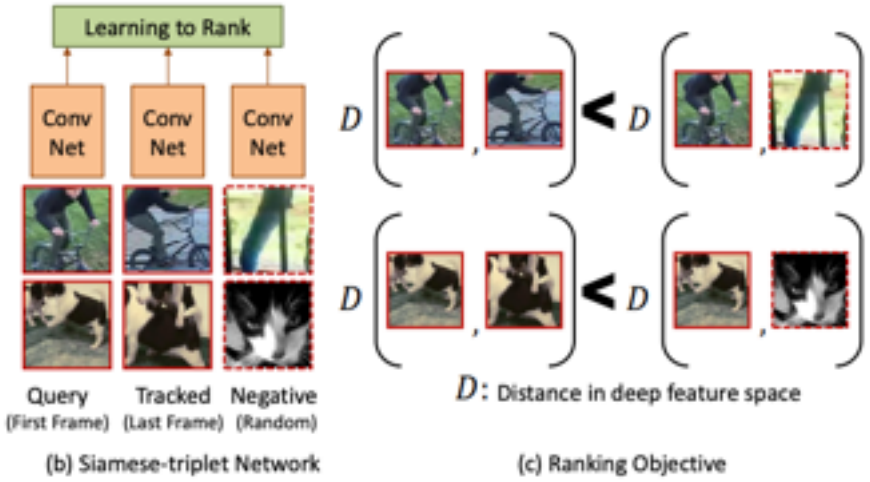
Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization

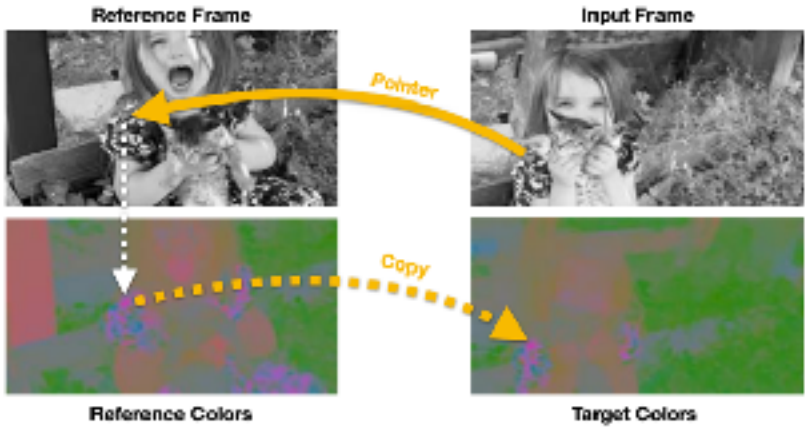
NeurIPS 2018

Bruno Korbar, Du Tran, Lorenzo Torresani

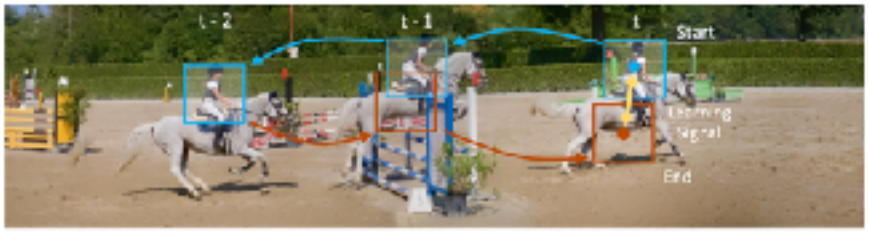
Learning Video Representations from Visual Data



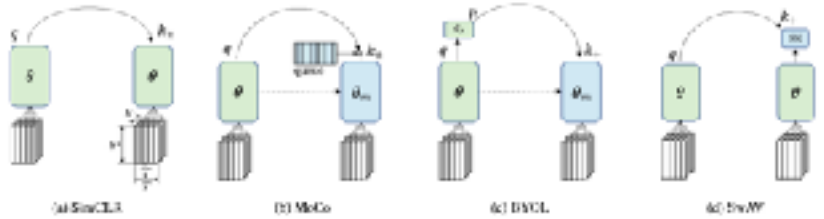
Wang et al. “Unsupervised Learning of Visual Representations using Videos”, ICCV 2015



