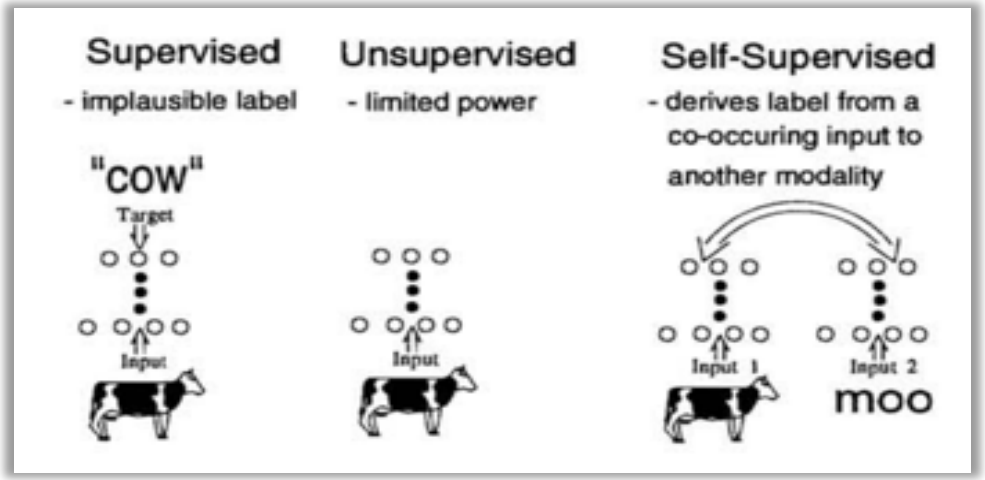
Audio-Visual Instance Discrimination

Pedro Morgado, Nuno Vasconcelos, Ishan Misra

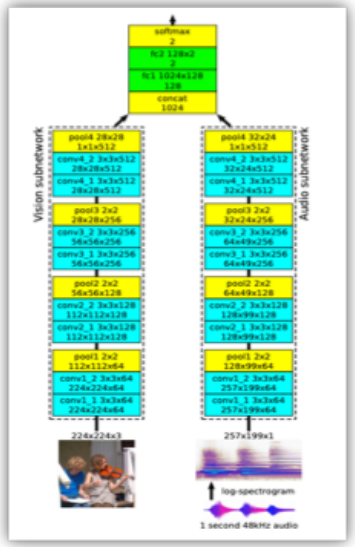
UC San Diego Facebook AI Research

Learning representations from audio-visual co-occurrence

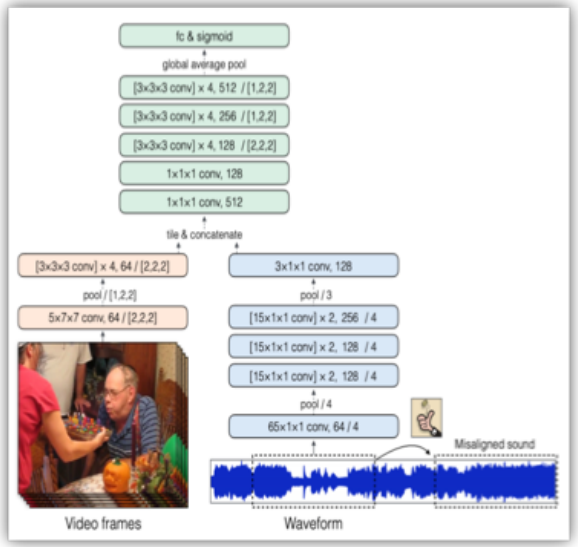


Learning Classification with Unlabeled Data
Virginia de Sa, 1994.

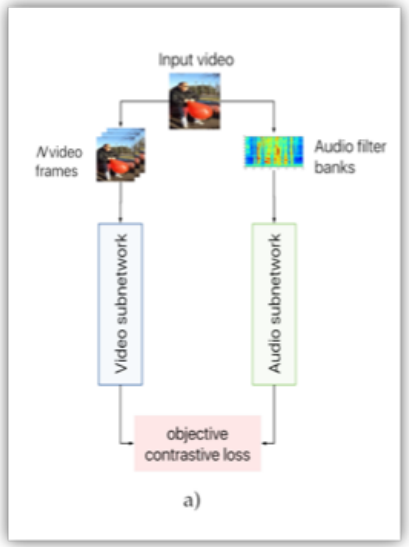
Learning representations from audio-visual co-occurrence



Look, listen & learn
Arandjelovic et al 2017.

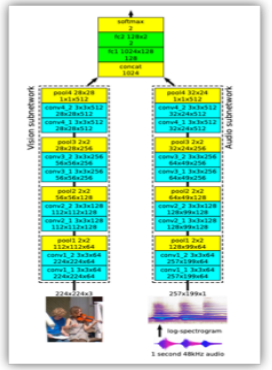


Multisensory synchronization
Owens et al 2018.



