SpeedNet: Learning the Speediness in Videos

Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman,
Michael Rubinstein, Michal Irani, Tali Dekel
Google Research, Tel Aviv University, Weizmann Institute
CVPR, 2020

Speediness

Slower

Normal

Faster

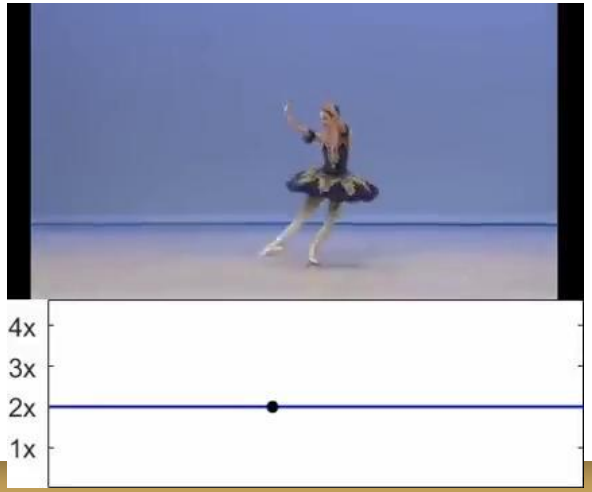
Speediness

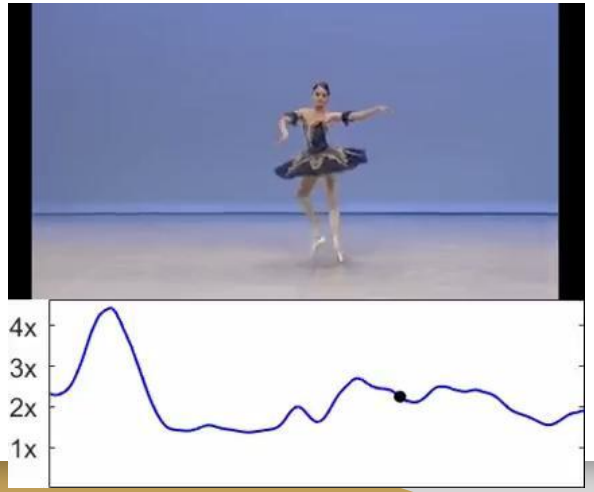
Motivation

- Video playback speed classification
  - Detecting replays in sports

Motivation

- Video playback speed classification
  - Detecting replays in sports
- Video time remapping
  - Generate adaptive speedup videos, depending on the speediness score, so that when sped up, their motion looks more natural to the viewer





Motivation

- Video playback speed classification
  - Detecting replays in sports
- Video time remapping
  - Generate adaptive speedup videos, depending on the speediness score, so that when sped up, their motion looks more natural to the viewer
- Self-supervised learning from videos
  - SpeedNet learns a powerful spacetime representation that can be used for self-supervised action recognition and for video retrieval

Challenges

- high-level reasoning is necessary for solving this task

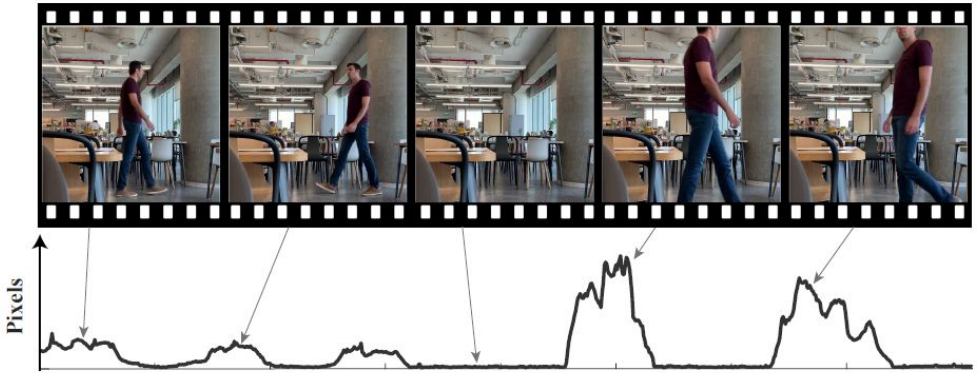
Challenges

- high-level reasoning is necessary for solving this task
- Avoid tendency to detect easy shortcuts specific to the dataset
  - Rely on artificial, low-level cues, such as compression artifacts

Challenges

- high-level reasoning is necessary for solving this task
- Avoid tendency to detect easy shortcuts specific to the dataset
  - Rely on artificial, low-level cues, such as compression artifacts

- Trivial case of motion magnitude => speediness (optical flow)
  - Two people walking normally at two different distances from the camera
  - Normal video of hare vs Sped up video of turtle

