CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Tutorial on Action Classification and Video Modeling
https://actionclassification-videomodelling.github.io/

Action Recognition in Long Videos

Lorenzo Torresani

facebook research

DARTMOUTH

facebook
Artificial Intelligence Research

SCSampler: Sampling Salient Clips from Video for Efficient Action Recognition


Bruno Korbar

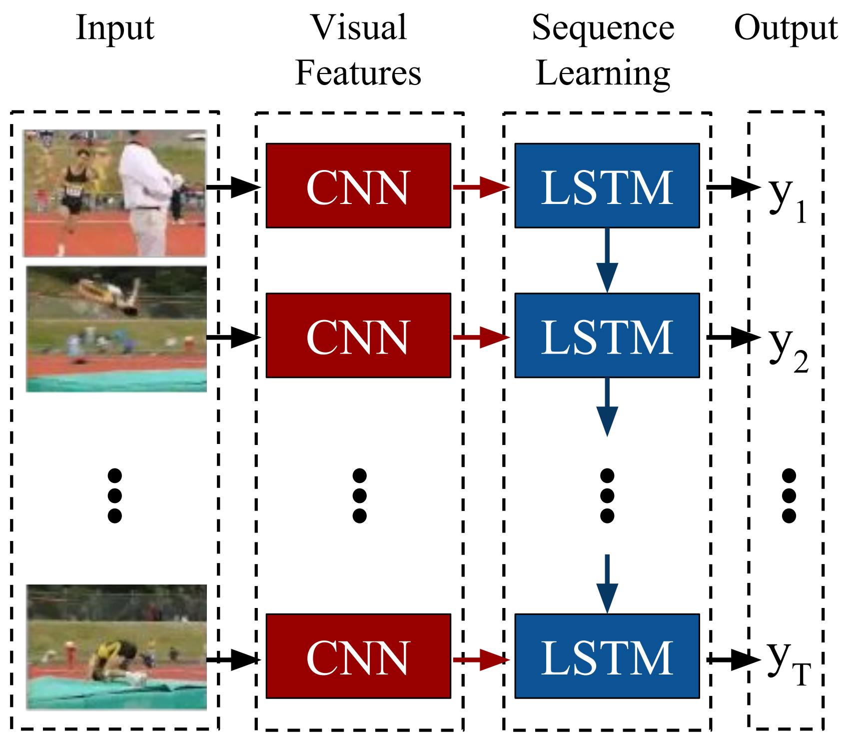

Du Tran


Lorenzo Torresani

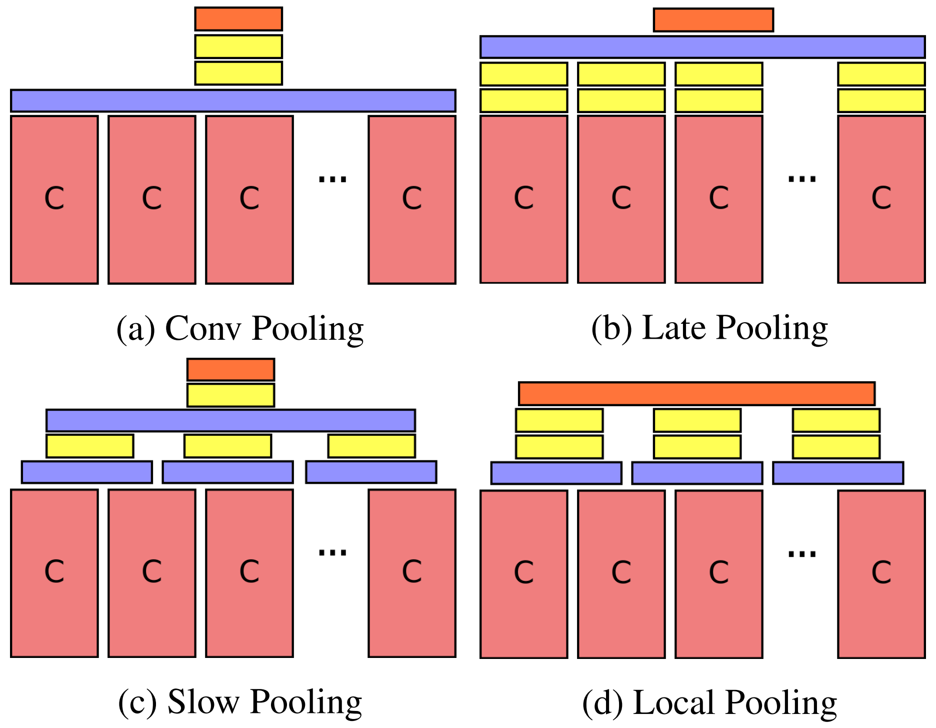
Preprint available at https://arxiv.org/abs/1904.04289

