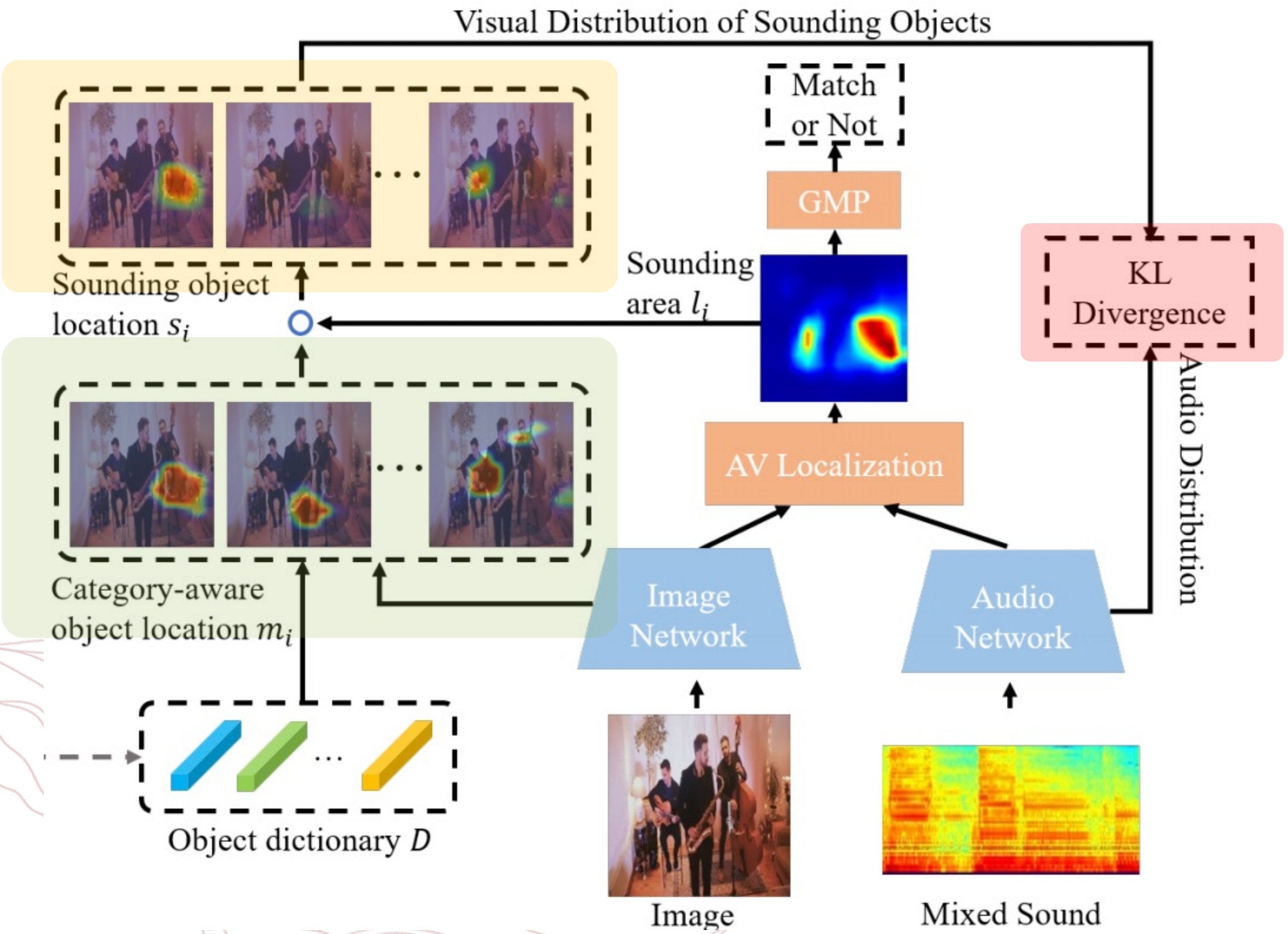


Learn visual knowledge from sound-source localization

Then finetuning to object detection or segmentation related task

[16] D. Hu, R. Qian, M. Jiang et al.,, "Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching," *Advances in Neural Information Processing Systems*, 2020.

