ACAV100M: Automatic Curation of Large-Scale Datasets for Audio-Visual Video Representation Learning


Sangho Lee*
SNU


Jiwan Chung*
SNU


Youngjae Yu
SNU


Gunhee Kim
SNU


Thomas Breuel
NVIDIA


Gal Chechik
NVIDIA


Yale Song
MSR

*: Equal Contribution

2021 ICCV OCTOBER 11-17 VIRTUAL

SEOUL NATIONAL UNIV. VISION & LEARNING

NVIDIA

Microsoft Research

Are existing audio-visual datasets large enough?

Visual-Audio datasets

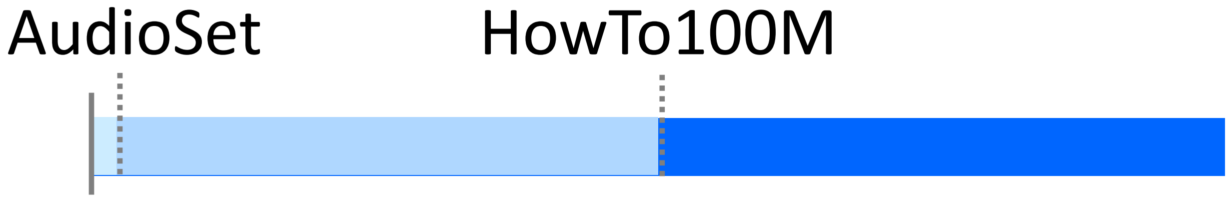
- Kinetics-Sounds
- VGG-Sound
- AudioSet

Visual-Text datasets

- HowTo100M

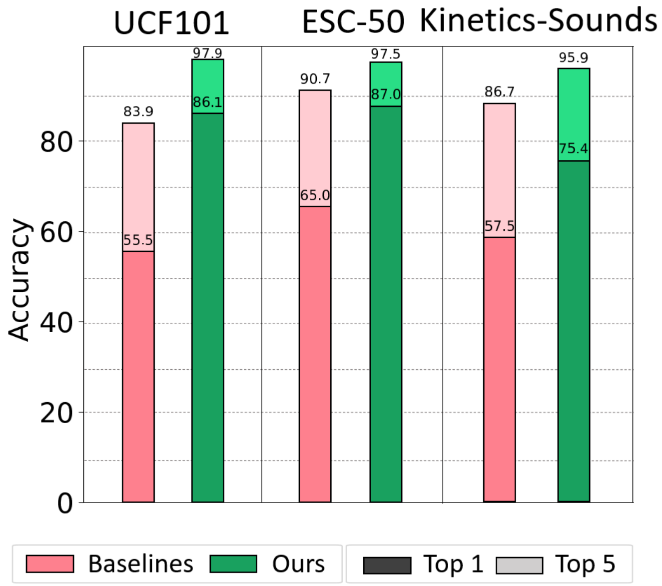
ACAV100M: A new video dataset for **audio-visual** learning

ACAV100M (31 years)



- **Two orders of magnitude larger** than the current largest video dataset used in the audio-visual learning literature: AudioSet **(8 months)**
- **Twice as large** as the largest video dataset: HowTo100M **(15 years)**

- Best performance in downstream tasks



The curation process should be **automatic** for **scalability**

There is no **large-scale (100M) audio-visual** dataset

Visual-Text

HowTo100M (136M clips)

Visual-Audio

?

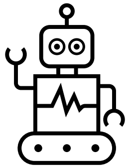
Miech et al. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *ICCV*

The curation process should be **automatic** for **scalability**

There is no **large-scale (100M) audio-visual** dataset
since it is hard to **scale up the curation process**

Visual-Text

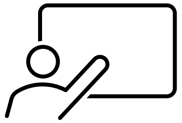
Videos
↓ ← YouTube ASR
Subtitles
↓ ← Keyword Search
Curated Data