From clip-level to video-level prediction

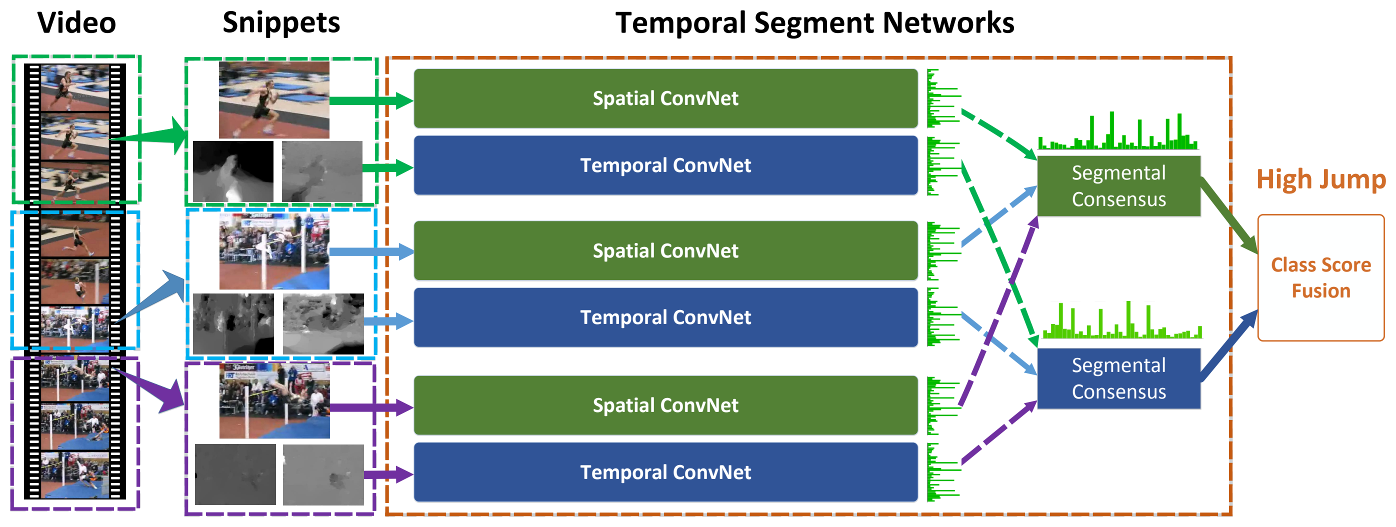
- Many different scheme have been proposed:



From [Donahue et al., CVPR 2015]

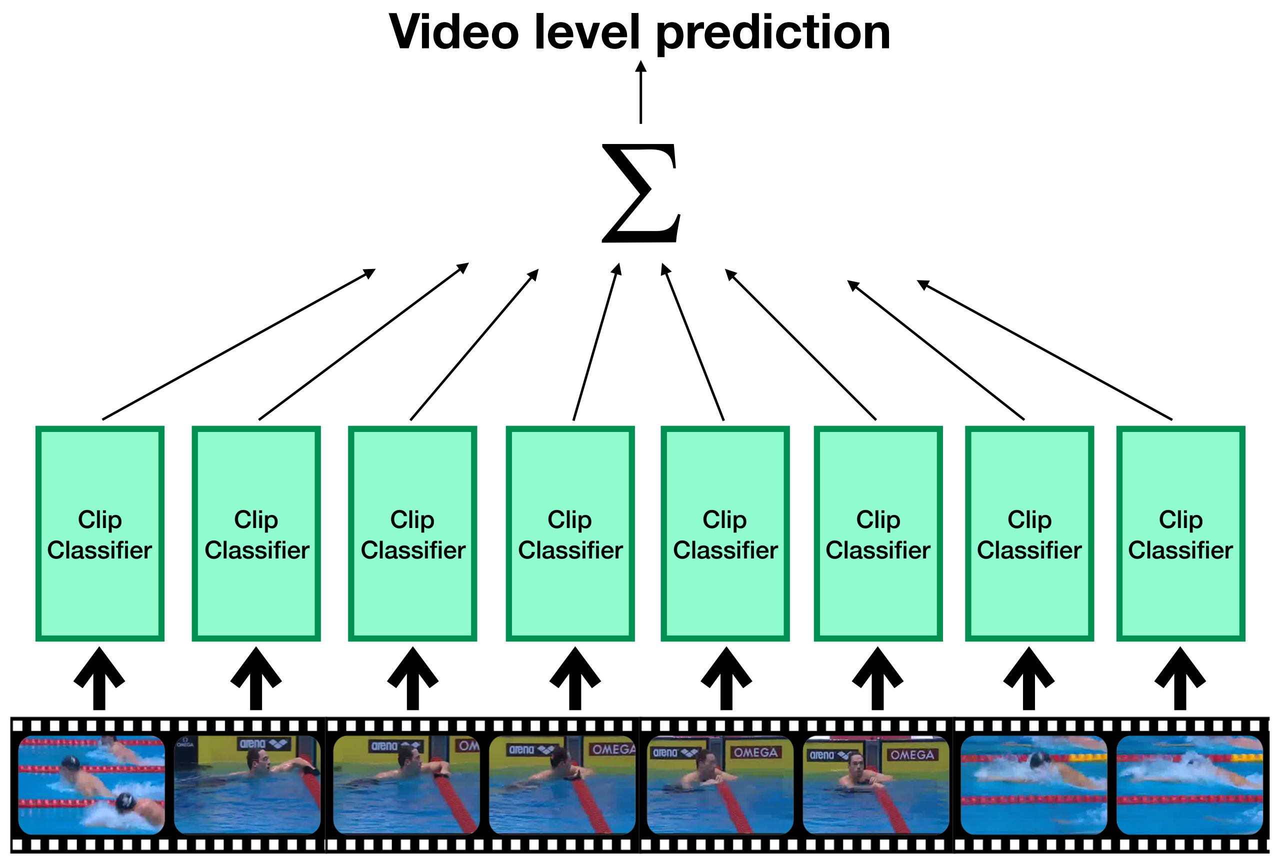


From [Ng et al., CVPR 2015]