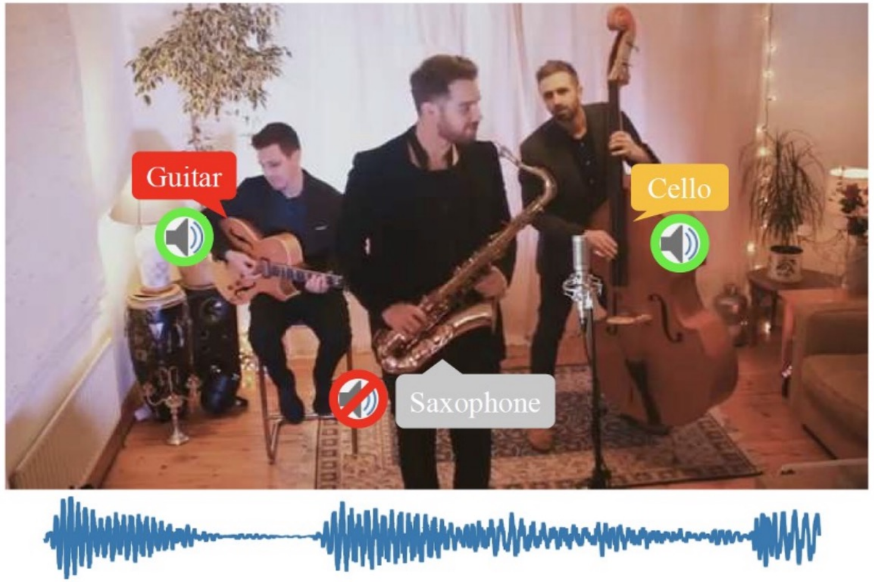
Self-supervised audiovisual object perception



Supervision still from correspondence but in the object-category level!

[16] D. Hu, R. Qian, M. Jiang et al.,, "Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching," *Advances in Neural Information Processing Systems*, 2020.

Self-supervised audiovisual object perception



A typical cocktail-party scenario

Object localization in cocktail-party[14]

[16] D. Hu, R. Qian, M. Jiang et al.,, "Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching," *Advances in Neural Information Processing Systems*, 2020.

Self-supervised audiovisual object perception

Object localization [5]

Methods	IoU@0.5	AUC
Sound-of-pixel	40.5	43.3
Object-that-sound	26.1	35.8
Attention	37.2	38.7
DMC	29.1	38.0
Ours	**51.4**	**43.6**