Vondrick et al. “Tracking Emerges by Colorizing Videos”, ECCV 2018



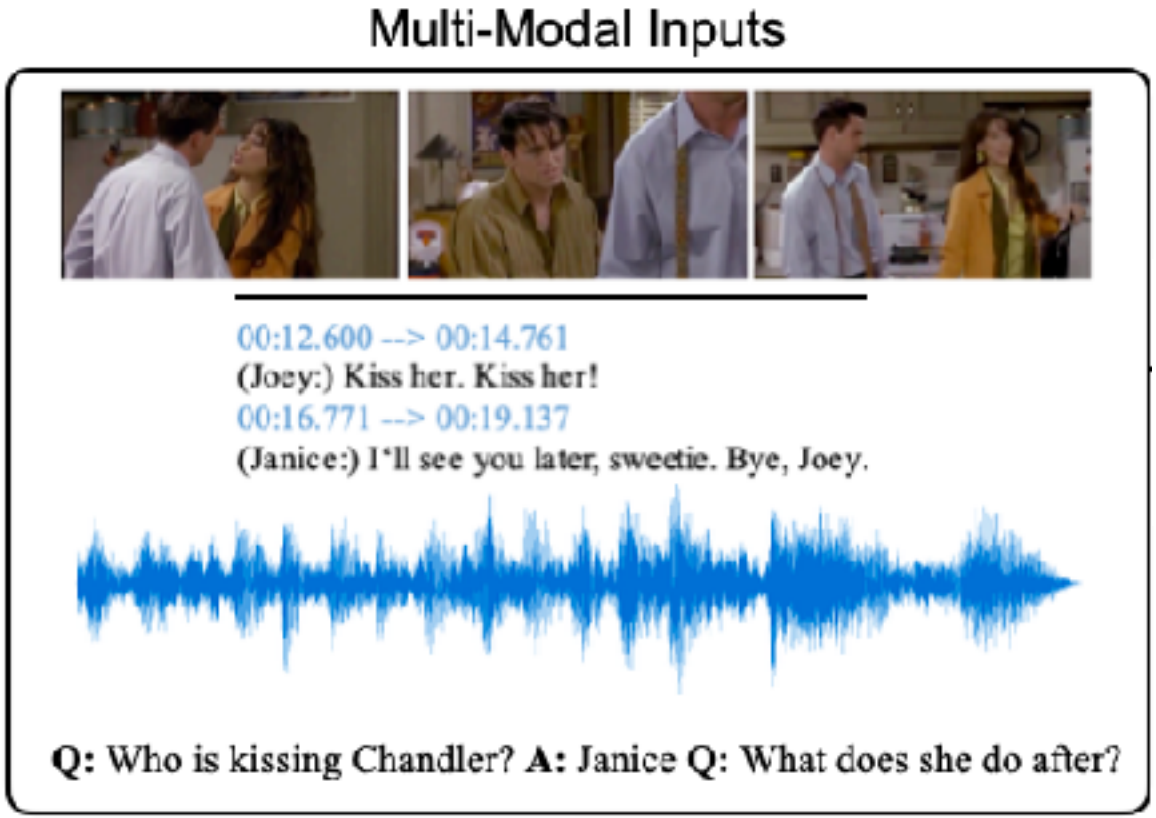
Wang et al. “Learning Correspondence from the Cycle-consistency of Time”, CVPR 2019



Feichtenhofer et al. “A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning”, CVPR 2021