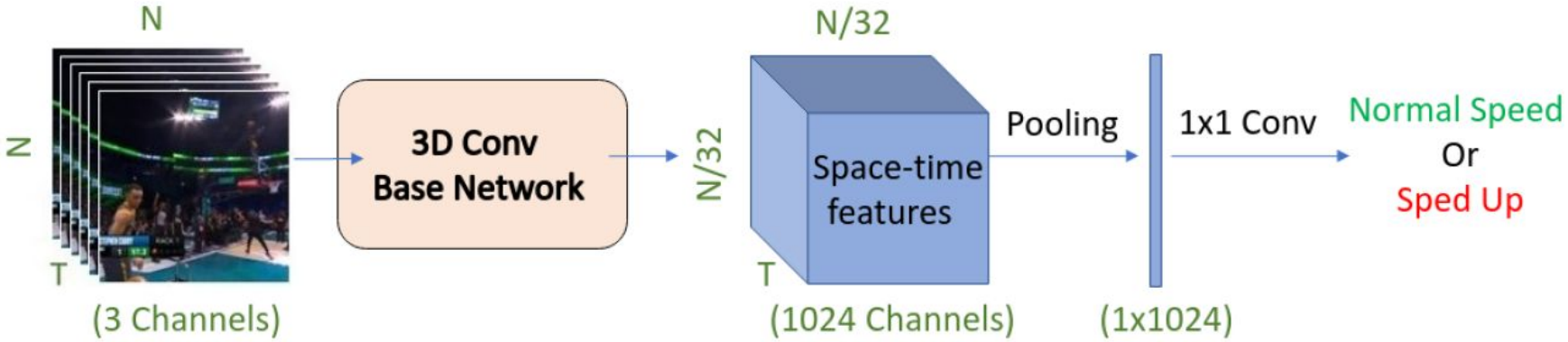


(a) Motion Magnitude (per frame)

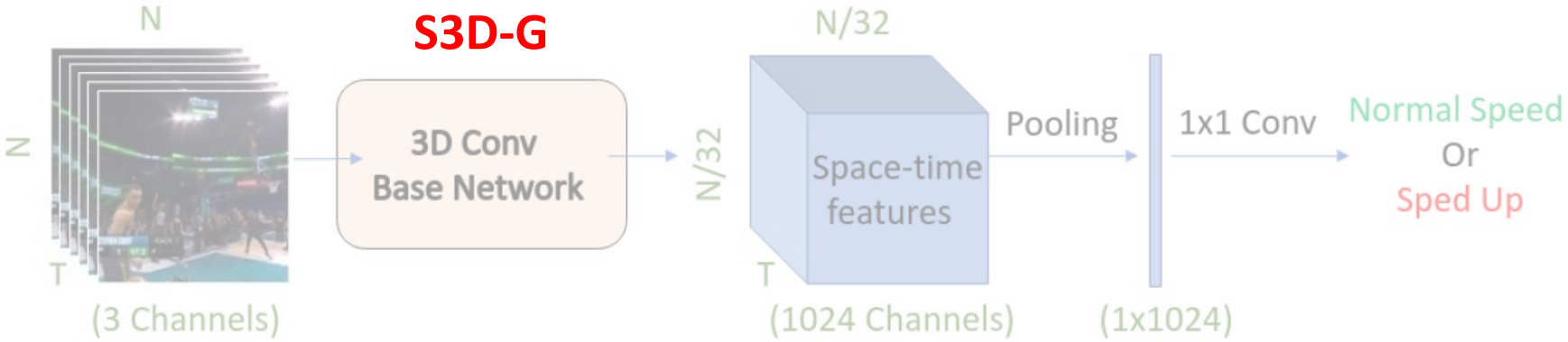
SpeedNet

- Classify whether an object in an input video sequence is moving at its normal speed, or faster than the normal speed
  - Given a set of L frames in an L-fps video, predict whether those frames depict 1 second of the object's motion (normal), or more than 1 second (sped up) [specifically 2 seconds]
- To determine whether motion in a video is natural or not, a regression objective may be unnecessarily difficult to learn

SpeedNet Architecture



SpeedNet Architecture



most dominant spatially moving object
∴ spatially apply global max pooling

temporally avoid sensitivity to instantaneous “spiky” motions
∴ temporally apply global average pooling.

