Synchformer: Efficient Synchronization from Sparse Cues

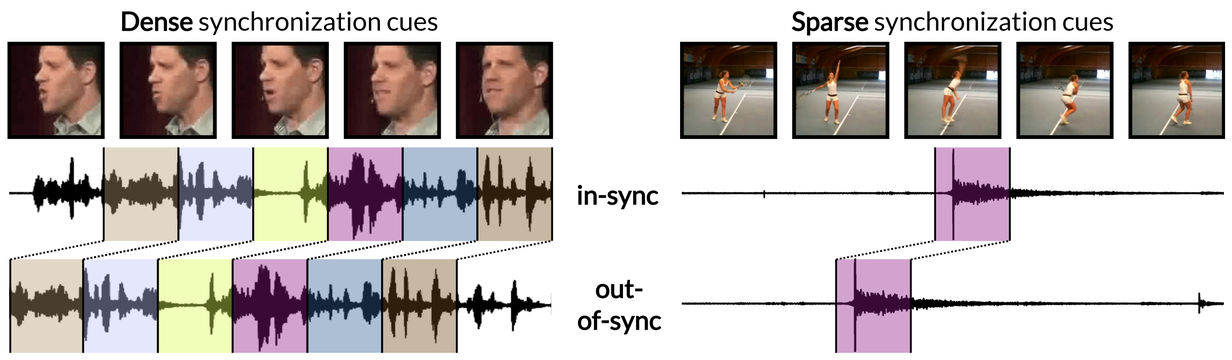
Vladimir Iashin [1,3] Weidi Xie [2,3] Esa Rahtu [1] Andrew Zisserman [3]

[1]Tampere University [2]Shanghai Jiao Tong University [3]University of Oxford

Goal

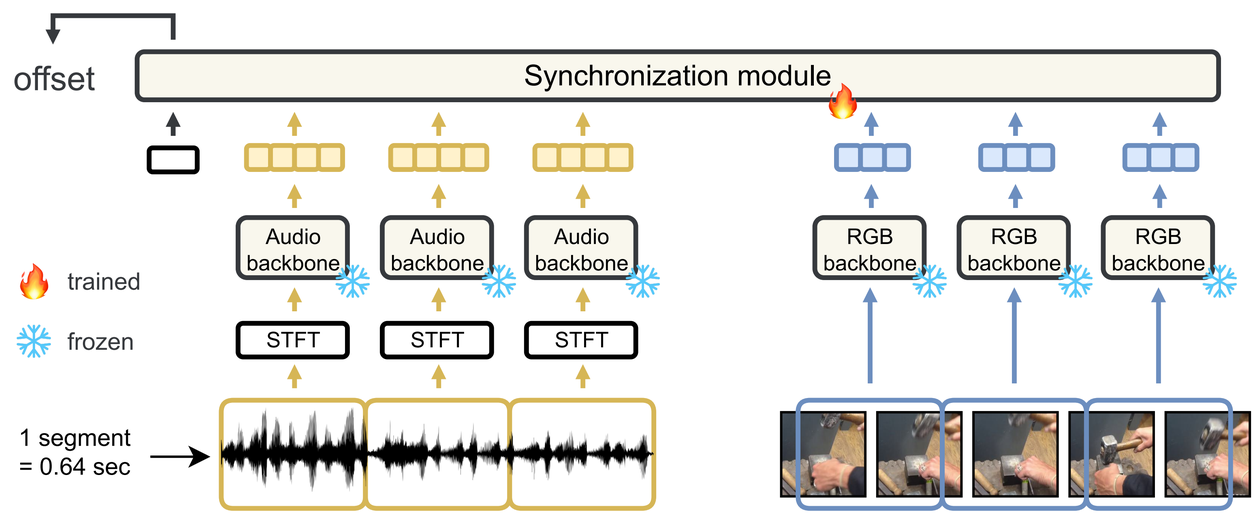
- Detect if a video is synchronized and, if not, what is the offset between audio and video
- Tell if a video is synchronizable at all