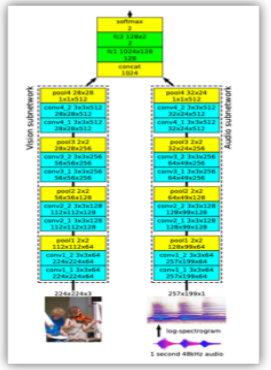
Synchronization w/ curriculum
Korbar et. al 2018.

Audio-visual correspondence



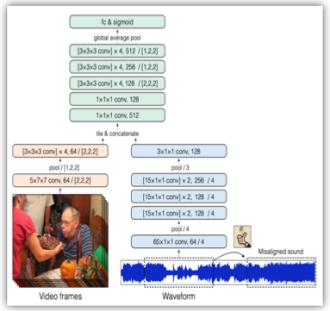
Look, listen & learn
Arandjelovic et al., 2017.

Audio-visual correspondence

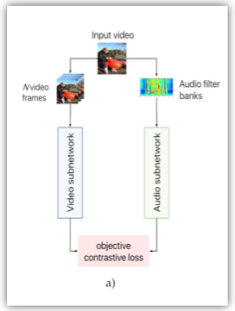


Look, listen & learn
Arandjelovic et al., 2017.

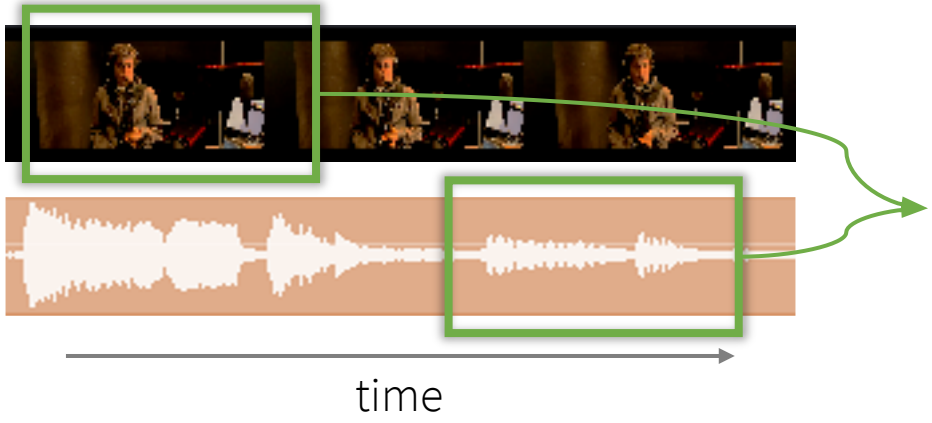
Audio-visual synchronization



Multisensory synchronization
Owens et al 2018.



Synchronization w/ curriculum
Korbar et. al 2018.



Pros & Cons

Audio-visual correspondence

- Negative audio-video pairs are often too easy.

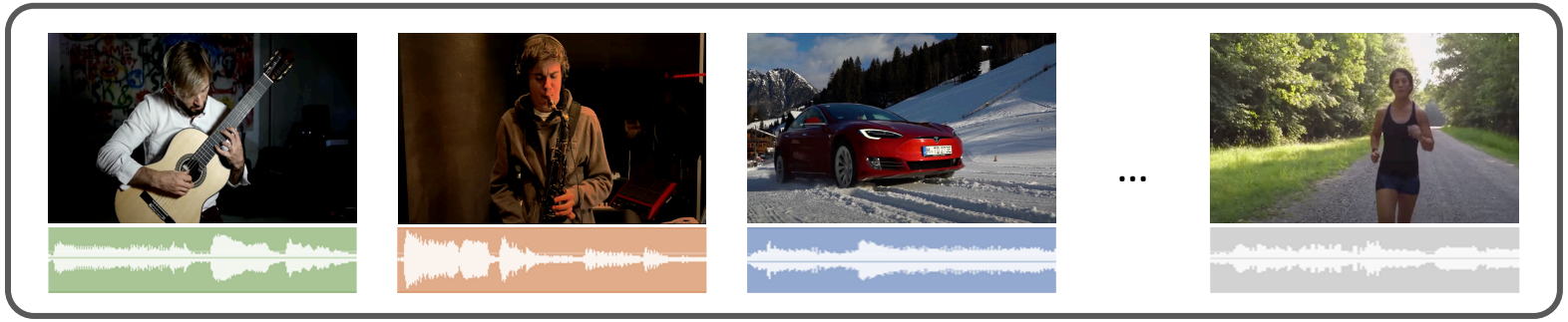
Audio-visual synchronization

- Off-sync audio-video pairs are harder to spot.
- However, synchronization can rely on correlations between loudness and movement, without learning more high-level semantic features.

