Scalable Frame Sampling for Video Classification: A Semi-Optimal Policy Approach with Reduced Search Space

SEOUL NATIONAL UNIVERSITY
Visual Information Processing Lab

Junho Lee[1], Jeongwoo Shin[1], Seung Woo Ko[1,2], Seongsu Ha[1,3], Joonseok Lee[1,4]

[1] Seoul National University [2] LG AI Research [3] Twelve Labs [4] Google Research

BMVC 2024

Introduction

Problem statement

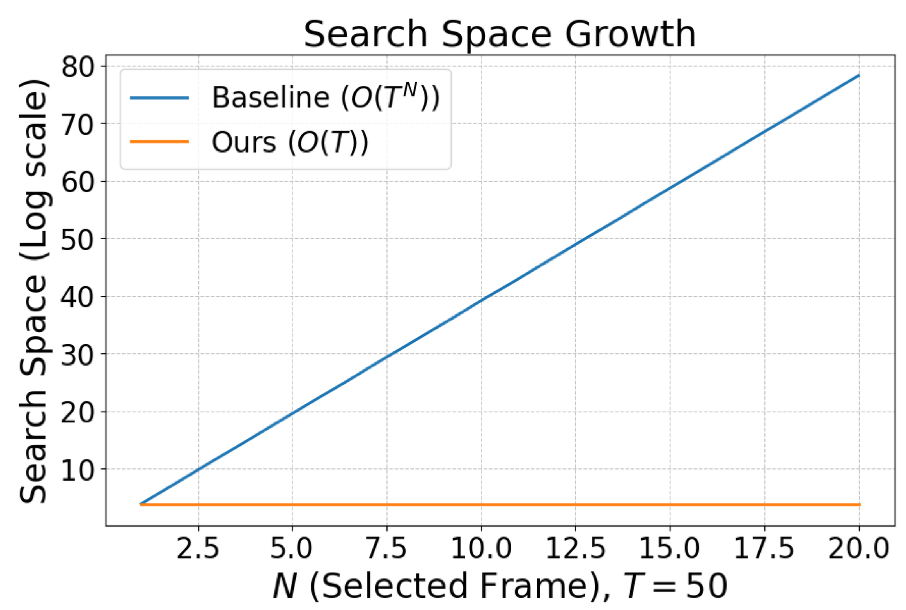
- **Selecting N frames from T candidates**, to avoid redundant computation and to enhance video understanding capability.
- However, **exploring all possible combinations** of frames requires $O(T^N)$ operations, which is computationally infeasible for **large N and T**.

Limitations of Previous Research

- Limited to small-scale scenarios
  - Most studies have focused on small N and T settings ($N \le 6, T \le 10$).
  - Even for these limited cases, exploring the complete search space is still complex.
- Reinforcement Learning (RL) approaches
  - Previous works tried to overcome the search space challenge using **reinforcement learning (RL)**, treating the sampler as an agent and the classifier as the environment, optimizing frame selection through rewards.
  - Challenge: **RL approaches still operate within the same $O(T^N)$ space**, limiting their scalability for large-scale video datasets.

Our Solution: Semi-Optimal Policy (π_s)

- We propose a semi-optimal policy (π_s) that reduces the search space to O(T) by **evaluating frames independently**.
- This approach allows for scalable frame sampling even for large N and T values.



Our Contributions

- **New Sampling Policy:** We propose the **Semi-Optimal Policy (π_s)**, which reduces search space from $O(T^N)$ to $O(T)$ by independently evaluating frames.
- **New Sampler:** We propose **SOSampler**, which learns the semi-optimal policy (π_s) instead of the optimal policy.
- **Performance:** Our method achieves state-of-the-art performance across multiple datasets and backbone architectures.
- **Scalability:** Unlike previous methods, our approach demonstrate robust performance gains even with large N and T values.

Sampling Policy

Optimal Policy (π_o)

- Definition: Select N frames from T candidates that maximize the classifier's confidence on the correct label.
- Optimal set: The N frames selected by the optimal policy (π_o) is defined as optimal set.
- Finding the optimal set and using it as a label to train the model can be the easy way to learn π_o, but its $O(T^N)$ complexity makes it computationally infeasible.

Semi-Optimal Policy (π_s)

- **Frame Independence Assumption**: Assessing the importance score of each frame independently, the sampling problem is simplified to $O(T)$.
- Definition: Select N frames from T candidates with the highest importance scores, when evaluated independently.
- The importance score of each frame is defined as $c(v_t) = \max_{\{i=1,\dots,C\}} [f_c(v_t)]_i$

Experimental Validation of π_s

- We estimate the relevance between adjacent frames (Figure 1) and conclude that **independence can reliably be assumed up to 1 fps**, with 5 fps being borderline.
- To verify that π_s approximates π_o, we compare performance (Table 1) and selected frames (Table 2).

