3D CNNs with Adaptive Temporal Feature Resolutions

Mohsen Fayyaz[1,*], Emad Bahrami[1,*], Ali Diba[2], Mehdi Noroozi[3], Ehsan Adeli[4], Luc Van Gool[2,5], Jürgen Gall[1]

[1]University of Bonn, [2]KU Leuven, [3]Bosch Center for Artificial Intelligence, [4] Stanford University, [5] ETH Zurich

[*]Contributed equally to this work

https://similarityguidedsampling.github.io

CVPR VIRTUAL JUNE 19-25

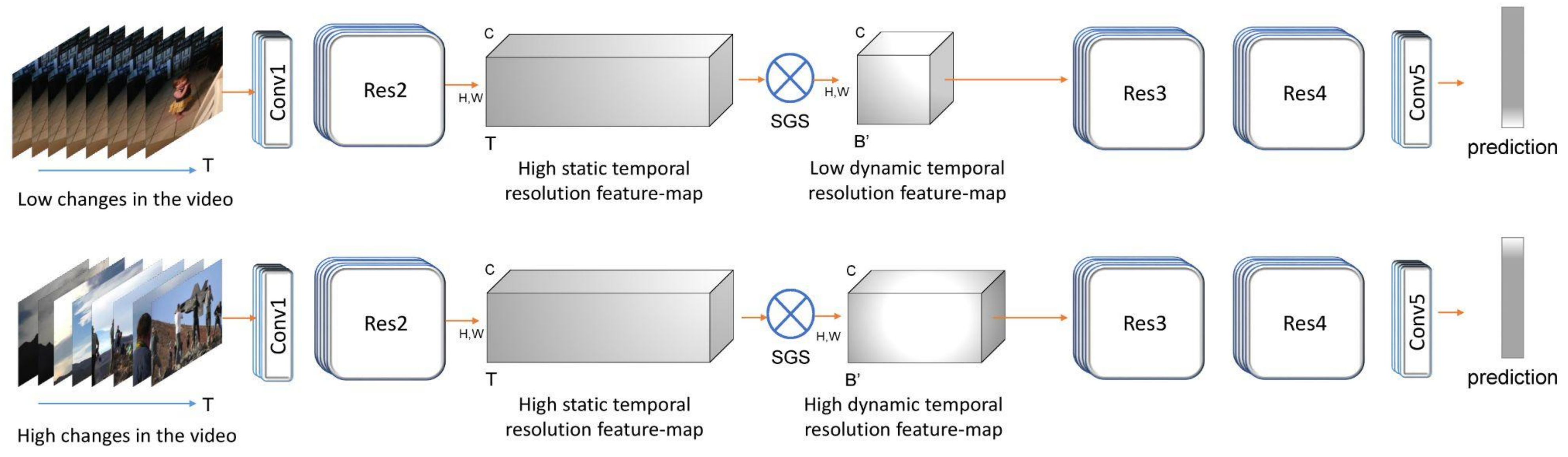
UNIVERSITÄT BONN

1.Handling Redundancy In 3D CNNs

Problem:

- Static temporal feature resolution leads to unnecessary computational overhead
- Various amount of changes across videos needs an adaptive method to handle the redundancy of temporal features