Fully
Automatic

Visual-Audio

Metadata
↓ ← Keyword Search
Raw Videos
↓ ← Human Filtering
Curated Data

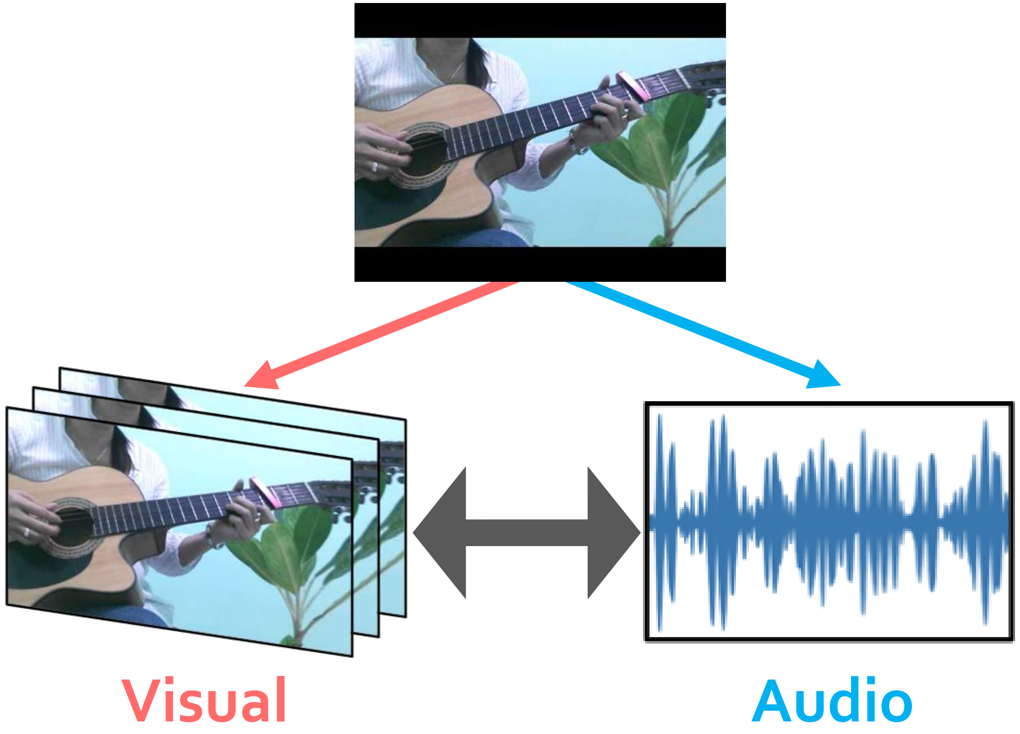


Relying on
Human Effort

What **criterion** should we use for **data construction?**

Recent self-supervised learning tasks leverage audio-visual correspondence

Goal:
Find a **subset** of videos with maximum **AV Correspondence**

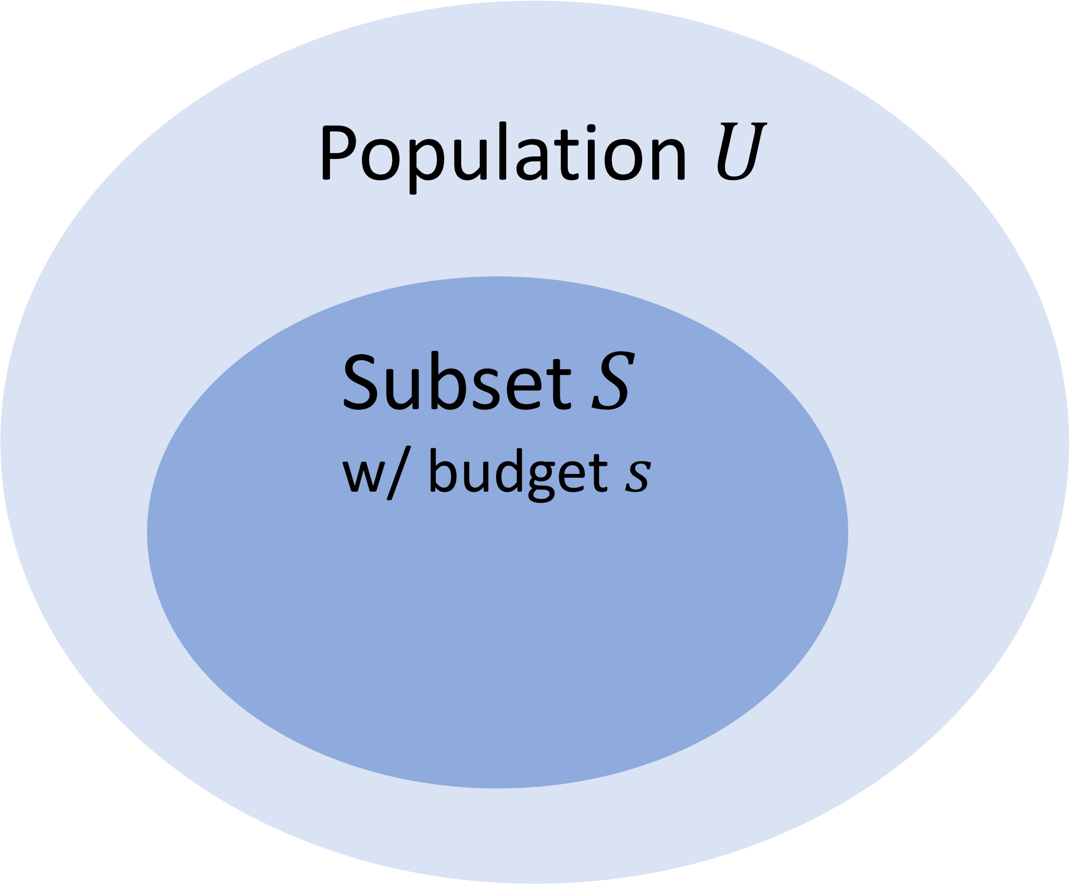


High AV Correspondence

Korbar et al. 2018. Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization. *NeurIPS*