From [Wang et al., CVPR 2016]

- Yet, most state-of-the-art action recognition models today simply average clip-level predictions (e.g., I3D, R(2+1)D, Non-Local nets)
  Benefits:
  🙂 simple
  🙂 effective in most scenarios



.. but averaging clip scores does not scale

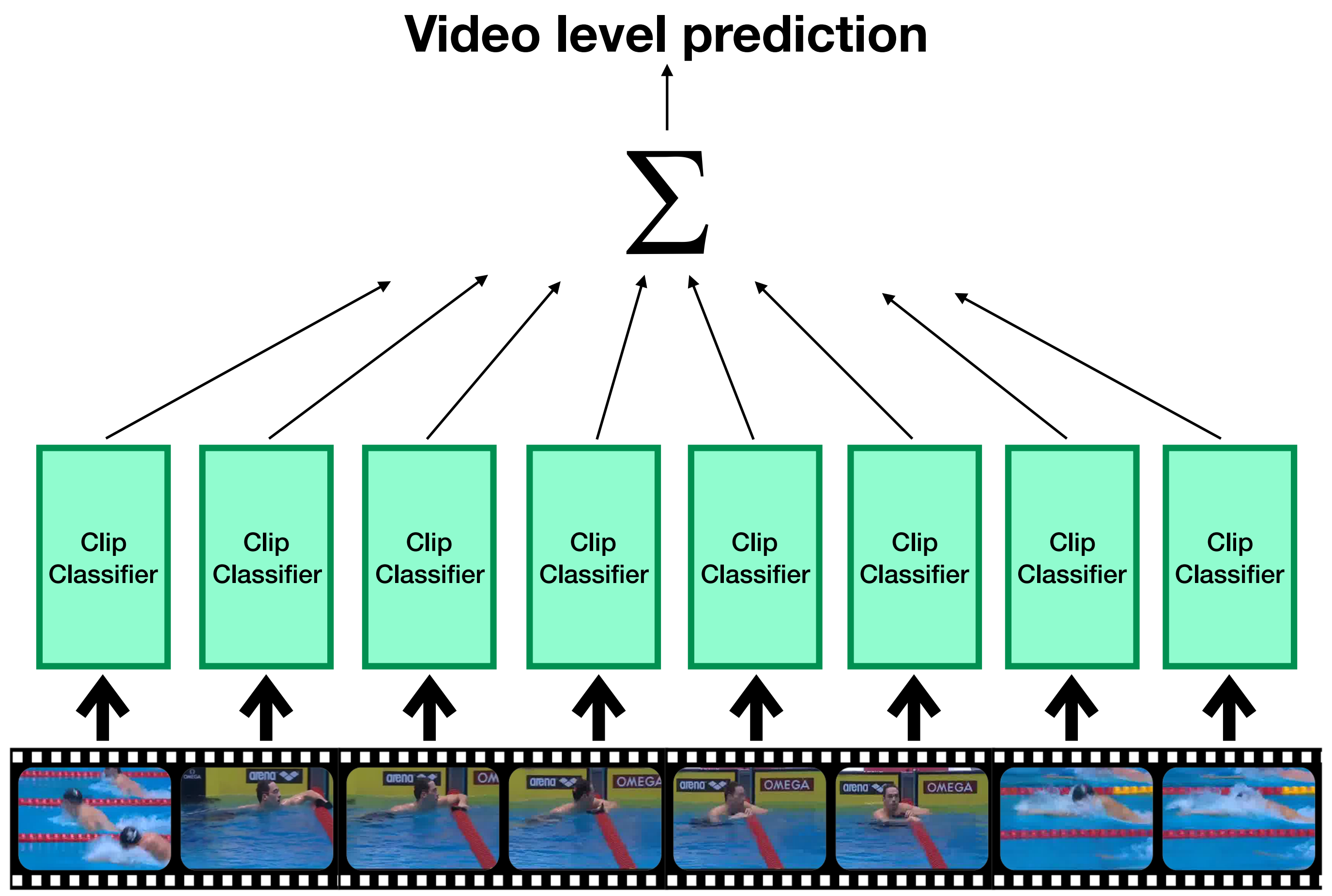
- Real-world videos are several minutes long but often contain few salient segments

[Karpathy et al., CVPR14]

Average video length: 3 minutes and 48 seconds

Average video length: 5 minutes and 36 seconds
Maximum video length: 1 hour and 34 minutes