SpeedNet - avoiding artificial cues

Spatial augmentations

- randomly resize the input video clip to a spatial dimension N
  - blurring during resize process can help mitigate potential pixel intensity jitter caused by compression
  - Since the input is of variable size, space-time features correspond to differently sized regions in the unresized input. This forces network not to rely only on size-dependent factors, such as motion magnitude

SpeedNet - avoiding artificial cues

Spatial augmentations

Temporal augmentations

- Introduce variability in the time domain
- Normal speed => rate of 1-1.2
- Spedup version => rate of 1.7-2.2

SpeedNet – avoiding artificial cues

Spatial augmentations

Temporal augmentations

Same-batch training

- Each batch contains both normal-speed and sped-up versions of each video clip

Experiments on SpeedNet

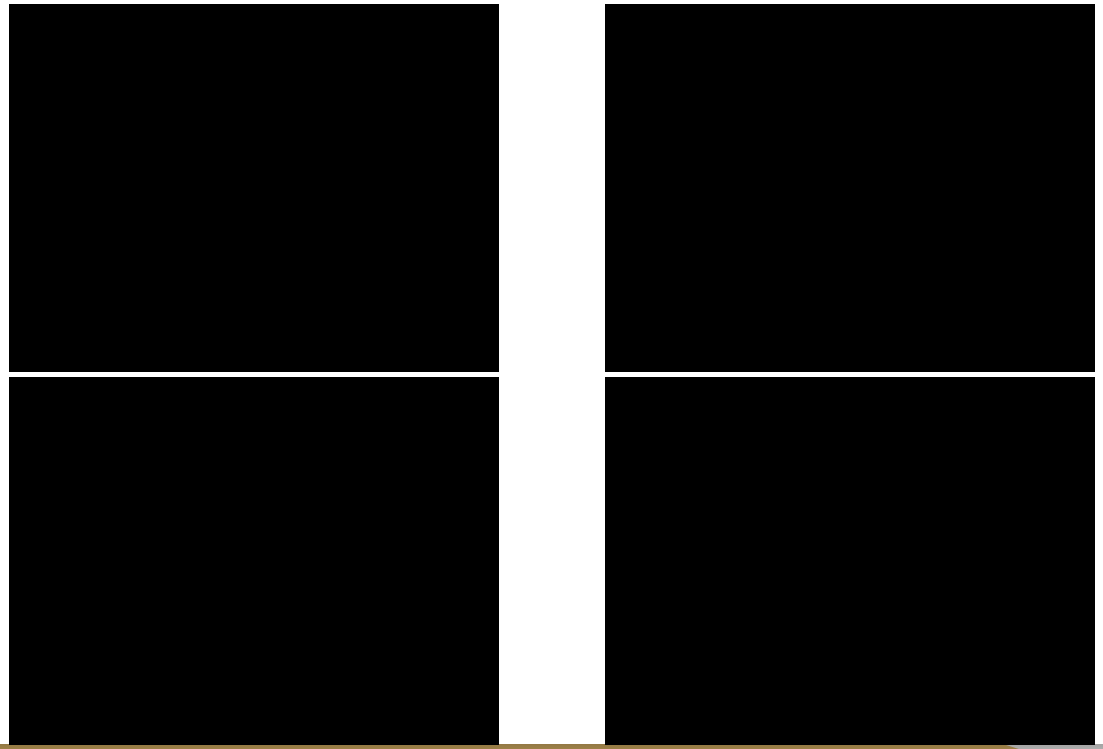
Train

- **Kinetics train set** (Normal = 1X | Spedup = 2X)

Test

- **Kinetics test set** (Normal = 1X | Spedup = 2X)
- **Need for Speed dataset (NFS)** (Normal = 10X | Spedup = 20X)

Experiments on SpeedNet



Experiments on SpeedNet

Model Type			Accuracy	
Batch	**Temporal**	**Spatial**	**Kinetics**	**NFS**
Yes	Yes	Yes	75.6%	73.6%
No	Yes	Yes	88.2%	59.3%
No	No	Yes	90.0%	57.7%
No	No	No	96.9%	57.4%
	Mean Flow		55.8%	55.0%