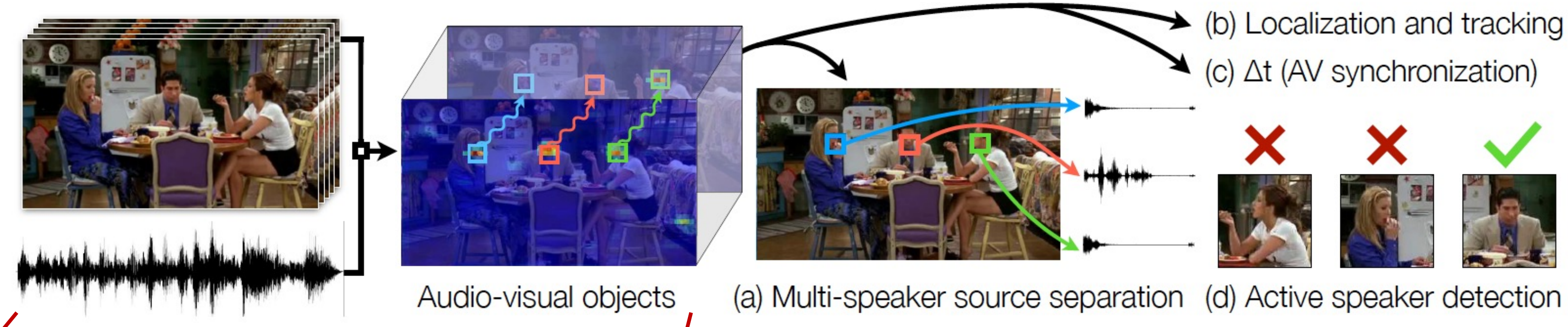
Single sound scene

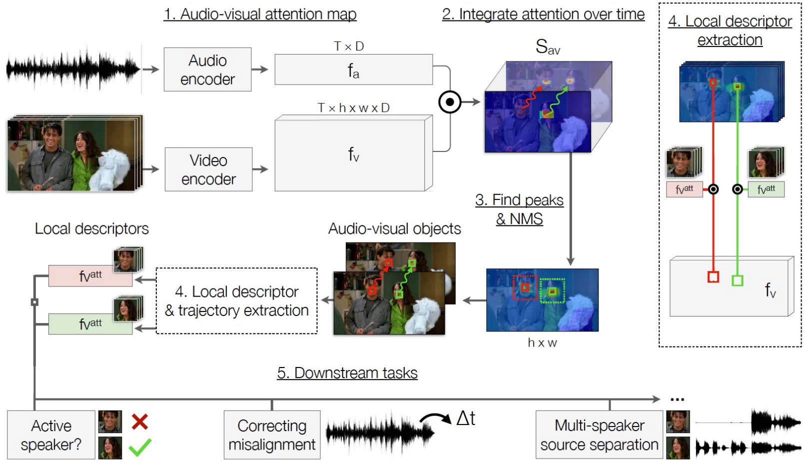
Data	MUSIC-Synthetic			MUSIC-Duet			AudioSet-multi		
Methods	CIoU	AUC	NSA	CIoU	AUC	NSA	CIoU	AUC	NSA
Sound-of-pixel	8.1	11.8	97.2	16.8	16.8	**92.0**	39.8	27.3	**88.8**
Object-that-sound	3.7	10.2	19.8	13.2	18.3	15.7	27.1	21.9	16.5
Attention	6.4	12.3	77.9	21.5	19.4	54.6	29.9	23.5	4.5
DMC	7.0	16.3	-	17.3	21.1	-	32.0	25.2	-
Ours	**32.3**	**23.5**	**98.5**	**30.2**	**22.1**	83.1	**48.7**	**29.7**	56.8

Cocktail-party scene

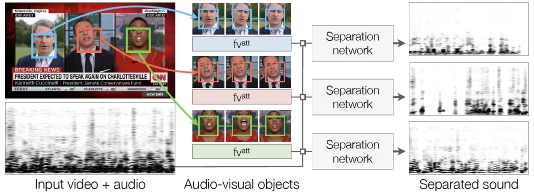
[16] D. Hu, R. Qian, M. Jiang et al.,, "Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching," *Advances in Neural Information Processing Systems*, 2020.