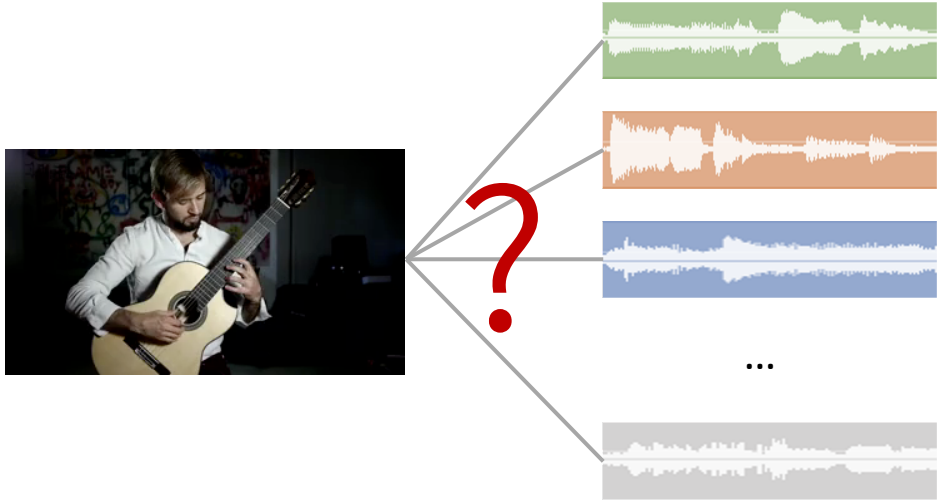
Question: Are there other ways to increase the difficulty of the AVC task?

Audio-Visual Instance Discrimination (AVID)

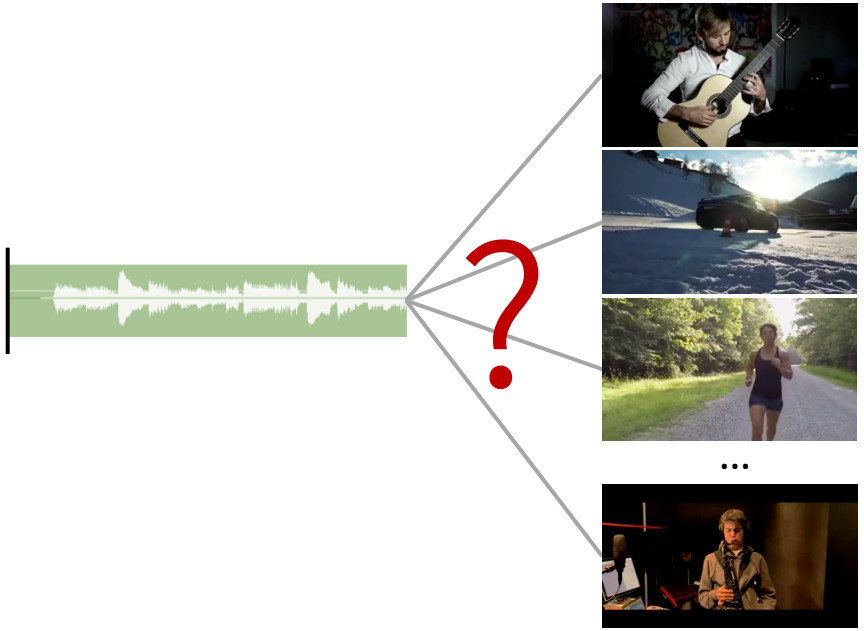
Video dataset



Multiple choices



Multiple choices



Audio-Visual Instance Discrimination (AVID)

As the number of negatives grows, the difficulty of some negatives increases naturally.

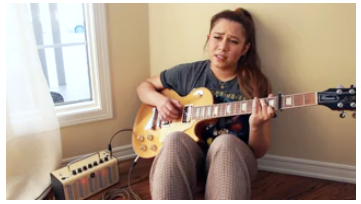
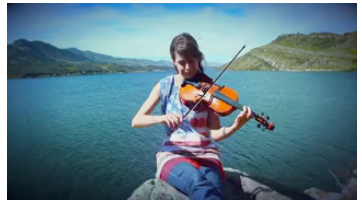
Base instance

Negatives




...

Hard Negatives (for free)


...

Increasing the number of negatives

Solution #1: Increasing the batch size

+ Loss is fully differentiable (allow back-propagation over anchors, positive and negative samples);

- Loss is batch size dependent;

- Requires distributed training for large number of negatives;

Solution #2: Using memory banks [1]