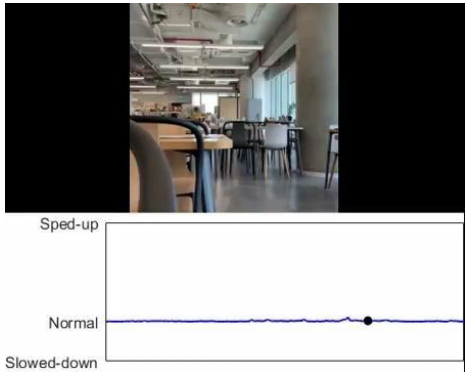
Experiments on SpeedNet

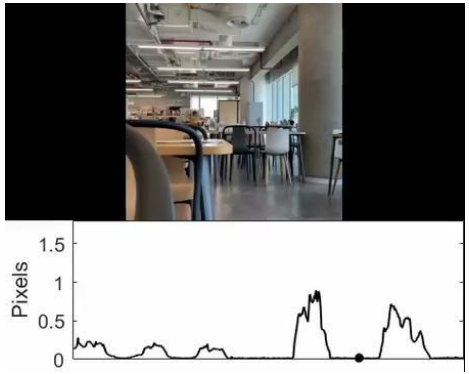
Batch	Model Type Temporal	Spatial	Accuracy Kinetics	NFS
Yes	Yes	Yes	75.6%	73.6%
No	Yes	Yes	88.2%	59.3%
No	No	Yes	90.0%	57.7%
No	No	No	96.9%	57.4%
	Mean Flow		55.8%	55.0%

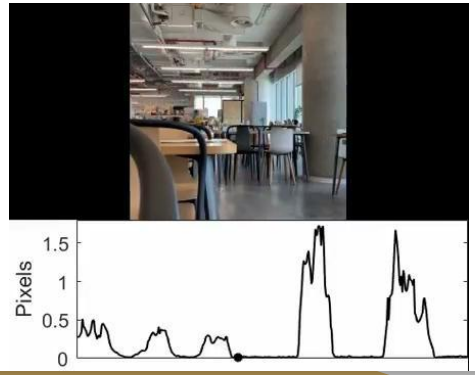
Experiments on SpeedNet

Batch	Model Type: Temporal	Model Type: Spatial	Accuracy: Kinetics	Accuracy: NFS
Yes	Yes	Yes	75.6%	73.6%
No	Yes	Yes	88.2%	59.3%
No	No	Yes	90.0%	57.7%
No	No	No	96.9%	57.4%
	Mean Flow		55.8%	55.0%