Challenges



- Cues rare in time and space → high resolution inputs in space and time
- Variety of classes → more training parameters
- Temporal artefacts in data → model learns a shortcut

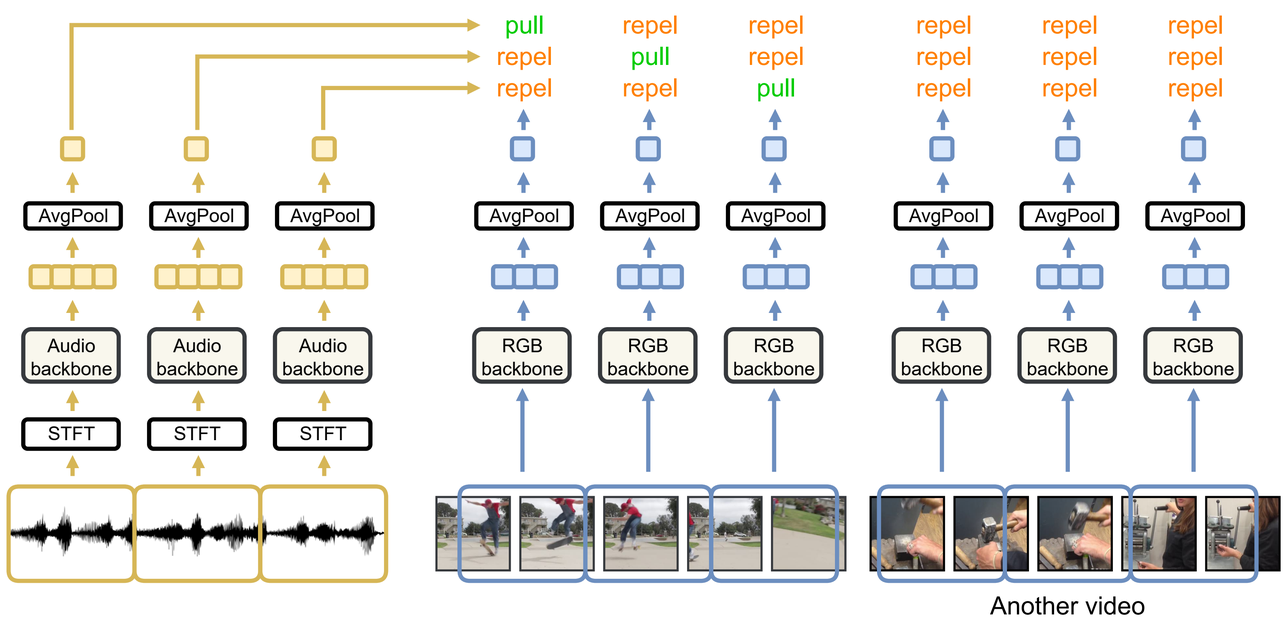
Synchformer



1. Audio and visual streams are split into short segments, e.g. 0.64 seconds
2. Spectrogram is extracted from an audio
3. Temporal features are extracted from the spectrogram and stacks of (16) RGB frames
4. Audio and visual features from all segments are concatenated
5. Sync module uses the features to predict the temporal offset for synchronization

Training

Step 1: Segment-level contrastive pre-training of feature extractors



- **Input**: 14 segments from one video + 14 segments from another video (negatives)
- **Outputs**: logits from each audio segment and visual segment
- **Loss**: Symmetric Info-NCE (from-audio-to-visual and vice-versa — averaged)

Step 2: Freeze (❄️) feature extractors and train (🔥) synchronization model (prev. section)

- **Input**: ≈5 sec video, randomly offset by one of 21 cls [−2.0, −1.8, ..., 0.0, ..., +1.8, +2.0]
- **Output**: an offset distribution across the 21 classes
- **Loss**: Cross-entropy

Datasets

Dataset	Type	Classes	Train size
LRS3 (“Full scene”)	Dense	1	≈43k
VGGSound	Sparse	≈300	≈180k
AudioSet	Sparse	≈500	≈1600k