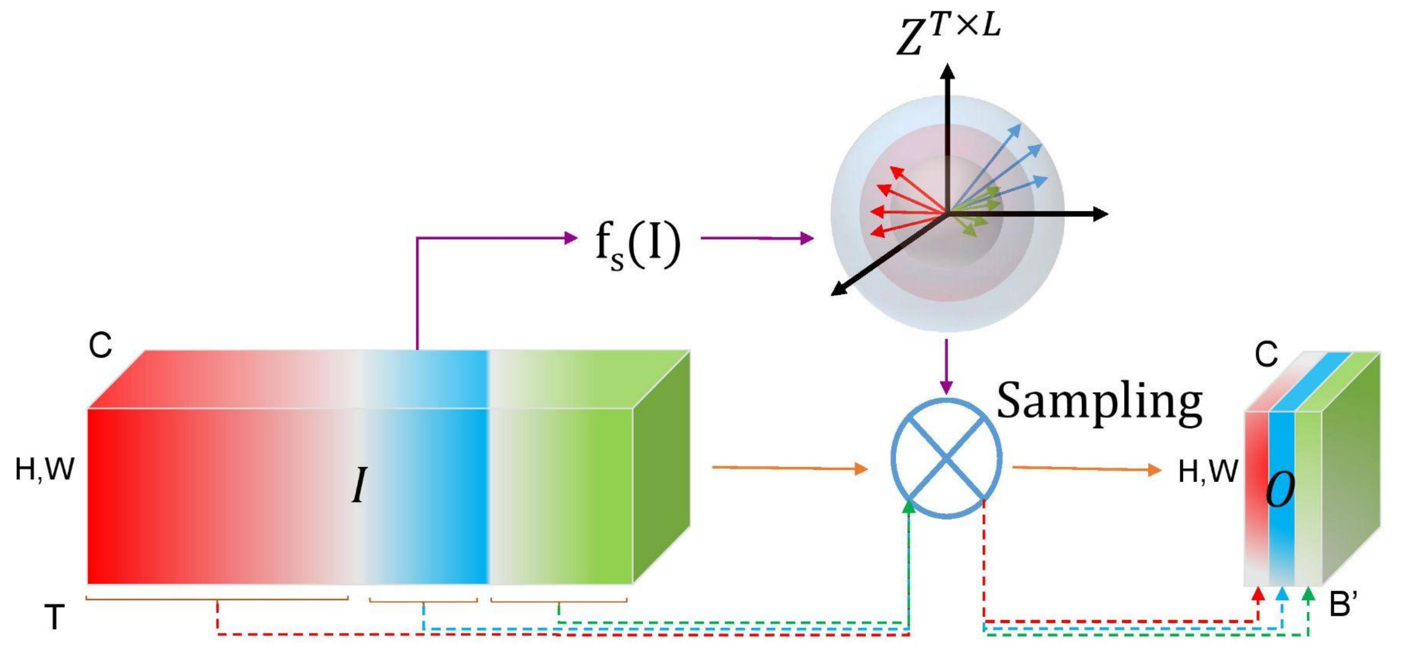
- **Similarity Guided Sampling:** discarding redundant information by grouping and aggregating temporally similar feature maps



- Complementary to existing 3D CNNs
- Reduction of GFLOPs in state-of-the art 3D CNNs by half while preserving the accuracy
- Adaptive temporal resolution to deal with variations in a dataset
- Evaluated on challenging datasets such as Kinetics and Something-Something V2

2.Similarity Guided Sampling

- Mapping temporal feature maps to learnt similarity space
- Sampling based on the similarity of feature maps
- Grouping similar maps and aggregating them



- Differentiable bin sampling for sampling temporal feature maps

$$\mathcal{O}_b = \frac{1}{\sum_{t=1}^{T}\delta\left(\left\lfloor\frac{|\Delta_t-\beta_b|}{\gamma}\right\rfloor\right)}\sum_{t=1}^{T}\mathcal{I}_t\,\delta\left(\left\lfloor\frac{|\Delta_t-\beta_b|}{\gamma}\right\rfloor\right)$$