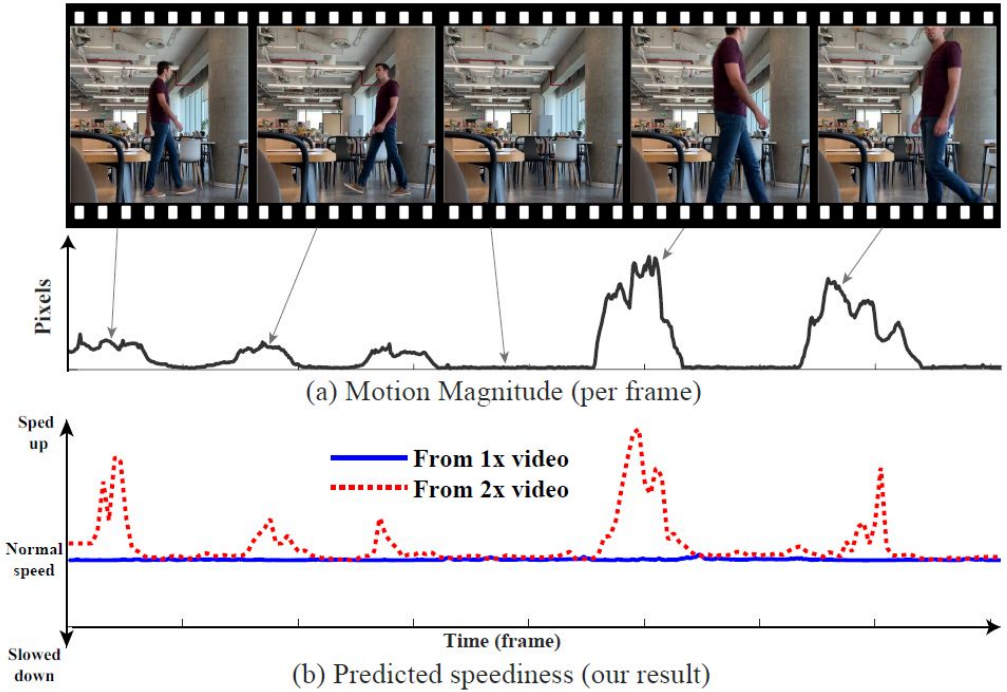
SpeedNet - motion magnitude



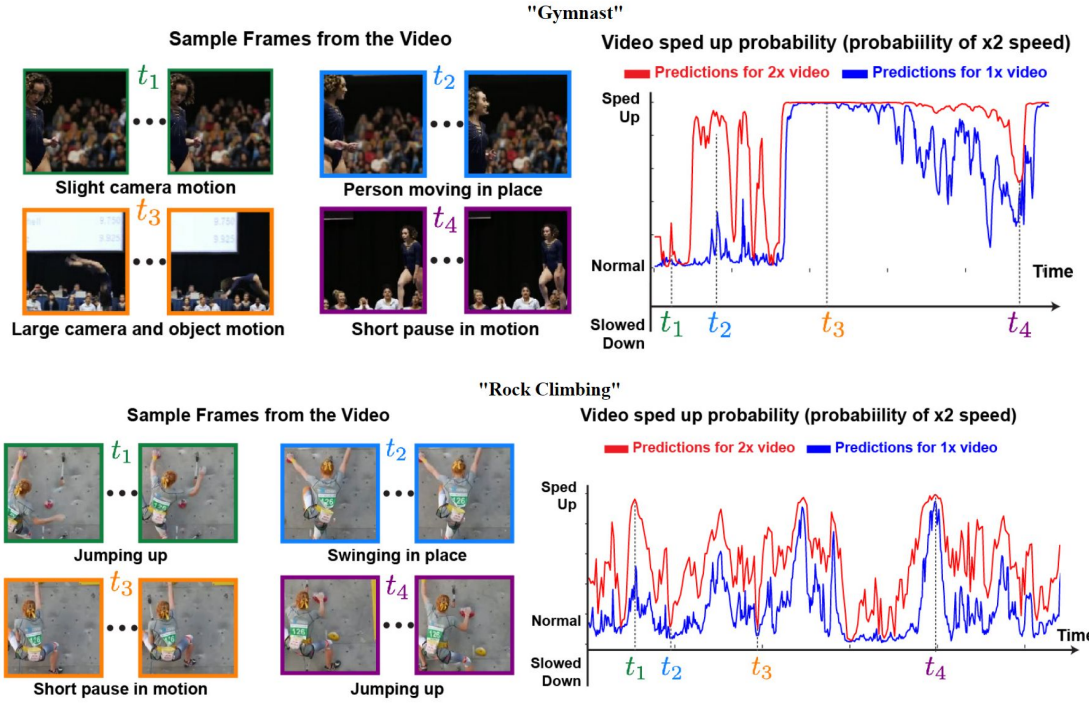




SpeedNet - motion magnitude



SpeedNet - motion magnitude



Adaptive video speedup

- **{V_i}** => Videos of different speeds
  - i -> 0 to k
- Speed of V_i = $\mathbf{X^i}$
  - X = 1.25

Adaptive video speedup

- $\{V_i\}$
- $\{P_{Vi}\}$ => probability of normal speed
- $\{P_{Vi}^{\delta}\}$ => threshold probability to get speed label
  - 0 => spedup
  - 1 => normal speed

Adaptive video speedup

- $\{V_i\}$
- $\{P_{Vi}\}$
- **$\{V_i(t)\} = \{P_{Vi}^{\delta}\} * X$**
- **$V(t) = \max(\{V_i(t)\})$** along i
  - contains the maximum possible speedup for each timestep that was still classified as not sped-up

Adaptive video speedup

Optimal speedup **S***

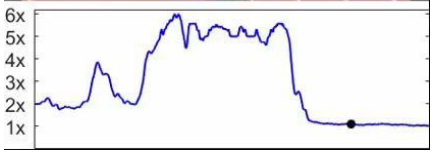
Match overall speedup

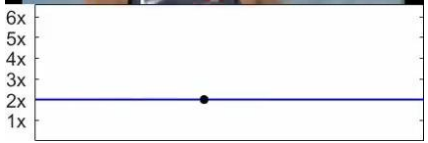
$$\arg\min_S E_{\text{speed}}(S, V) + \beta E_{\text{rate}}(S, R_o) + \alpha E_{\text{smooth}}(S')$$