Results

Dense: evaluated on the test set of LRS3 (“full scene”)

	Training Dataset	Acc@1 / Acc@1±1 cls
(Chen *et al.*, 2021)	LRS3 (“full scene”)	58.6 / 85.3
(Iashin *et al.*, 2022)	LRS3 (“full scene”)	80.8 / 96.9
Ours	LRS3 (“full scene”)	**86.5** / **99.6**

Metrics are accuracy across 21 classes with and without ±0.2 sec tolerance

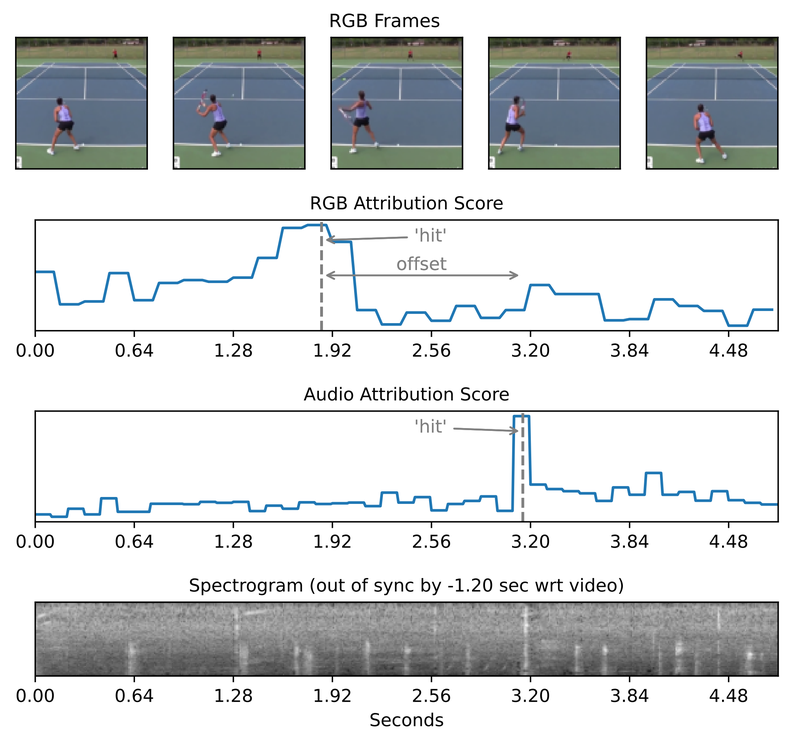
Sparse: the ‘sparse’ subset of videos from test set of VGGSound

	Training Dataset	Acc@1 / Acc@1±1 cls
(Iashin *et al.*, 2022)	VGGSound	43.4 / 62.1
Ours	VGGSound	**52.9** / **70.1**
(Iashin *et al.*, 2022)	AudioSet	40.0 / 63.0
Ours	AudioSet	**54.6** / **77.6**

Metrics are accuracy across 21 classes with and without ±0.2 sec tolerance

Visualizing evidence attribution

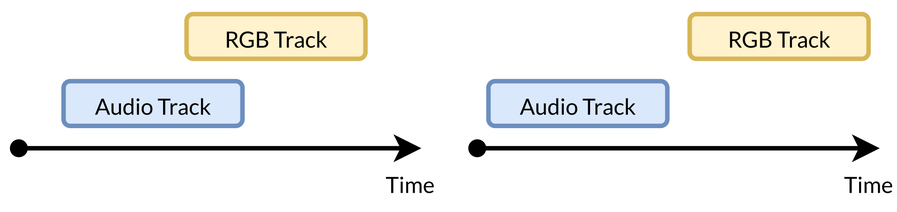
Given a video, which parts of it are important for prediction?



Attribution scores: Randomly corrupt inputs, observe model’s offset predictions.

New task: Predicting synchronizability

What makes a video synchronizable? → content overlap



Training for synchronizability

- Replace the classifier and fine-tune **Synchformer** with 2 classes
  - “*no content overlap*”, “*some content overlap*”

Results