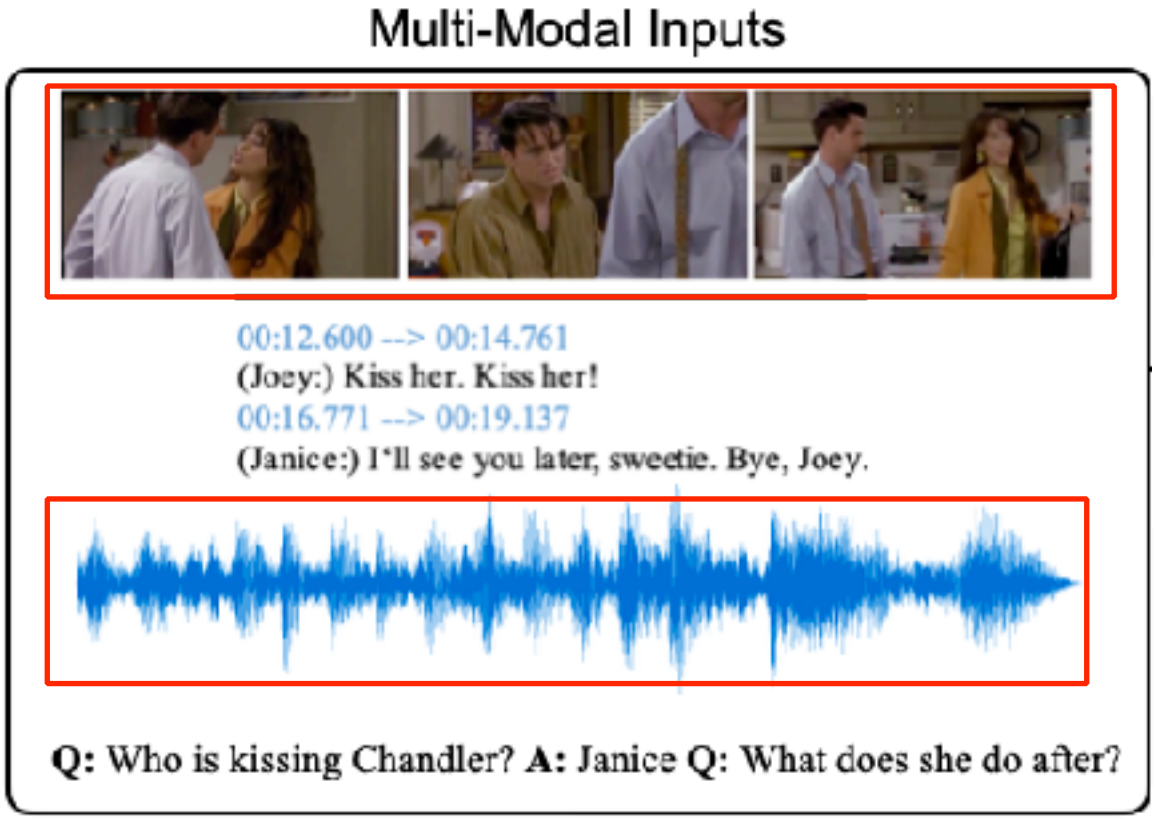
Motivation

- We want to use signals from different modalities (e.g., speech, audio, text, etc.) to learn better video representations.