Morgado et al. 2021. Audio-Visual Instance Discrimination with Cross-Modal Agreement. *CVPR*

Subset maximization idea:

Find a **subset that maximizes the MI** between audio and visual channels



$$\max_{S \subset U} \sum_{i \in S} MI(A_i, V_i) \;\; s.t. |S| = s$$

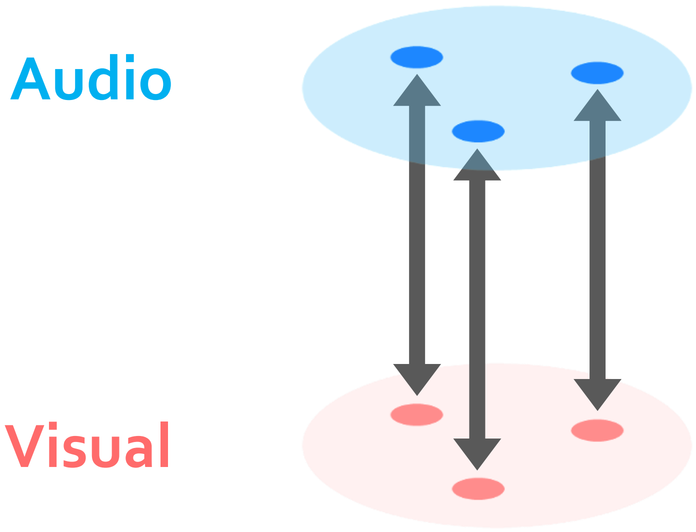
Challenge:
How to **estimate** MI
over high dimensional signals

MI **estimation**: instance-level vs. set-level

MI estimators can utilize instance-level or set-level information

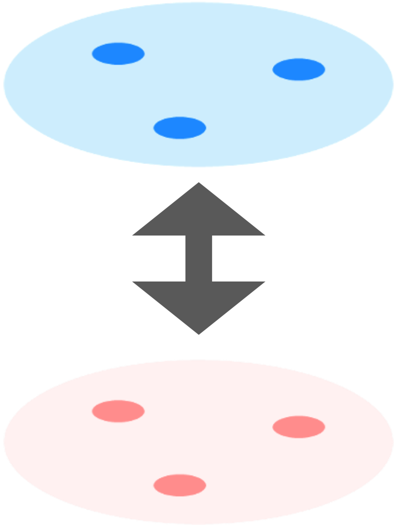
We opt for **set-level** method due to its superior empirical performance