$\Delta_t = ||\mathcal{Z}_t||$

Num. bins B=T

$\Delta_{max} = \max(\Delta_1,\ldots,\Delta_T)$

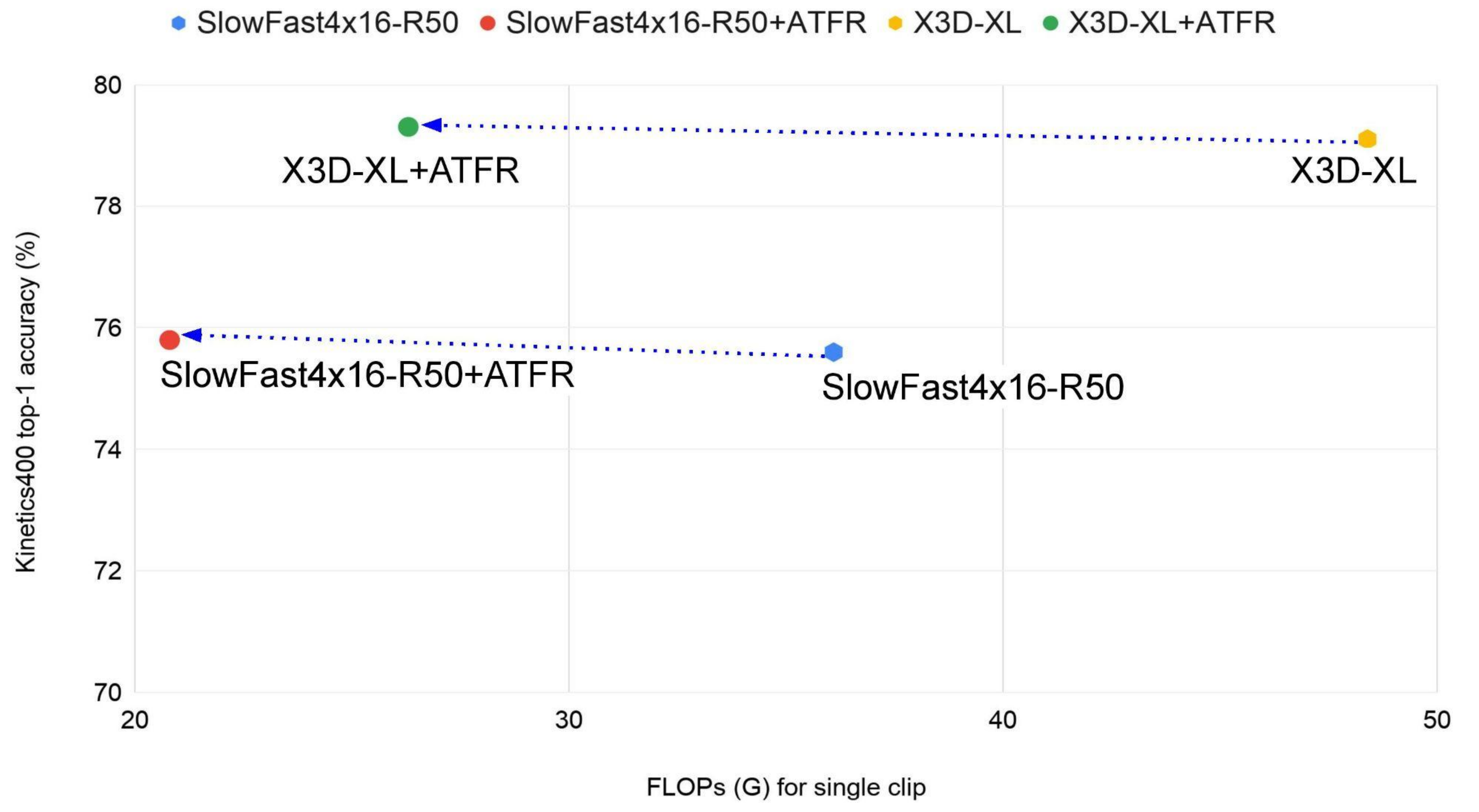
Half width of bins: $\gamma = \frac{\Delta_{max}}{2B}$ Center of bins: $\beta_b = (2b-1)\gamma \quad \forall b \in (1,\ldots,B)$

3.Temporal Resolution of Kinetics Action Classes

- Required adaptive temporal resolution for 3DResNet-50 + ATFR on Kinetics

Lowest Temporal Resolution	Highest Temporal Resolution
presenting weather forecast	passing American football (in game)
stretching leg	swimming breast stroke
playing didgeridoo	playing ice hockey
playing clarinet	pushing cart
golf putting	gymnastics tumbling

4.GFLOPs Improvement



5.Comparison with State-of-the-art

- Results on Kinetics-400

model	GFLOPs	top1	top5	Param
SlowFast4×16,R50	36.1×30	75.6	92.1	34.40M
SlowFast4×16,R50+ATFR	20.8 × 30 (↓ 42%)	75.8	92.4	34.40M
X3D-S$^\alpha$	1.9×10	72.9	90.5	**3.79M**
X3D-S+ATFR$^\alpha$	**1.0** × 10 (↓ 47%)	73.5	91.2	**3.79M**
X3D-XL$^\alpha$	35.8×10	78.4	93.6	11.09M
X3D-XL+ATFR$^\alpha$	20 × 10 (↓ 44%)	78.6	93.9	11.09M
X3D-XL$^\beta$	48.4×30	79.1	93.9	11.09M
X3D-XL+ATFR$^\beta$	26.3 × 30 (↓ 45%)	79.3	**94.1**	11.09M

- Accuracy and GFLOPs for Something-Something-V2

model	pretrain	GFLOPs	top1	top5
SlowFast-R50	Kinetics400	132.8	61.7	87.8
SlowFast-R50+ATFR	Kinetics400	**87.8** (↓ 33%)	**61.8**	**87.9**