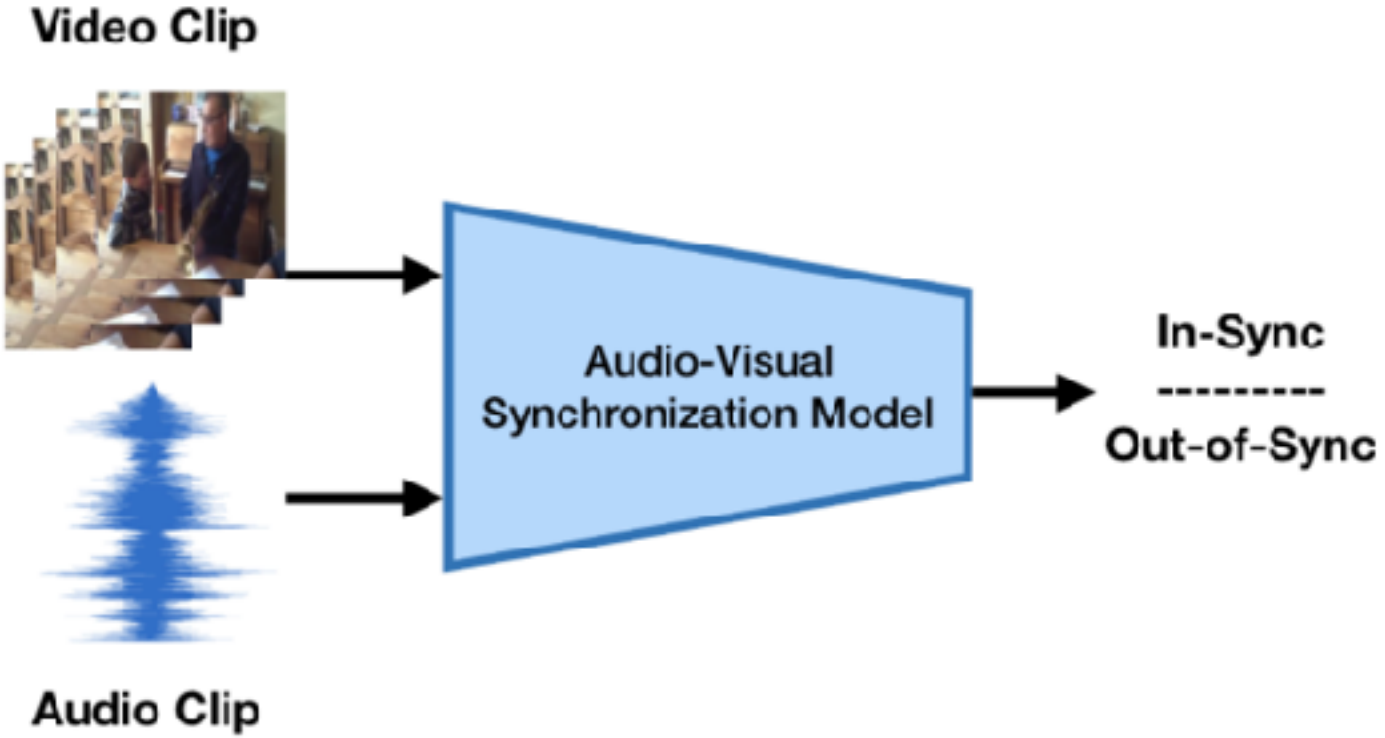
Motivation

- We want to use signals from different modalities (e.g., speech, audio, text, etc.) to learn better video representations.



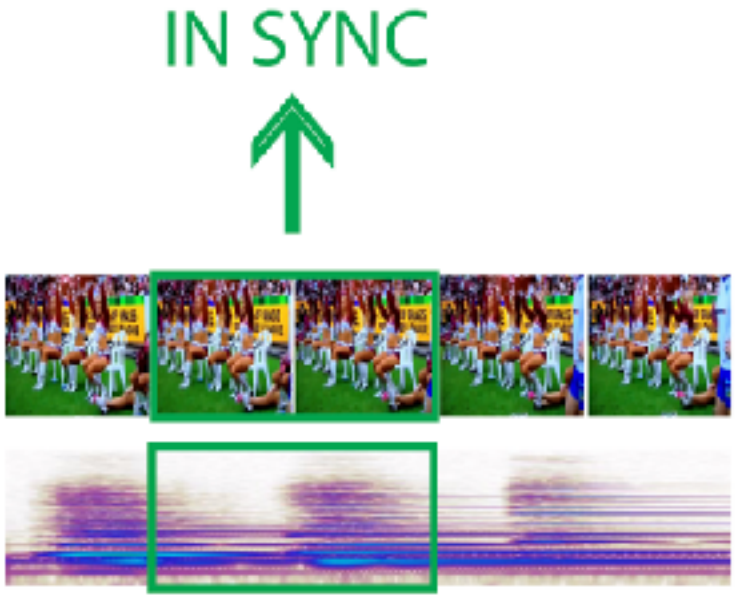
Audio-Visual Temporal Synchronization

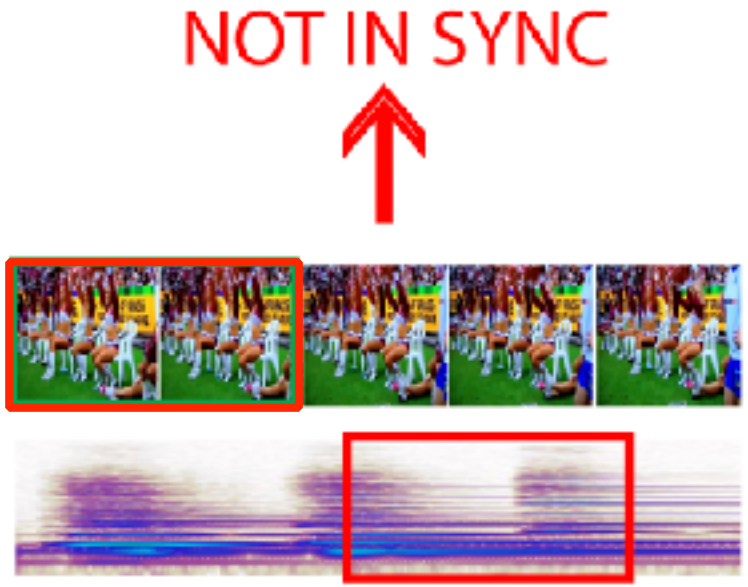
- Can we learn general audio and video models from self-supervised synchronization?