Instance-Level

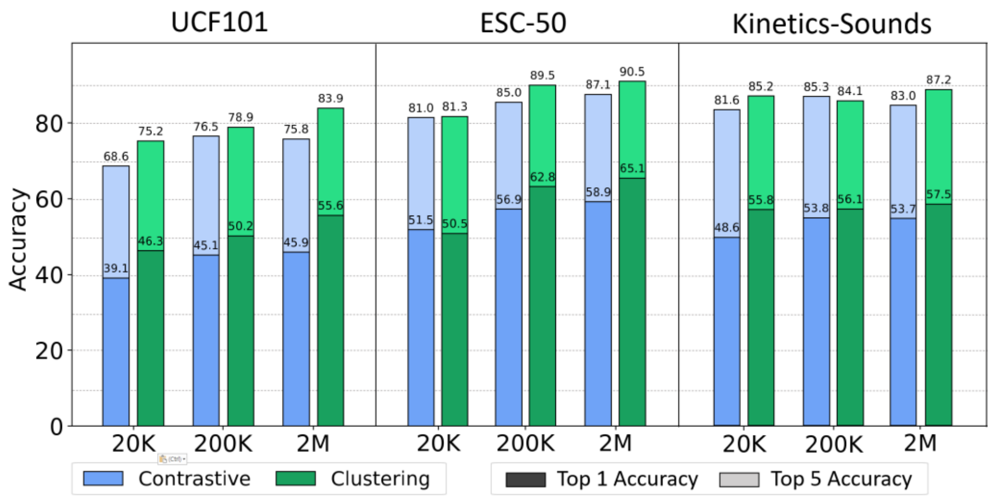


e.g. contrastive learning

Set-Level



e.g. clustering



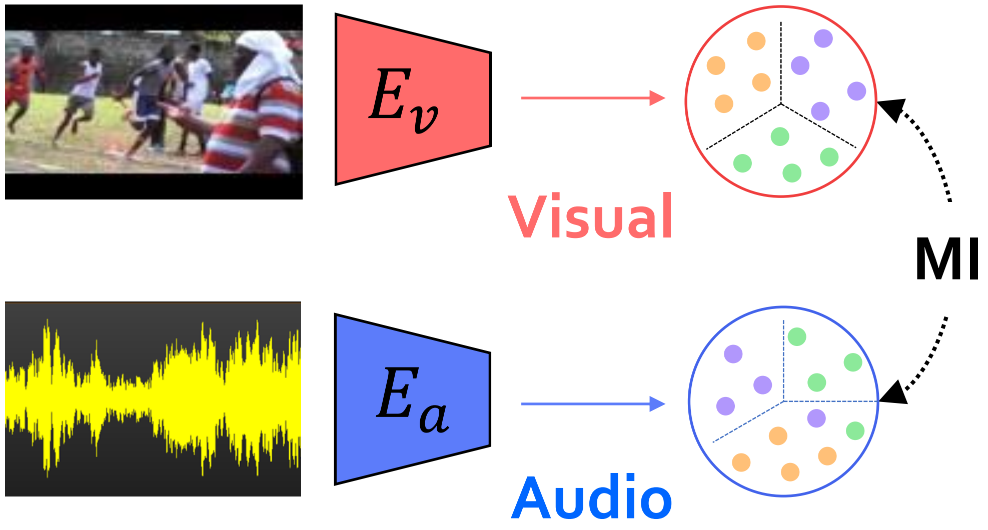
MI **estimation**: implementation

Estimate MI in a discrete space by clustering audio and visual signals, respectively

MI Estimator

$$MI(\mathcal{A},\mathcal{V}) = \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{V}|}\frac{|A_i \cap V_j|}{|X|}\log\frac{|X||A_i \cap V_j|}{|A_i||V_j|}$$

X: Raw dataset
$\mathcal{A} = \{A_i, \ldots, A_{|\mathcal{A}|}\}$: Partitions of X w.r.t. audio clustering
$\mathcal{V} = \{V_j, \ldots, V_{|\mathcal{V}|}\}$: Partitions of X w.r.t. visual clustering



Scalability of the selection algorithm

Estimate MI in a **discrete space** induced by clustering
-> Combinatorial subset selection problem (**NP-Hard**)

We exploit the most scalable approximation (batch-greedy)