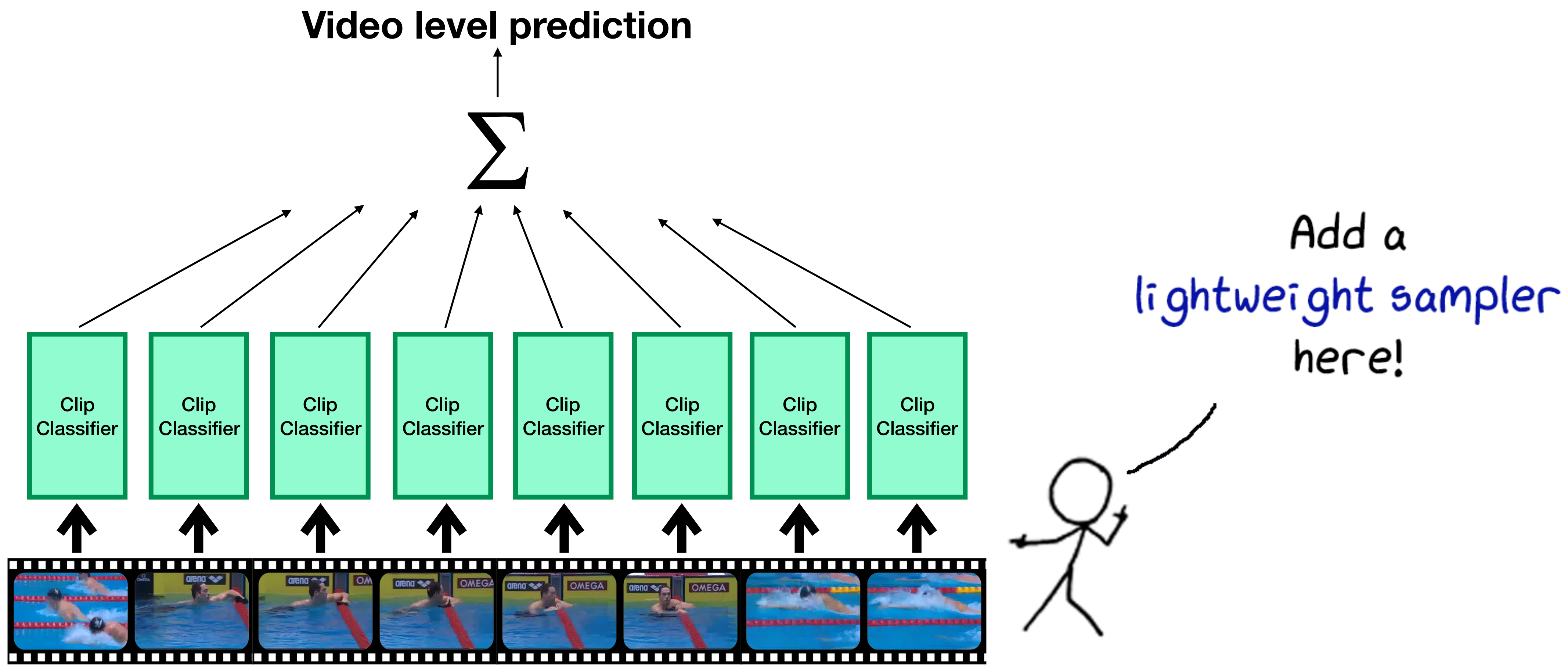
- Problems:
  🙁 computationally prohibitive for videos in the wild
  🙁 irrelevant clips outnumber salient segments



Approach [Korbar, Tran and Torresani, arXiv 2019]