Audio-Visual Temporal Synchronization

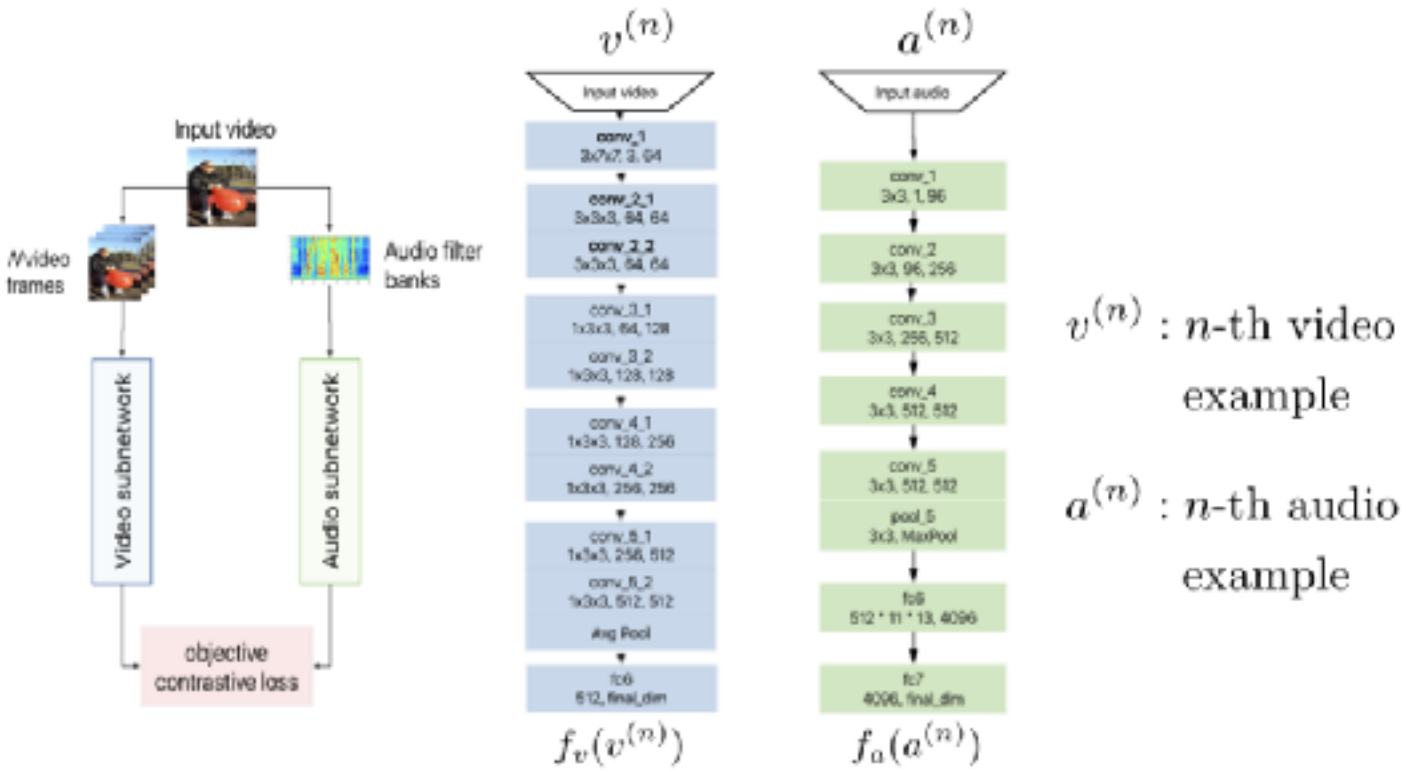
- The video model is trained to recognize temporal audio-visual synchronization.





Architecture

- Separate audio and video subnetworks allow feature extraction and finetuning on single modality.



Loss Function

- The model is trained by minimizing the contrastive loss function.
- The authors optimize the audio and video streams to produce small distance on positive pairs and larger distance on negative pairs.