Approximation
Scalability ↑

$\boldsymbol{O(2^N)}$: **Brute-Force**

$\boldsymbol{O(N^2)}$: **Greedy**

Batch Trick

$\boldsymbol{O(N \times B)}$: **Batch Greedy**

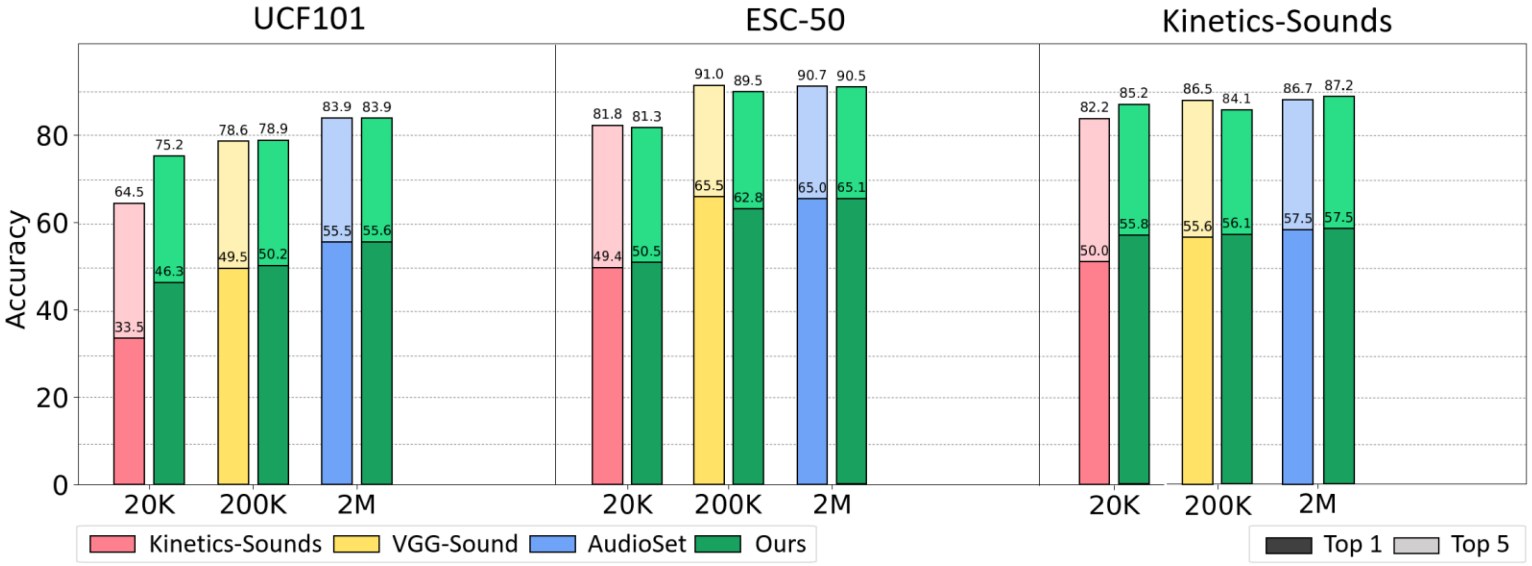
$\boldsymbol{B}$: Mini-batch size

Chen and Krause. 2013. Near-optimal Batch Mode Active Learning and Adaptive Submodular Optimization. *ICML*