Match speedup given by V

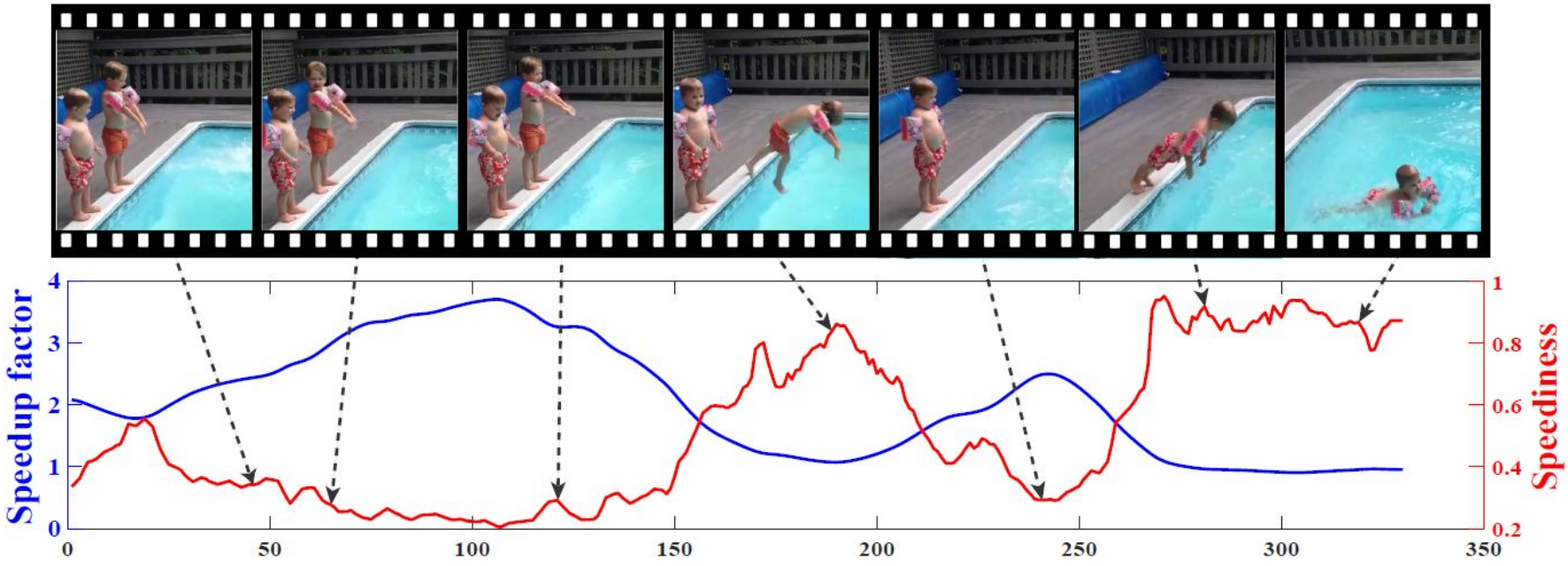
smoothness regularizer

Adaptive video speedup





Adaptive video speedup



Experiments on Adaptive video speedup

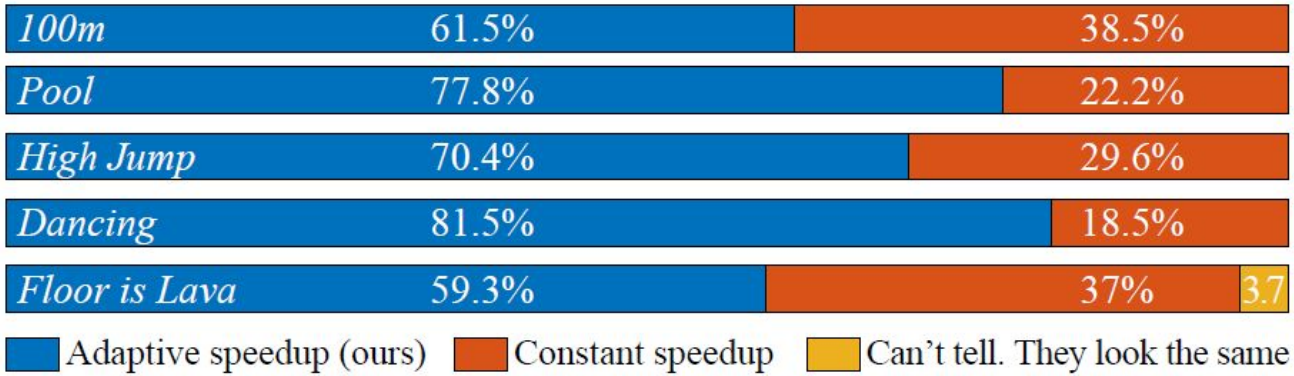
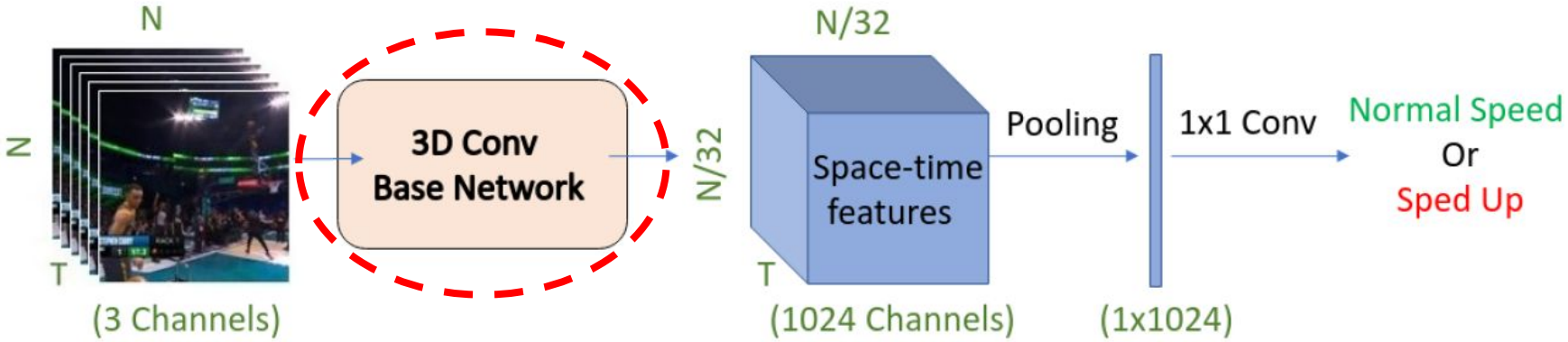