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

$$y^{(n)} = \begin{cases} 1: & \text{if the examples are in sync} \\ 0: & \text{otherwise} \end{cases}$$

Loss Function

- The model is trained by minimizing the contrastive loss function.
- The authors optimize the audio and video streams to produce small distance on positive pairs and larger distance on negative pairs.

Minimize distance on positive pairs

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

$$y^{(n)} = \begin{cases} 1: & \text{if the examples are in sync} \\ 0: & \text{otherwise} \end{cases}$$

Loss Function

- The model is trained by minimizing the contrastive loss function.
- The authors optimize the audio and video streams to produce small distance on positive pairs and larger distance on negative pairs.

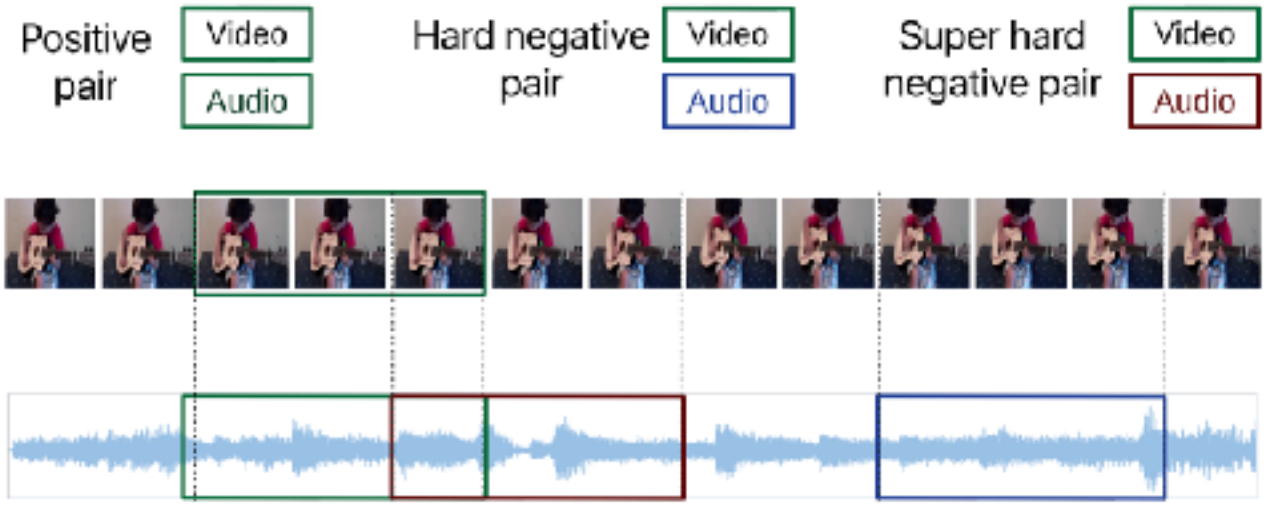
Maximize distance on negative pairs

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

$$y^{(n)} = \begin{cases} 1: & \text{if the examples are in sync} \\ 0: & \text{otherwise} \end{cases}$$

Selection of Negative Examples

- "easy" negatives (audio and video come from different samples).
- "hard" negatives (same sample, audio and video are non-overlapping).
- "super-hard" (same sample, audio and video are overlapping but still out of sync).



Curriculum Learning

- Progressively increasing the difficulty of the problem yields accuracy gains on downstream tasks.
- Audiovisual Temporal Synchronization accuracy is evaluated on the Kinetics-400 test set, which includes only negatives of “easy” type.

Method	Negative type	Epochs	Accuracy (%)
Single learning stage	easy	1 - 90	69.0
	75% easy, 25% hard	1 - 90	58.9
	hard	1 - 90	52.3
	easy	1 - 50	67.2
Curriculum learning (i.e., second learning stage applied after a first stage of 1-50 epochs with easy negatives only)	75% easy, 25% hard	51 - 90	**78.4**
	hard	51 - 90	65.7

Action Recognition Experiments

- The authors assess the effectiveness of the AVTS learned representation for downstream action recognition tasks.
- The pretrained model is fine-tuned on UCF101 and HMDB51.
- No labeled data is used during pretraining.