Results on Real-World Problems

Linear evaluation on visual, audio and audio-visual classification tasks

Visual **Audio** **Audio-Visual**



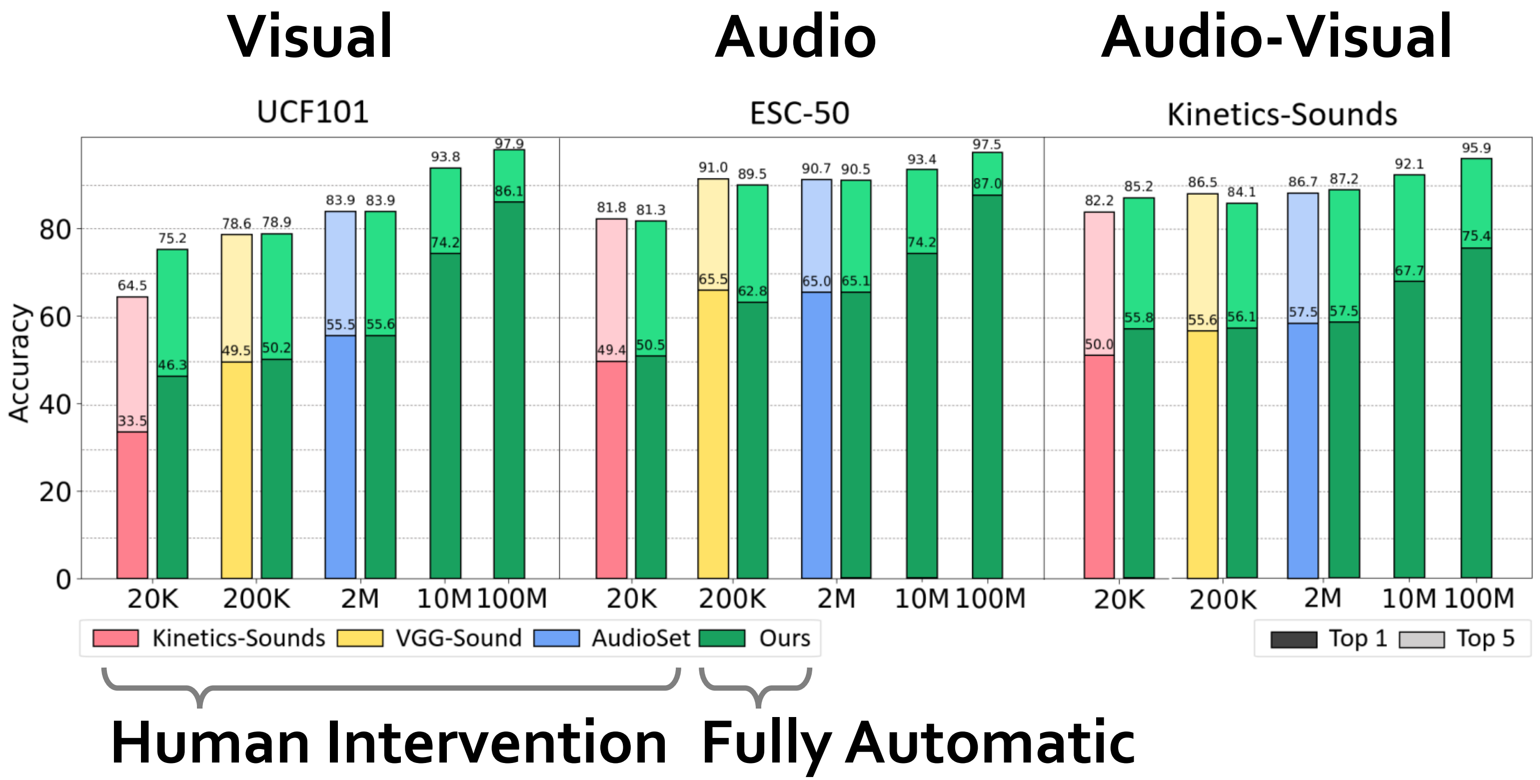
Human Intervention **Fully Automatic**

Our **automatic** pipeline achieves slightly better or comparable to the baselines **without human effort**

Soomro et al. 2012. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. *CRCV-TR-12-01*
Piczak. 2015. ESC: Dataset for Environmental Sound Classification. *ACM-MM*
Arandjelovic and Zisserman. 2017. Look, Listen and Learn. *ICCV*

Results on Real-World Problems

Linear evaluation on visual, audio and audio-visual classification tasks



Our **automatic** pipeline achieves slightly better or comparable to the baselines **without human effort**

Scalable to **10M/100M** videos with **best performances**

Soomro et al. 2012. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. *CRCV-TR-12-01*
Piczak. 2015. ESC: Dataset for Environmental Sound Classification. *ACM-MM*
Arandjelovic and Zisserman. 2017. Look, Listen and Learn. *ICCV*

Video Diversity

Our curation process
is not confined to
a human-defined
taxonomy of concepts

Thus, our datasets contain
diverse concepts
such as shoes unboxing

Project Webpage

We provide the dataset, paper, code and sample explorers from the webpage
https://acav100m.github.io/