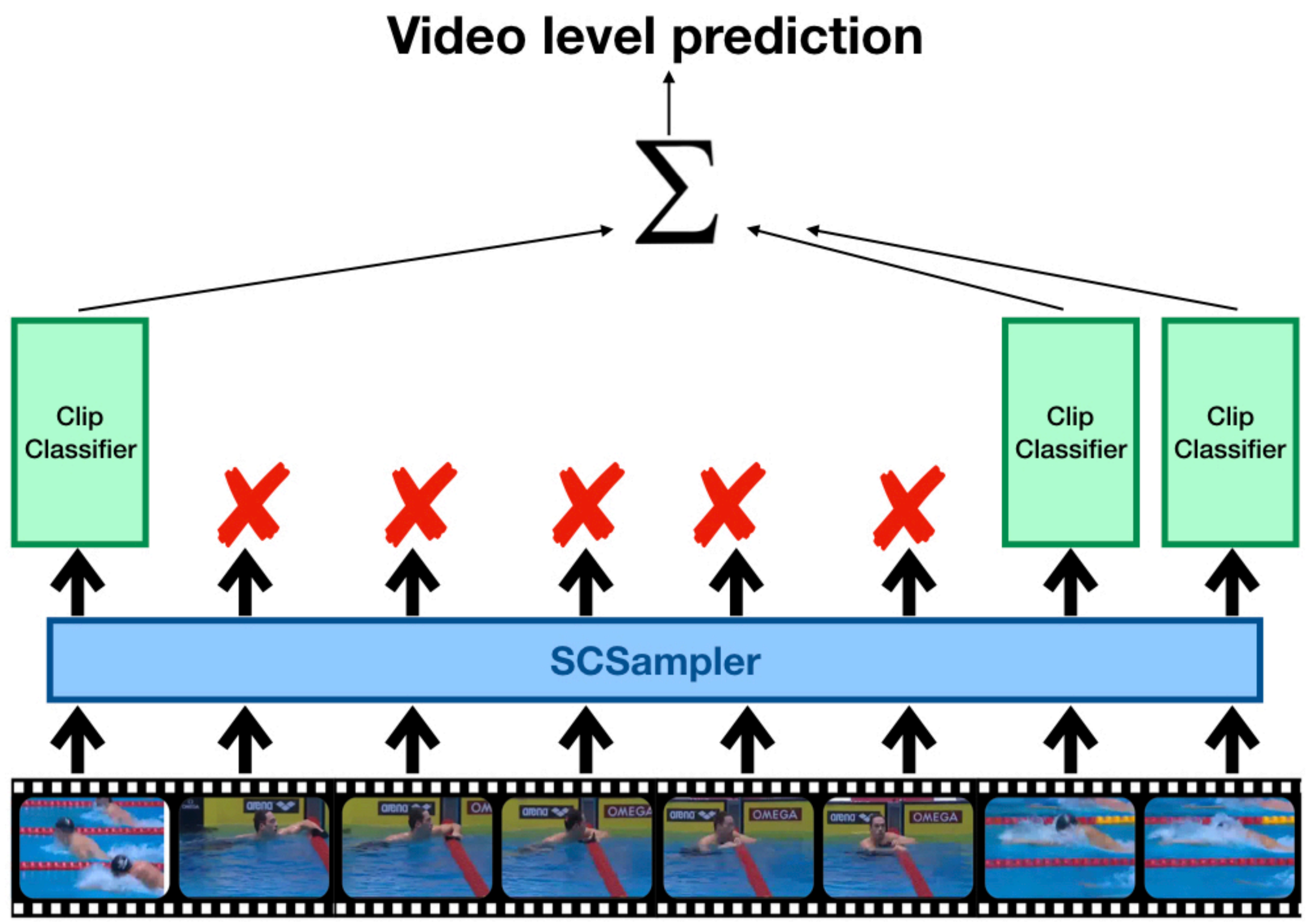
https://arxiv.org/abs/1904.04289

- Design a clip sampler to *efficiently* select the most salient clip of a video
- Run costly action classifier on this small subset of clips



Benefits:

- Improve both *efficiency* and *accuracy* of video-level classification by removing irrelevant clips from consideration