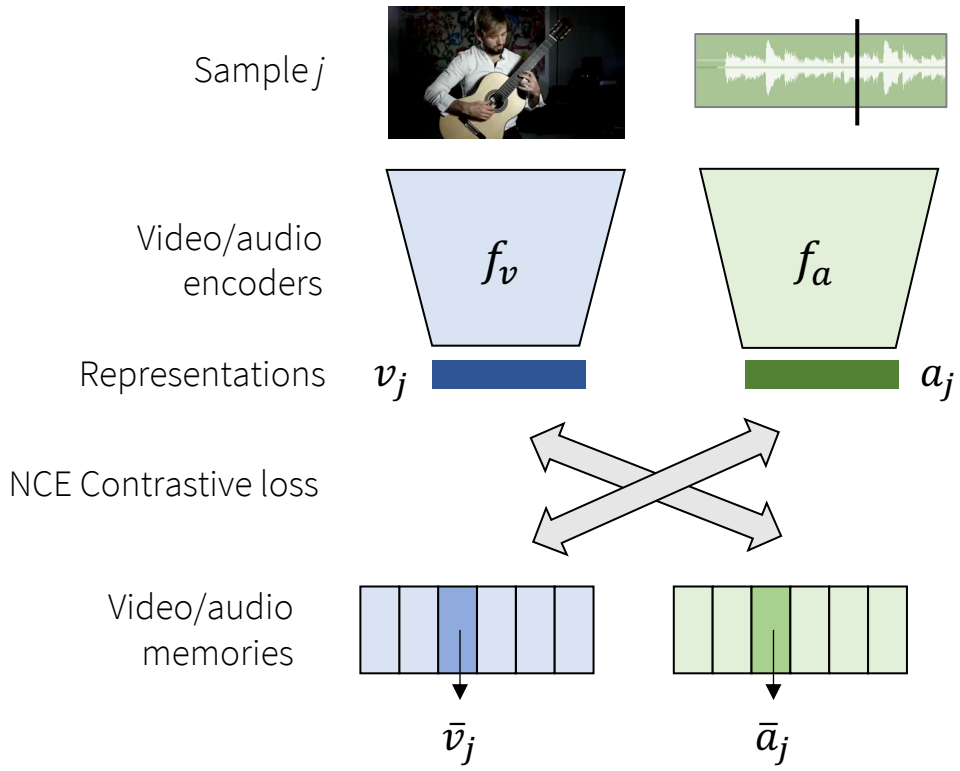
+ Loss independent of batch size

+ Number of negatives can be easily scaled up without distributed training.

- Size of memory bank becomes prohibitively large for *very* large datasets (>100M samples).

[1] Zhirong Wu et al. Unsupervised feature learning via non-parametric instance discrimination, CVPR, 2018.

AVID Training



Video encoder f_v

- R(2+1)D-CNN[1]

Audio encoder f_a

- 2D-CNN on log-spectrograms

Loss

- Noise contrastive estimation[2]

Memories $\bar{v}_j$ and $\bar{a}_j$

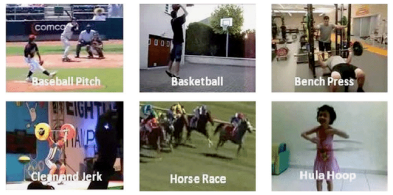
- Updated by an exponential moving average of representations v_j and a_j

[1] Tran et al. A closer look at spatiotemporal convolutions for action recognition, CVPR, 2018.
[2] Gutmann & Hyvarinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, ICAIS, 2010.