Figure 6. **Adaptive video speedup user study.** We asked 30 participants to compare our adaptive speedup results with constant uniform speedup for 5 videos (without saying which is which), and select the one they liked better. Our adaptive speedup results were consistently (and clearly) preferred over uniform speedup.

Action recognition



Pretraining

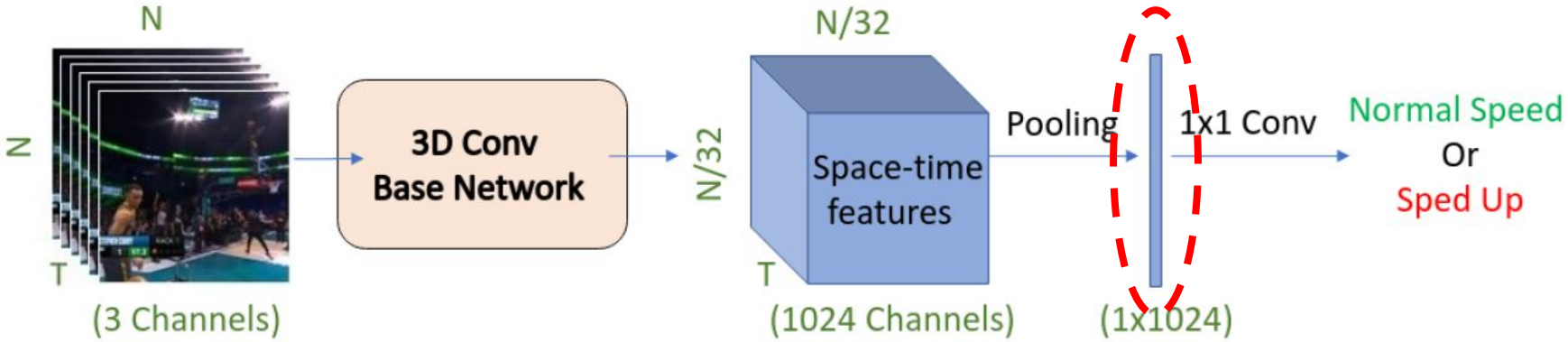
Action recognition

Initialization		Supervised accuracy	
Method	**Architecture**	**UCF101**	**HMDB51**
Random init	S3D-G	73.8	46.4
ImageNet inflated	S3D-G	86.6	57.7
Kinetics supervised	S3D-G	96.8	74.5
CubicPuzzle [19]	3D-ResNet18	65.8	33.7
Order [40]	R(2+1)D	72.4	30.9
DPC [13]	3D-ResNet34	75.7	35.7
AoT [38]	T-CAM	79.4	-
SpeedNet (Ours)	S3D-G	**81.1**	**48.8**
Random init	I3D	47.9	29.6
SpeedNet (Ours)	I3D	66.7	43.7