Self-supervised audiovisual object perception





(a) Multi-speaker separation



(b) Talking head detection and tracking

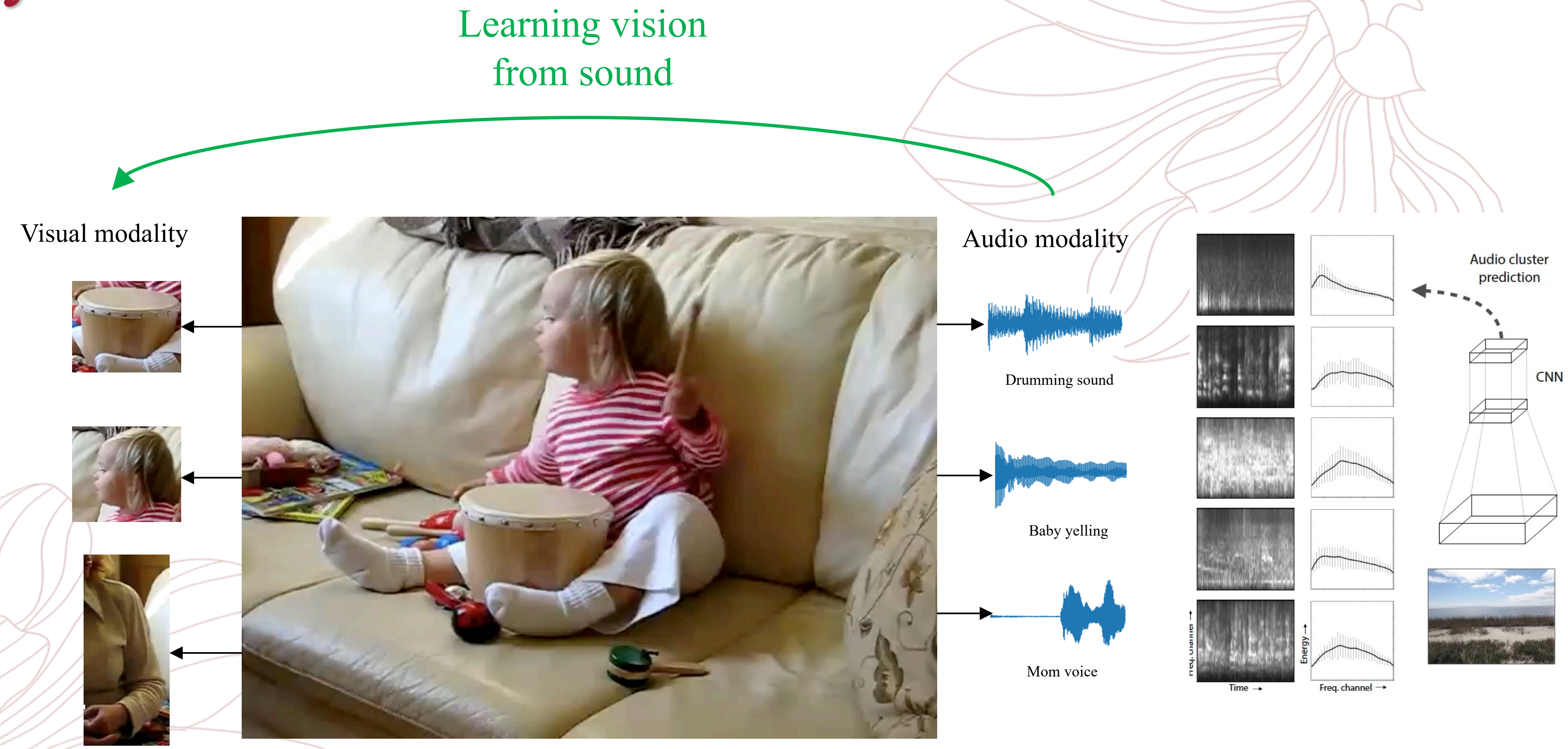
(c) Active speaker detection

[17] T. Afouras, A. Owens, J. Chung, A. Zisserman, "Self-Supervised Learning Of Audio-Visual Objects From Video," *Proc. European Conf. Computer Vision,* 2020.

Outline

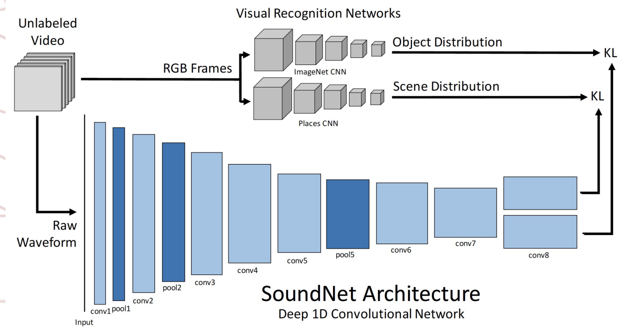
- Topic Overview
  - What is and why using self-supervised audiovisual learning
- Approaches Overview
  - What are the state of the art approaches and what is the inherent self-supervision
- Summary
  - What are the core challenges and future directions