Video Network Architecture	Pretraining Dataset	Pretraining Supervision	UCF101	HMDB51
MC2	none	N/A	67.2	41.2
MC2	Kinetics	self-supervised (AVTS)	83.6	54.3
MC2	Kinetics	fully supervised (action labels)	87.9	62.0
MC3	none	N/A	69.1	43.9
MC3	Kinetics	self-supervised (AVTS)	85.8	56.9
MC3	Audioset	self-supervised (AVTS)	89.0	61.6
MC3	Kinetics	fully supervised (action labels)	90.5	66.8
I3D-RGB	none	N/A	57.1	40.0
I3D-RGB	Kinetics	self-supervised (AVTS)	83.7	53.0
I3D-RGB*	Imagenet	fully supervised (object labels)	84.5	49.8
I3D-RGB*	Kinetics	fully supervised (action labels)	95.1	74.3
I3D-RGB*	Kinetics + Imagenet	fully supervised (object+action labels)	95.6	74.8

Action Recognition Experiments

- The authors assess the effectiveness of the AVTS learned representation for downstream action recognition tasks.
- The pretrained model is fine-tuned on UCF101 and HMDB51.
- No labeled data is used during pretraining.

Video Network Architecture	Pretraining Dataset	Pretraining Supervision	UCF101	HMDB51
MC2	none	N/A	67.2	41.2
MC2	Kinetics	self-supervised (AVTS)	83.6	54.3
MC2	Kinetics	fully supervised (action labels)	87.9	62.0
MC3	none	N/A	69.1	43.9
MC3	Kinetics	self-supervised (AVTS)	85.8	56.9
MC3	Audioset	self-supervised (AVTS)	89.0	61.6
MC3	Kinetics	fully supervised (action labels)	90.5	66.8
I3D-RGB	none	N/A	57.1	40.0
I3D-RGB	Kinetics	self-supervised (AVTS)	83.7	53.0
I3D-RGB*	Imagenet	fully supervised (object labels)	84.5	49.8
I3D-RGB*	Kinetics	fully supervised (action labels)	95.1	74.3
I3D-RGB*	Kinetics + Imagenet	fully supervised (object+action labels)	95.6	74.8

Action Recognition Experiments

- The authors assess the effectiveness of the AVTS learned representation for downstream action recognition tasks.
- The pretrained model is fine-tuned on UCF101 and HMDB51.
- No labeled data is used during pretraining.

Video Network Architecture	Pretraining Dataset	Pretraining Supervision	UCF101	HMDB51
MC2	none	N/A	67.2	41.2
MC2	Kinetics	self-supervised (AVTS)	83.6	54.3
MC2	Kinetics	fully supervised (action labels)	87.9	62.0
MC3	none	N/A	69.1	43.9
MC3	Kinetics	self-supervised (AVTS)	85.8	56.9
MC3	Audioset	self-supervised (AVTS)	89.0	61.6
MC3	Kinetics	fully supervised (action labels)	90.5	66.8
I3D-RGB	none	N/A	57.1	40.0
I3D-RGB	Kinetics	self-supervised (AVTS)	83.7	53.0
I3D-RGB*	Imagenet	fully supervised (object labels)	84.5	49.8
I3D-RGB*	Kinetics	fully supervised (action labels)	95.1	74.3
I3D-RGB*	Kinetics + Imagenet	fully supervised (object+action labels)	95.6	74.8

Action Recognition Experiments

- The authors assess the effectiveness of the AVTS learned representation for downstream action recognition tasks.
- The pretrained model is fine-tuned on UCF101 and HMDB51.
- No labeled data is used during pretraining.

Video Network Architecture	Pretraining Dataset	Pretraining Supervision	UCF101	HMDB51
MC2	none	N/A	67.2	41.2
MC2	Kinetics	self-supervised (AVTS)	83.6	54.3
MC2	Kinetics	fully supervised (action labels)	87.9	62.0
MC3	none	N/A	69.1	43.9
MC3	Kinetics	self-supervised (AVTS)	85.8	56.9
MC3	Audioset	self-supervised (AVTS)	89.0	61.6
MC3	Kinetics	fully supervised (action labels)	90.5	66.8
I3D-RGB	none	N/A	57.1	40.0
I3D-RGB	Kinetics	self-supervised (AVTS)	83.7	53.0
I3D-RGB*	Imagenet	fully supervised (object labels)	84.5	49.8
I3D-RGB*	Kinetics	fully supervised (action labels)	95.1	74.3
I3D-RGB*	Kinetics + Imagenet	fully supervised (object+action labels)	95.6	74.8

Audio Classification

- Evaluation of audio features learned with AVTS on two audio classification benchmarks: ESC-50 and DCASE2014.
- The audio sub-network is not fine-tuned on the target datasets.