Action recognition

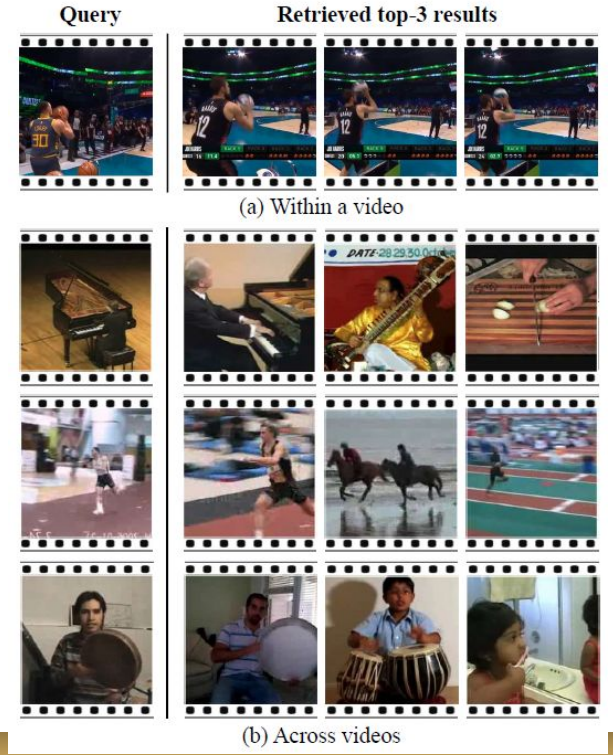
Initialization		Supervised accuracy	
Method	**Architecture**	**UCF101**	**HMDB51**
Random init	S3D-G	73.8	46.4
ImageNet inflated	S3D-G	86.6	57.7
Kinetics supervised	S3D-G	96.8	74.5
CubicPuzzle [19]	3D-ResNet18	65.8	33.7
Order [40]	R(2+1)D	72.4	30.9
DPC [13]	3D-ResNet34	75.7	35.7
AoT [38]	T-CAM	79.4	-
SpeedNet (Ours)	S3D-G	**81.1**	**48.8**
Random init	I3D	47.9	29.6
SpeedNet (Ours)	I3D	66.7	43.7

Nearest neighbor retrieval



Video representation

Nearest neighbor retrieval



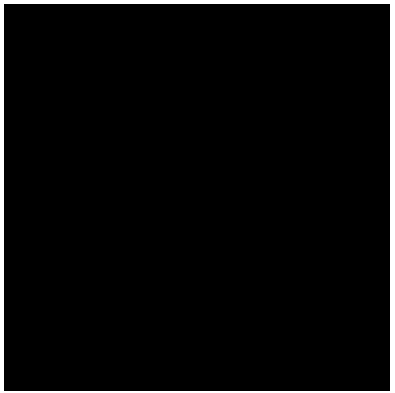
Not necessarily same class

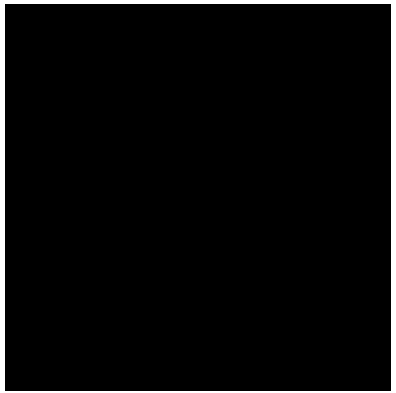
Nearest neighbor class correlation

Method	Architecture	1	5	10	20	50
Jigsaw [28]	CFN	19.7	28.5	33.5	40.0	49.4
OPN [22]	OPN	19.9	28.7	34.0	40.6	51.6
Buchler [3]	CaffeNet	25.7	36.2	42.2	49.2	59.5
Order [40]	C3D	12.5	29.0	39.0	50.6	66.9
Order [40]	R(2+1)D	10.7	25.9	35.4	47.3	63.9
Order [40]	R3D	14.1	30.3	40.0	51.1	66.5
Ours	S3D-G	13.0	28.1	37.5	49.5	65.0

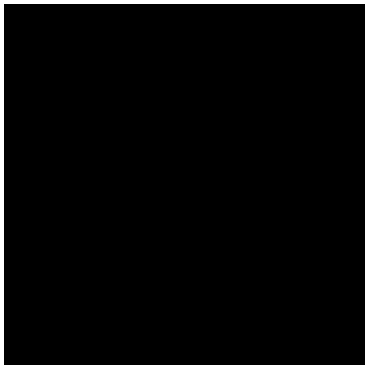
Table 3. **Recall-at-topK.** Top-K accuracy for different values of K for UCF101.

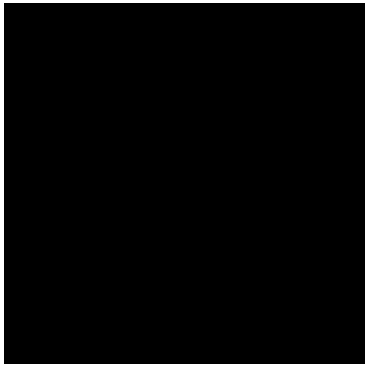
Visualizing salient spacetime regions





Spatially-varying speediness





Strengths

- Novel idea to detect video speed
- Theory and details behind the method well explained in the paper
- Great experimentation and ablation
- Learns general representations useful for multiple applications
- Great project page with details and examples - https://speednet-cvpr20.github.io/

Weaknesses

- Could have used backbones other than S3D to compare with the action recognition task baselines
- NFS dataset framerate adjusted so that it’s similar to that of Kinetics. Could have tried with original framerate
- Adaptive video speedup user study is human based and subjective

Discussion