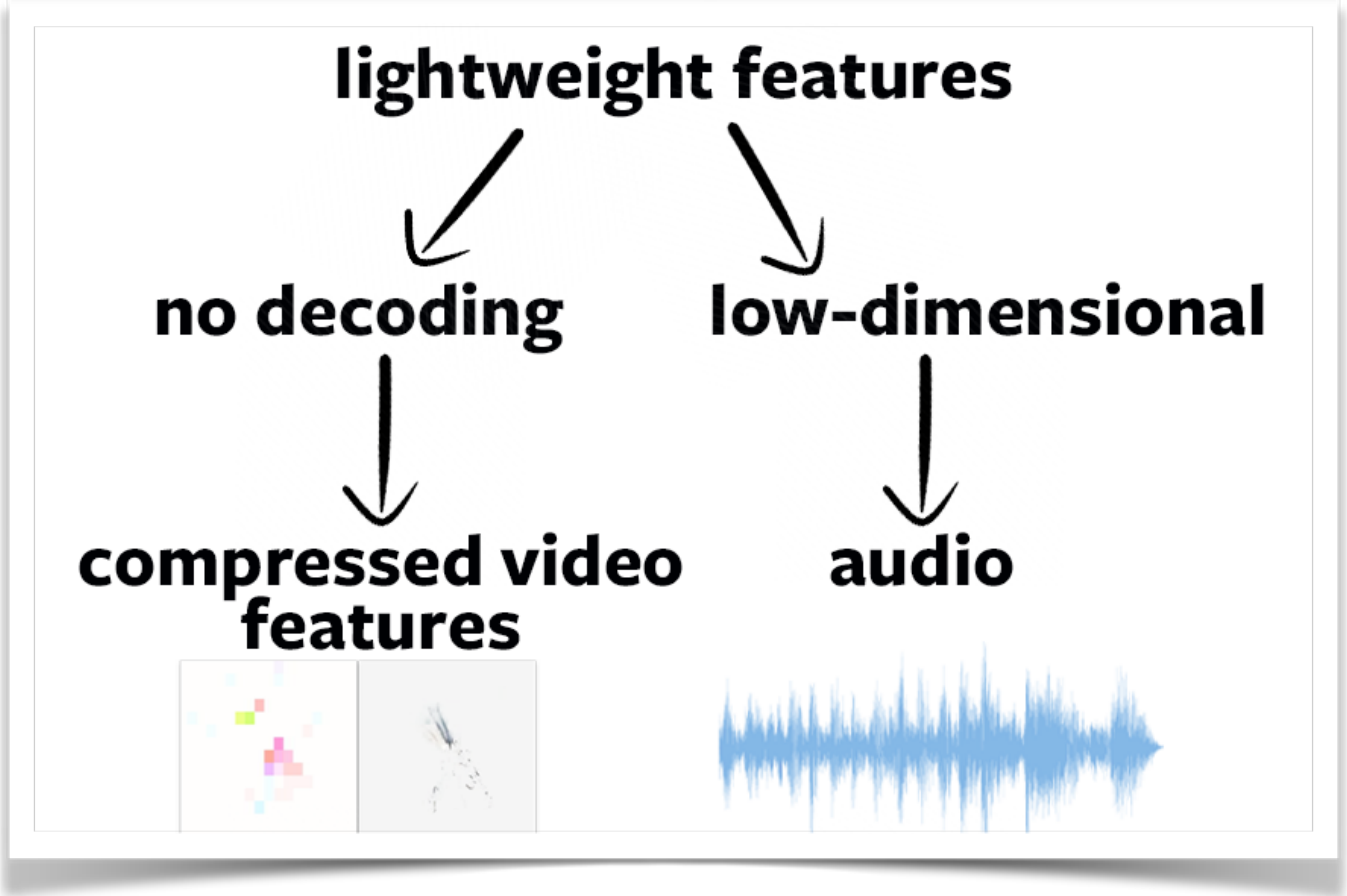


Salient Clip Sampler (SCSampler) Design

SCSampler requirements:

- Must have high precision
- Must be orders of magnitude faster than the action classifier

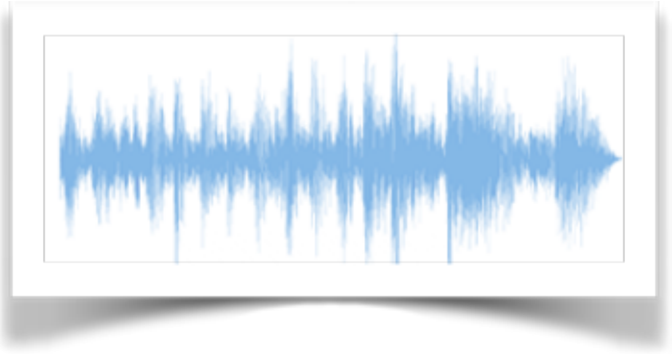
These requirements can be met by leveraging semantically-rich features that can be extracted without costly video decoding

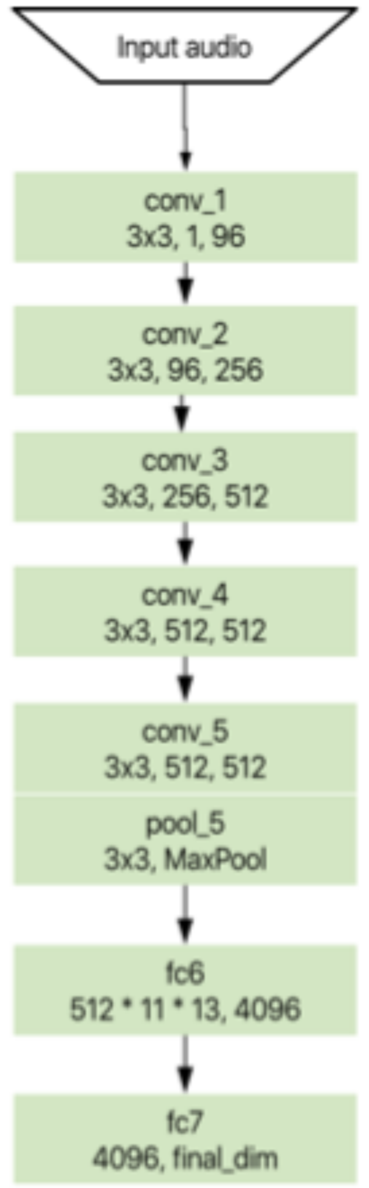


Audio SCSampler

- Audio channel is separately encoded from the video
- Audio had been shown to be semantically correlated to the content in the video [Aytar et al., NIPS16; Arandjelovic and Zisserman, ICCV17; Owens and Efros, ECCV18; Gao et al., ECCV18]

$\phi_A^{(i)}$ audio clip of 2 seconds





$s(\phi_A^{(i)})$ *VGG* applied to audio MEL-spectrogram [Chung and Zisserman, ACCVW16]

SCSampler on Compressed Video

- MPEG-4/H.264 video encodings:

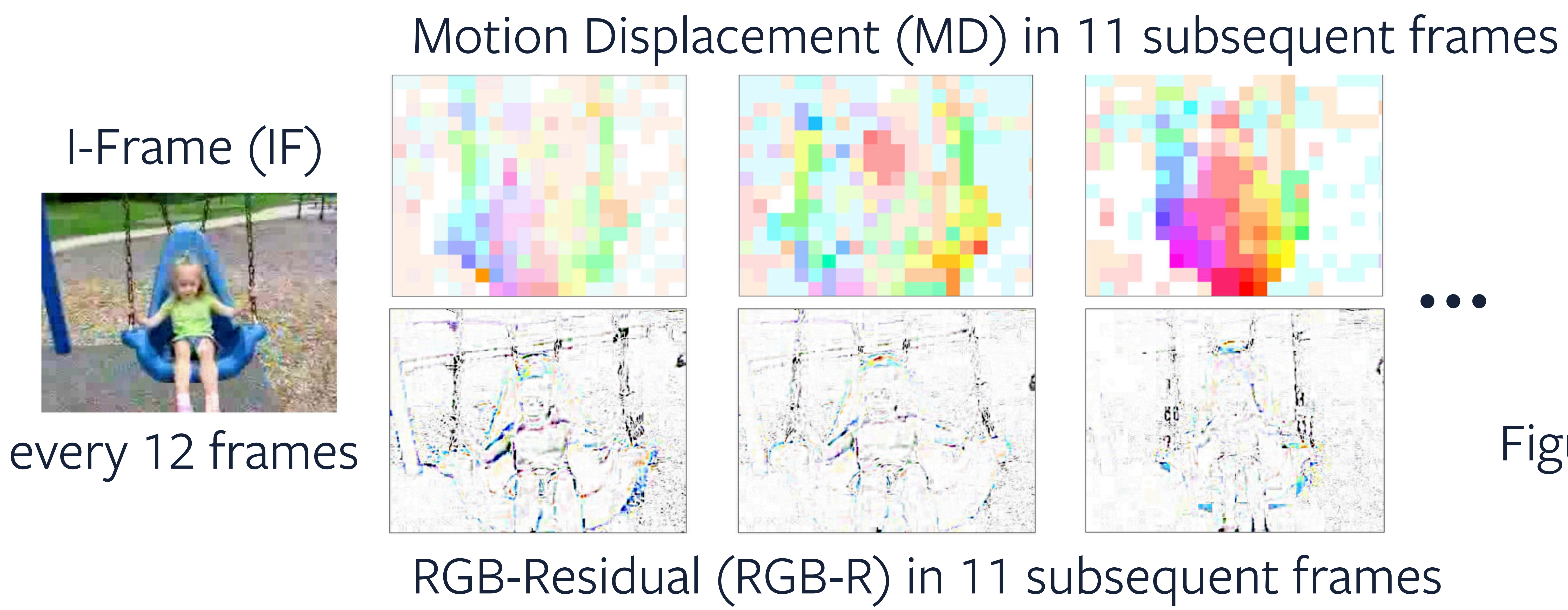


Figure from [Wu et al. CVPR18]

- CNNs trained on IF/RGB-R/MD for action recognition were shown to achieve good accuracy [Wu et al., CVPR18]
- We adopted a lightweight SCSampler CNN (ResNet-18) trained on each individual modality (IF/RGB-R/MD)

SCSampler learning objectives

Two variants:

- Action Classification (AC) loss

  # action classes

  1. Train SCSampler as an action classifier $s_{AC}(\phi^{(i)}) \in [0,1]^C$ using cross-entropy loss
  2. At test time, compute SCSampler saliency score as $s(\phi^{(i)}) = \max_{c \in \{1,\ldots,C\}} s_c^{AC}(\phi^{(i)})$

  max over action classes

- Ranking (RANK) loss

  ✓ Train SCSampler to rank higher clips that are *better* classified by action classifier $\mathbf{f}(v^{(i)})$

desired ranking wrt ground-truth action class $c*$

$$z^{(i,j)} = \begin{cases} 1 & \text{if} \quad f_{c^*}(v^{(i)}) > f_{c*}(v^{(j)}) \\ -1 & \text{otherwise} \end{cases}$$

ranking loss

$$\ell(\phi^{(i)}, \phi^{(j)}, z^{(i,j)}) = \max\left(0, -z^{(i,j)}[s(\phi^{(i)}) - s(\phi^{(j)}) + \eta]\right)$$

Experimental evaluation

- Assessing design choices on miniSports (136K/133K training/testing subset of Sports1M)
- Clip-level action classifier $\mathbf{f}(v^{(i)})$ is MC3-18, a 3D CNN from [Tran et al, CVPR18]

Clip sampling method	accuracy (%)	runtime (min)
Random	59.51	15.1
Uniform	59.87	15.1
Dense	61.6	2293.5 (38.5 hrs)
Audio SCSampler	67.82	22.0
Visual SCSampler	73.05	20.9
Audio-Visual SCS - Joint Training	75.53	23.4

Video-level recognition accuracy on miniSports with MC-13 by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

Experimental evaluation

- Assessing design choices on miniSports (136K/133K training/testing subset of Sports1M)
- Clip-level action classifier $\mathbf{f}(v^{(i)})$ is MC3-18, a 3D CNN from [Tran et al, CVPR18]

Clip sampling method	accuracy (%)	runtime (min)
Random	59.51	15.1
Uniform	59.87	15.1
Dense	61.6	2293.5 (38.5 hrs)
Audio SCSampler	67.82	22.0
Visual SCSampler	73.05	20.9
Audio-Visual SCS - Joint Training	75.53	23.4

Video-level recognition accuracy on miniSports with MC-13 by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

Audio SCSampler yields gain of 8% over Uniform

Experimental evaluation

- Assessing design choices on miniSports (136K/133K training/testing subset of Sports1M)
- Clip-level action classifier $\mathbf{f}(v^{(i)})$ is MC3-18, a 3D CNN from [Tran et al, CVPR18]

Clip sampling method	accuracy (%)	runtime (min)
Random	59.51	15.1
Uniform	59.87	15.1
Dense	61.6	2293.5 (38.5 hrs)
Audio SCSampler	67.82	22.0
Visual SCSampler	73.05	20.9
Audio-Visual SCS - Joint Training	75.53	23.4