Downstream tasks

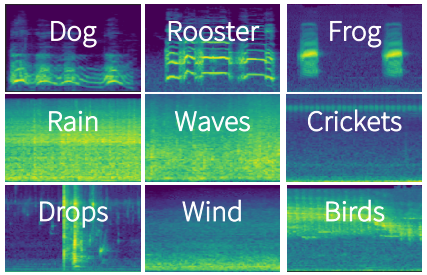
Action Recognition

UCF-101



HMDB-51

Sound Classification

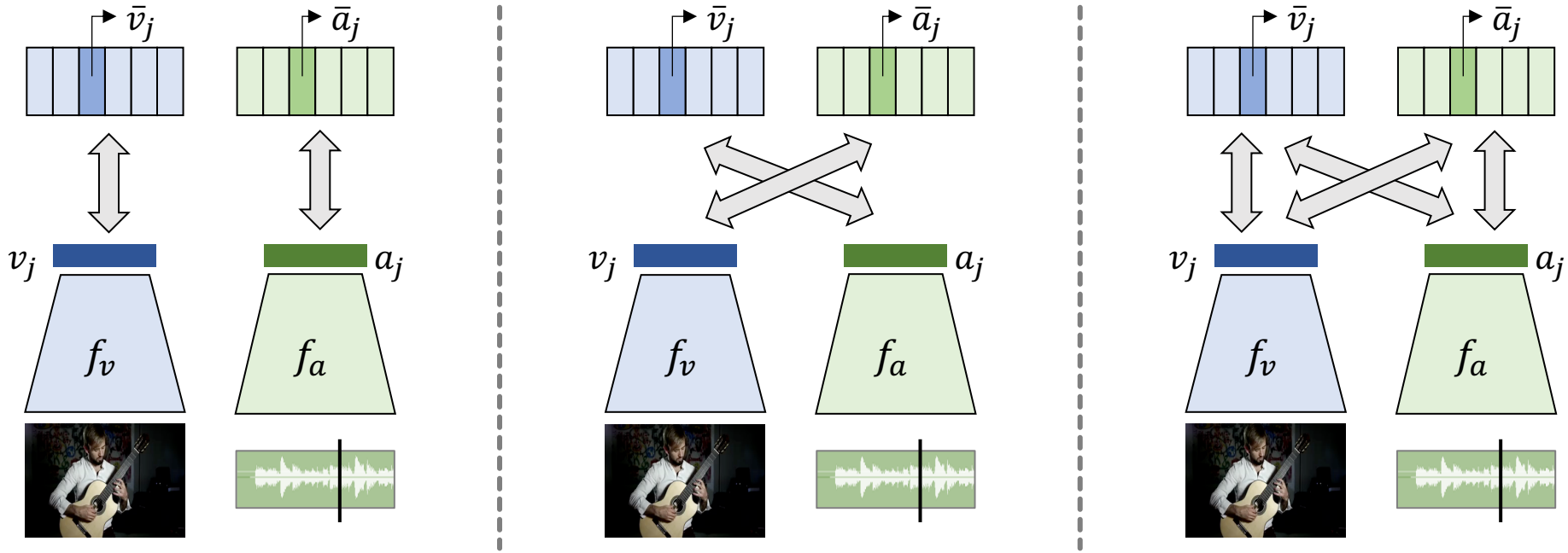


ESC-50

DCASE

Pre-trained models can be transferred to the downstream task, either by training a simple **linear classifier**, or by **finetuning** using the pre-trained weights as initialization.