Summary

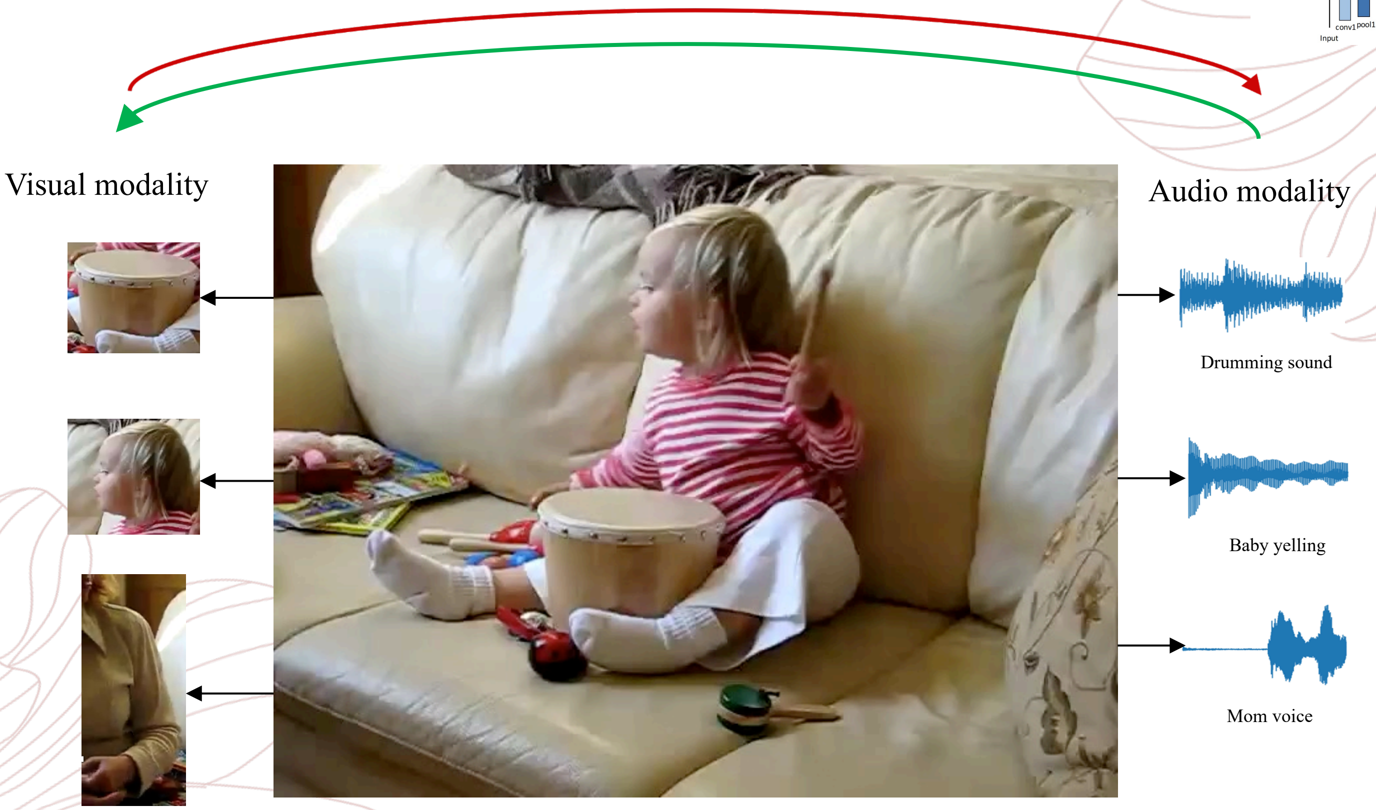


Learning vision
from sound
Visual modality
Audio modality
Drumming sound
Baby yelling
Mom voice
Audio cluster
prediction
CNN
Time
Energy
Freq. channel