Method	Auxiliary dataset	Auxiliary supervision	# auxiliary examples	ESC-50 accuracy (%)	DCASE2014 accuracy (%)
SVM-MFCC [29]	none	none	none	39.6	-
Random Forest [29]	none	none	none	44.3	-
Our audio subnet	none	none	none	61.6	72
SoundNet [20]	SoundNet	self	2M+	74.2	88
L^3-Net [21]	SoundNet	self	2M+	79.3	93
Our AVTS features	Kinetics	self	230K	76.7	91
Our AVTS features	AudioSet	self	1.8M	80.6	93
Our AVTS features	SoundNet	self	2M+	**82.3**	**94**
Human performance [21]	n/a	n/a	n/a	*81.3*	-
State-of-the-art (RBM)[31]	none	none	none	**86.5**	-

Audio Classification

- Evaluation of audio features learned with AVTS on two audio classification benchmarks: ESC-50 and DCASE2014.
- The audio sub-network is not fine-tuned on the target datasets.

Method	Auxiliary dataset	Auxiliary supervision	# auxiliary examples	ESC-50 accuracy (%)	DCASE2014 accuracy (%)
SVM-MFCC [29]	none	none	none	39.6	-
Random Forest [29]	none	none	none	44.3	-
Our audio subnet	none	none	none	61.6	72
SoundNet [20]	SoundNet	self	2M+	74.2	88
L^3-Net [21]	SoundNet	self	2M+	79.3	93
Our AVTS features	Kinetics	self	230K	76.7	91
Our AVTS features	AudioSet	self	1.8M	80.6	93
Our AVTS features	SoundNet	self	2M+	**82.3**	**94**
Human performance [21]	n/a	n/a	n/a	*81.3*	-
State-of-the-art (RBM)[31]	none	none	none	**86.5**	-

Multi-modal Action Recognition

- The results are evaluated on UCF-101.

Model	Accuracy (%)
Owens et al. (vision only) [5]	77.6
AVTS (vision only)	85.8
Owens et al. (multisensory) [5]	82.1
AVTS (multisensory)	**87.0**

Curriculum Learning

- Impact of curriculum learning on the downstream tasks performance.
- The evaluation is done for both audio classification and action recognition.

Method	AVTS-Kinetics	ESC-50	DCASE	HMDB51	UCF101
Our AVTS - single stage	69.8	70.6	89.2	46.4	77.1
Our AVTS - curriculum	78.4	82.3	94.1	56.9	85.8
L^3-Net	74.3	79.3	93	40.2	72.3

Curriculum Learning

- Impact of curriculum learning on the downstream tasks performance.
- The evaluation is done for both audio classification and action recognition.

Method	AVTS-Kinetics	ESC-50	DCASE	HMDB51	UCF101
Our AVTS - single stage	69.8	70.6	89.2	46.4	77.1
Our AVTS - curriculum	78.4	82.3	94.1	56.9	85.8
L^3-Net	74.3	79.3	93	40.2	72.3

Curriculum learning helps downstream audio classification tasks

Curriculum Learning

- Impact of curriculum learning on the downstream tasks performance.
- The evaluation is done for both audio classification and action recognition.

Method	AVTS-Kinetics	ESC-50	DCASE	HMDB51	UCF101
Our AVTS - single stage	69.8	70.6	89.2	46.4	77.1
Our AVTS - curriculum	78.4	82.3	94.1	56.9	85.8
L^3-Net	74.3	79.3	93	40.2	72.3

Curriculum learning helps downstream action recognition tasks

Contributions

- Simple and elegant self-supervised pretraining scheme on audiovisual video data.
- Effective curriculum learning approach.
- Impressive performance on both video-only, audio-only, and video+audio classification benchmarks.

Discussion Questions

- Why use two separate video and audio streams as opposed to a single unified audiovisual model?

Discussion Questions

- Why use two separate video and audio streams as opposed to a single unified audiovisual model?
- What's the advantage of using video+audio data for self-supervised learning compared to using video-only data?