Within-modal vs cross-modal supervision

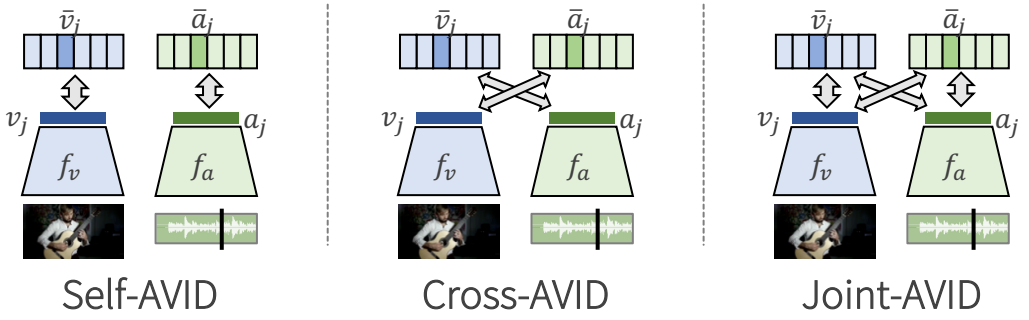


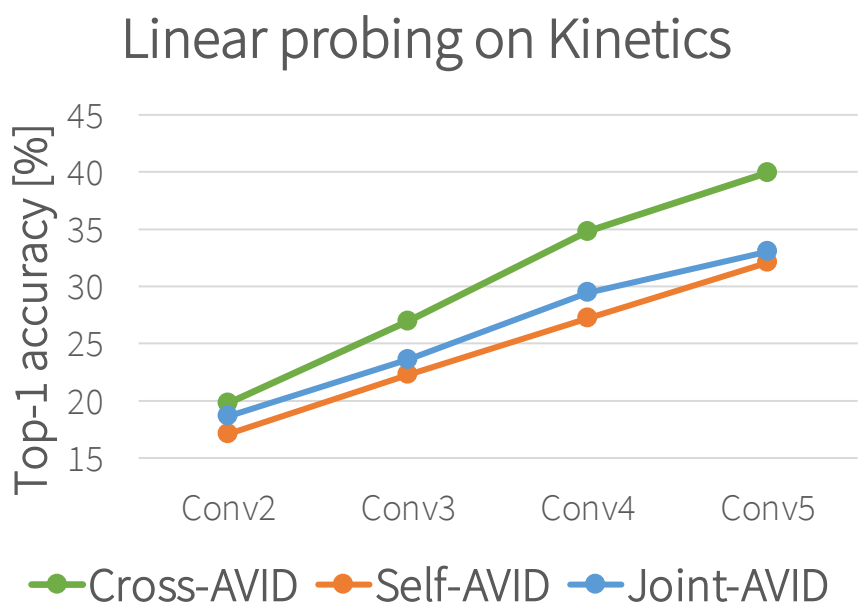
Self-AVID

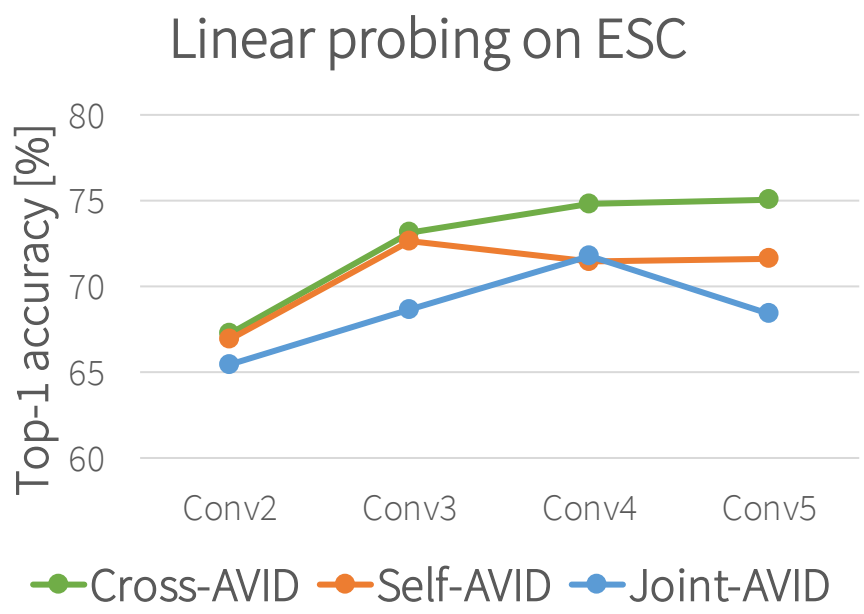
Cross-AVID

Joint-AVID

Within vs cross-modal supervision







Comparison to prior work

Finetuning for action recognition on UCF and HMDB.