Video-level recognition accuracy on miniSports with MC-13 by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

Audio SCSampler yields gain of more than 6% over dense prediction

Experimental evaluation

- Assessing design choices on miniSports (136K/133K training/testing subset of Sports1M)
- Clip-level action classifier $\mathbf{f}(v^{(i)})$ is MC3-18, a 3D CNN from [Tran et al, CVPR18]

Clip sampling method	accuracy (%)	runtime (min)
Random	59.51	15.1
Uniform	59.87	15.1
Dense	61.6	2293.5 (38.5 hrs)
Audio SCSampler	67.82	22.0
Visual SCSampler	73.05	20.9
Audio-Visual SCS - Joint Training	75.53	23.4

Video-level recognition accuracy on miniSports with MC-13 by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

Visual SCSampler gives accuracy boost of 11.4% over dense prediction

Experimental evaluation

- Assessing design choices on miniSports (136K/133K training/testing subset of Sports1M)
- Clip-level action classifier $\mathbf{f}(v^{(i)})$ is MC3-18, a 3D CNN from [Tran et al, CVPR18]

Clip sampling method	accuracy (%)	runtime (min)
Random	59.51	15.1
Uniform	59.87	15.1
Dense	61.6	2293.5 (38.5 hrs)
Audio SCSampler	67.82	22.0
Visual SCSampler	73.05	20.9
Audio-Visual SCS - Joint Training	75.53	23.4

combining audio and visual information elevates the gain over dense prediction to 15%!

Video-level recognition accuracy on miniSports with MC-13 by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

Experimental evaluation

- Assessing design choices on miniSports (136K/133K training/testing subset of Sports1M)
- Clip-level action classifier $\mathbf{f}(v^{(i)})$ is MC3-18, a 3D CNN from [Tran et al, CVPR18]

Clip sampling method	accuracy (%)	runtime (min)
Random	59.51	15.1
Uniform	59.87	15.1
Dense	61.6	2293.5 (38.5 hrs)
Audio SCSampler	67.82	22.0
Visual SCSampler	73.05	20.9
Audio-Visual SCS - Joint Training	75.53	23.4

combining audio and visual information elevates the gain over dense prediction to 15%!

98x faster than dense evaluation

Video-level recognition accuracy on miniSports with MC-13 by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

Large-scale experiment

- Evaluation on full Sports1M using state-of-the-art action classification models (currently, CSN-152 [2] has the best reported classification accuracy on Sports1M)

	SCSampler $\mathcal{S}$ (K clips)		Uniform (K clips)		Dense (*all* clips)	
	acc. (%)	runtime (day)	acc. (%)	runtime (day)	acc. (%)	runtime (days)
R(2+1)D-34 [1]	77.96	0.9	71.49	0.6	70.90	14.2
CSN-152 [2]	83.98	0.9	75.80	0.5	76.97	14.0

Video-level recognition accuracy on Sports1M by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

[1] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In CVPR 2018.
[2] D. Tran, H. Wang, L. Torresani, and M. Feiszli. Classification with channel-separated convolutional networks. arXiv preprint, 2019

Large-scale experiment

- Evaluation on full Sports1M using state-of-the-art action classification models (currently, CSN-152 [2] has the best reported classification accuracy on Sports1M)

	SCSampler $\mathcal{S}$ (K clips)		Uniform (K clips)		Dense (*all* clips)	
	acc. (%)	runtime (day)	acc. (%)	runtime (day)	acc. (%)	runtime (days)
R(2+1)D-34 [1]	77.96	0.9	71.49	0.6	70.90	14.2
CSN-152 [2]	83.98	0.9	75.80	0.5	76.97	14.0

Video-level recognition accuracy on Sports1M by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

SCSampler elevates the best reported accuracy on Sports1M by 7%
while reducing by 15x the computational cost!

[1] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In CVPR 2018.
[2] D. Tran, H. Wang, L. Torresani, and M. Feiszli. Classification with channel-separated convolutional networks. arXiv preprint, 2019

Experiment on Kinetics

- Evaluation on Kinetics-400 using state-of-the-art action classification models

	SCSampler $\mathcal{S}$ (K clips)		Uniform (K clips)		Dense (*all* clips)	
	acc. (%)	runtime (hr)	acc. (%)	runtime (hr)	acc. (%)	runtime (hours)
R(2+1D)-34 [1]	76.71	1.6	73.26	1.5	74.11	3.1
I3D-RGB [3]	75.12	1.5	71.18	1.3	72.75	2.9
CSN-152 [2]	80.23	1.6	77.53	1.5	78.81	3.0

Video-level recognition accuracy on Sports1M by averaging predictions over K=10 clips per video (except for Dense, which uses all clips)

even though Kinetics videos are short (~10 seconds)
SCSampler provides a boost in accuracy
for all models, albeit small (1.4-2.4%)

[1] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In CVPR 2018.
[2] D. Tran, H. Wang, L. Torresani, and M. Feiszli. Classification with channel-separated convolutional networks. arXiv preprint, 2019.
[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In CVPR 2017.

Top-ranked and bottom-ranked clips

Top-3 clips sampled by SCSampler | Bottom-ranked clips by SCSampler