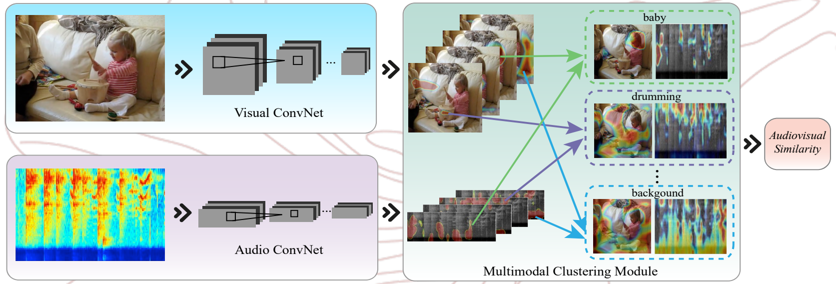
Summary



Learning sound from vision

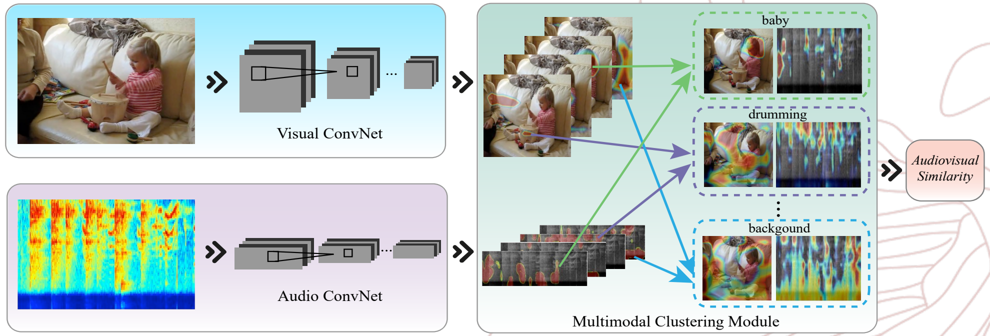


Summary



Self-supervised learning from
both sound and vision

Summary

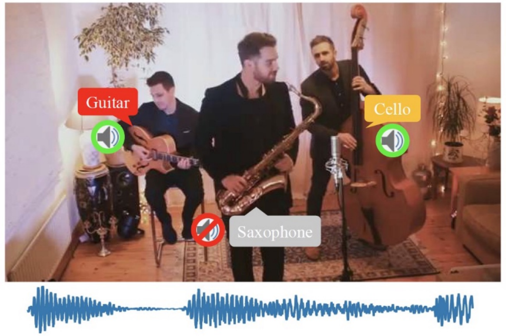


Self-supervised
Audiovisual Learning

Keys

Effective Modeling
for
Audiovisual Scenes

Effective Supervision
for
Audiovisual Learning



Thank You!

Di Hu
Gaoling School of Artificial Intelligence
Renmin University of China

Email: dihu@ruc.edu.cn
19-06-2021

Reference

[1] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," *Proc. IEEE Conf. Computer Vision*, 2017.

[2] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems*, 2018.

[3] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video," *Advances in Neural Information Processing Systems*, 2016.

[4] D. Harwath, A. Torralba, and J. Glass, "Unsupervised Learning of Spoken Language with Visual Context," *Advances in Neural Information Processing Systems*, 2016.

[5] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," *Proc. Annual Meeting of the Association for Computational Linguistics,* 2017.

[6] A. Owens, J. Wu, J. McDermott, W. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," *Proc. European Conf. Computer Vision*, 2016.

[7] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," arXiv:1706.00932, 2017.

[8] A. Owens, and A.Efros, "Audio-visual scene analysis with self-supervised multisensory features," *arXiv: 1804.03641*, 2018.

[9] D. Hu, F. Nie, and X. Li, "Deep Multimodal Learning for Unsupervised Audiovisual Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

[10] P. Morgado, I. Misra, and N. Vasconcelos, "Robust Audio-Visual Instance Discrimination," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2021.

[11] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *Advances in Neural Information Processing Systems*, 2020.

[12] Y. Asano*, M. Patrick*, C. Rupprecht, and A. Vedaldi, "Labelling unlabelled videos from scratch with multi-modal self-supervision," *Advances in Neural Information Processing Systems*, 2020.

[13] R. Arandjelovic and A. Zisserman, "Objects that Sound," *Proc. European Conf. Computer Vision*, 2018.

[14] A. Senocak, T. Oh, J. Kim, M. Yang, and I. Kweon, "Learning to localize sound source in visual scenes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018.

[15] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, "Multiple Sound Sources Localization from Coarse to Fine," *Proc. European Conf. Computer Vision,* 2020.

[16] D. Hu, R. Qian, M. Jiang et al.,, "Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching," *Advances in Neural Information Processing Systems*, 2020.

[17] T. Afouras, A. Owens, J. Chung, A. Zisserman, "Self-Supervised Learning Of Audio-Visual Objects From Video," *Proc. European Conf. Computer Vision,* 2020.