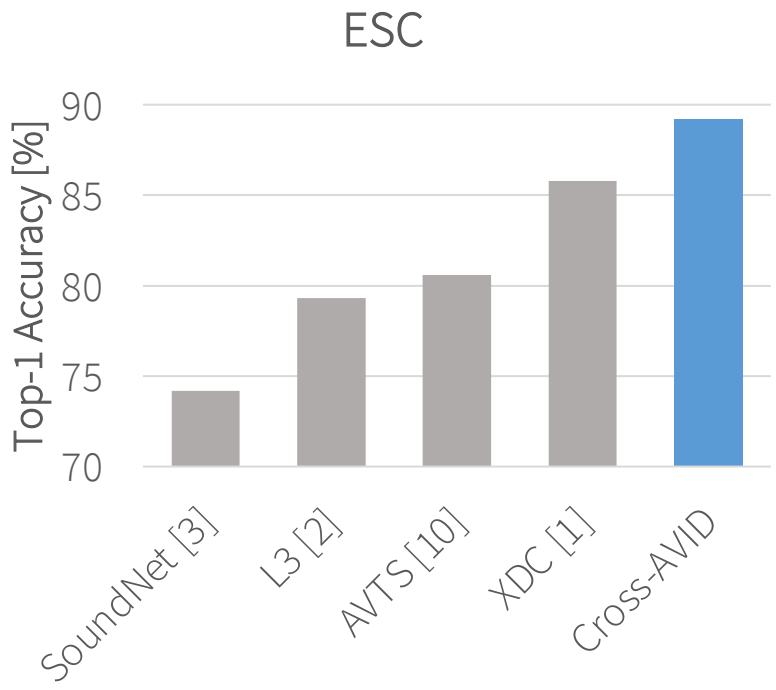
Audio-Visual Correspondence

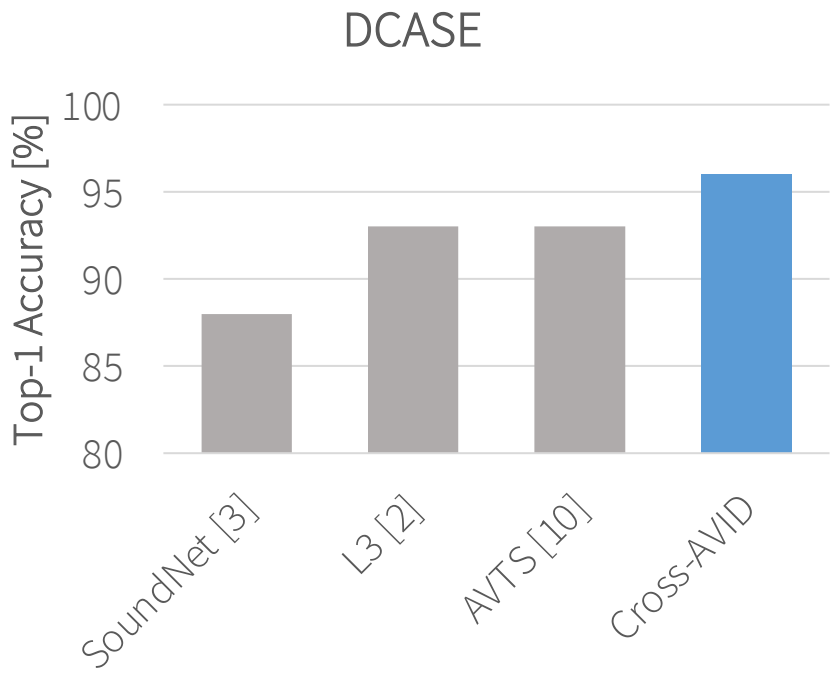
Audio-Visual Synchronization

Method	Backbone	Input	UCF	HMDB
		Pre-training DB: Kinetics		
DPC [9]	3D ResNet-34	25×128^2	75.7	35.7
CBT [20]	S3D Inception	16×112^2	79.5	44.6
L3* [2]	R(2+1)D-18	16×224^2	74.4	47.8
AVTS [10]	MC3-VGGish-9	25×224^2	85.8	56.9
XDC [1]	R(2+1)D-18	8×224^2	74.2	39.0
		32×224^2	84.2	47.1
Cross-AVID	R(2+1)D-18	8×224^2	82.3	49.1
		32×224^2	**86.9**	**59.9**
		Pre-training DB: Audioset		
L3* [2]	R(2+1)D-18	16×224^2	82.3	51.6
Multisensory [13]	3D-Resnet-18	64×224^2	82.1	–
AVTS [10]	MC3-VGGish-9	25×224^2	89.0	61.6
XDC [1]	R(2+1)D-18	8×224^2	84.9	48.8
		32×224^2	**91.2**	61.0
Cross-AVID	R(2+1)D-18	8×224^2	88.3	57.5
		32×224^2	91.0	**64.1**

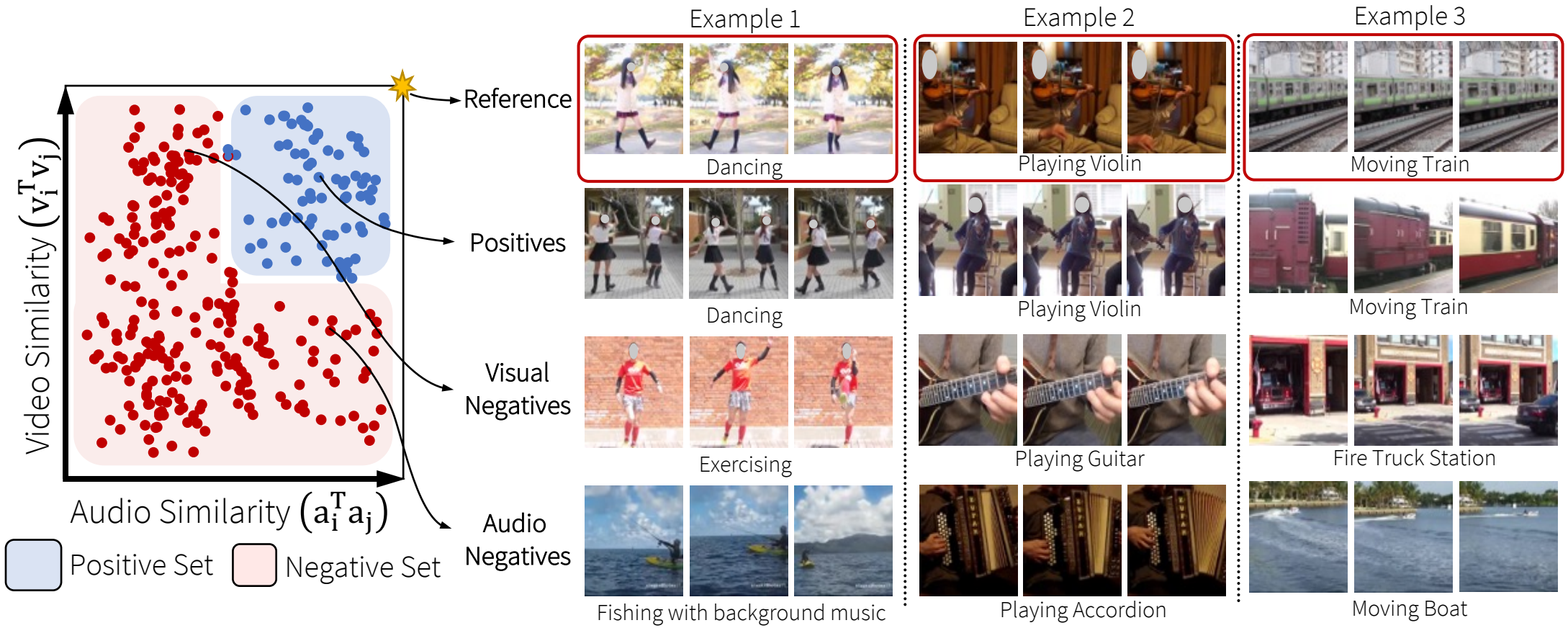
Comparison to prior work

Linear classifier for sound classification on ESC and DCASE.

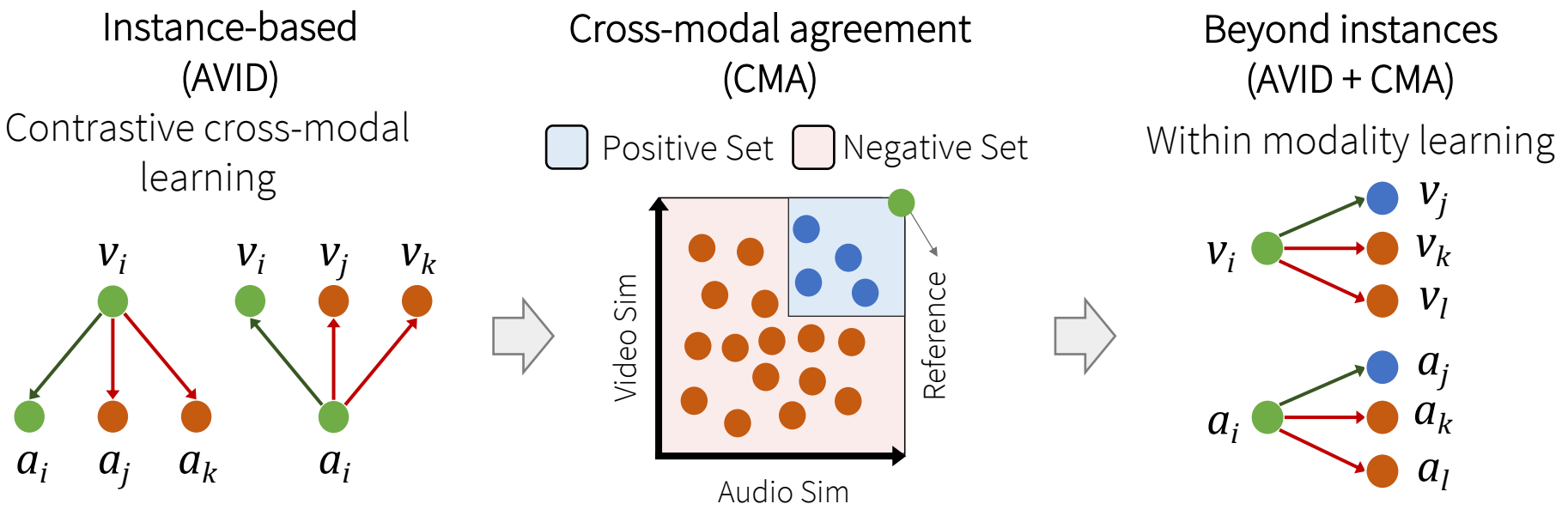




Visualizing agreements



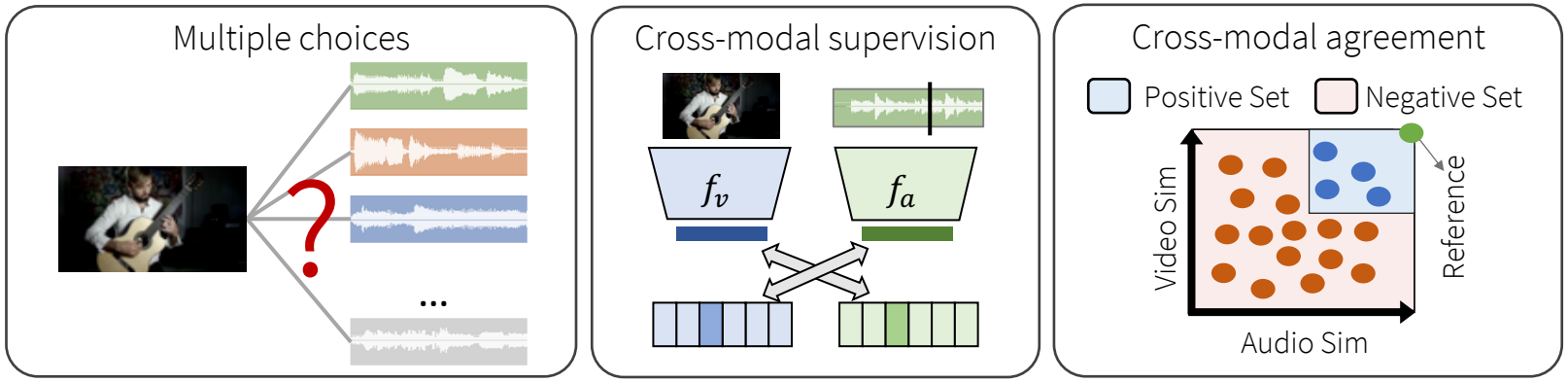
Supervision beyond instances



Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. "Audio-visual instance discrimination with cross-modal agreement." arXiv:2004.12943, 2020.

Thanks for listening!

Summary



Extended version: P Morgado, N Vasconcelos, and I Misra. "Audio-visual instance discrimination with cross-modal agreement." arXiv:2004.12943, 2020.

People

Pedro Morgado
UC San Diego

Nuno Vasconcelos
UC San Diego

Ishan Misra
Facebook AI Reseach