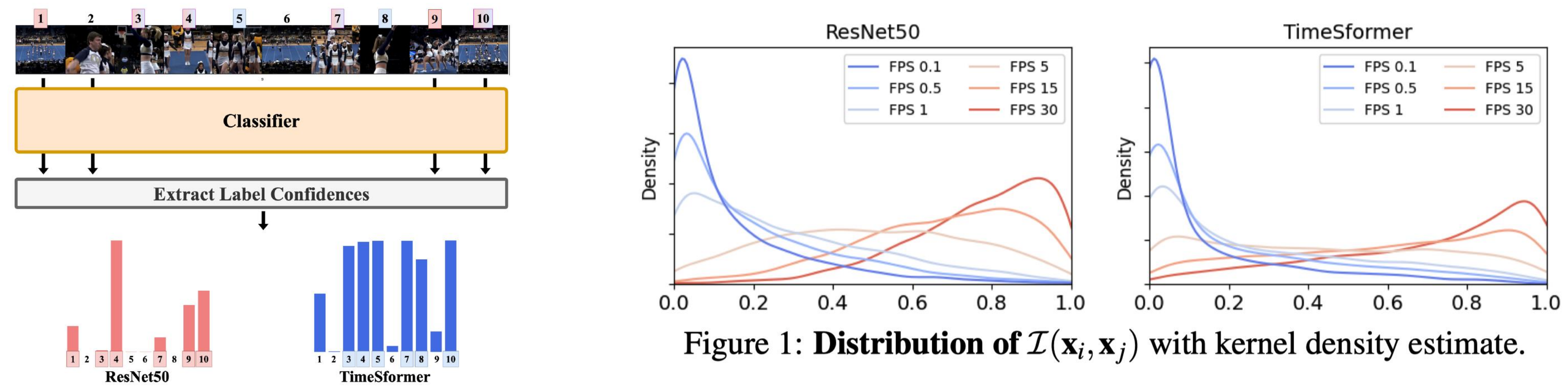


Figure 1: **Distribution of $\mathcal{I}(\mathbf{x}_i, \mathbf{x}_j)$** with kernel density estimate.

f_c	Policy	A-Net (mAP)		M-Kin. (Top-1)	
TimeSformer	π_u	87.0%	–	79.6%	–
	π_o	91.5%	+4.5	89.3%	+9.7
	π_s	89.4%	+2.4	84.8%	+5.2
	All	89.0%	+2.0	81.2%	+1.6
ResNet50	π_u	75.3%	–	72.5%	–
	π_o	90.5%	+15.2	83.8%	+11.3
	π_s	87.4%	+12.1	80.3%	+7.8
	All	77.8%	+2.5	73.6%	+1.1

Table 1: **Performance of π_o and π_s on ActivityNet and Mini-Kinetics**. Relative improvement from π_u is provided on the right.

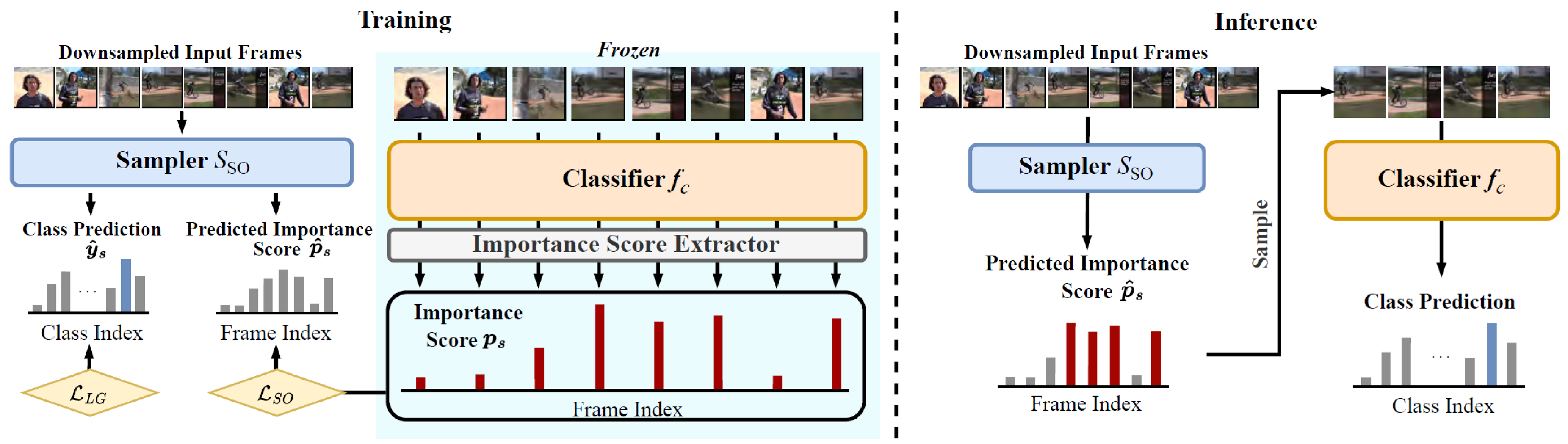
Dataset	Sampler	Sampling Fidelity (%) $N=1$	$N=2$	$N=3$	$N=4$	$N=5$	$N=6$
A-Net	Random	10.0	20.0	30.0	40.0	50.0	60.0
	FrameExit	10.2	19.3	29.3	39.3	49.6	59.3
	π_s	100.0	74.6	73.2	75.1	78.5	81.0
M-Kin.	Random	10.0	20.0	30.0	40.0	50.0	60.0
	FrameExit	9.8	22.5	32.0	42.5	51.5	62.5
	π_s	100.0	61.5	65.2	70.6	71.8	80.5

Table 2: **Sampling Fidelity**. Note that we report the expected value of the sampling fidelity for random sampling.

SOSampler

Training Strategy

- SOSampler: A lightweight sampler that learns the semi-optimal policy (π_s) instead of the optimal policy (π_o).
- L_{SO}: Penalizes when the estimated importance score differs from the pretrained classifier's score.
- L_{LG}: Penalizes when the predicted frame class differs from the true label, ensuring each frame reflects the video-level class.



Performance Comparison

- **Performance**: SOSampler outperforms other methods on various dataset for both small and large N, T settings.
- **Scalability**: As N and T increase, OCSampler[1] struggles to optimize effectively, often performing worse than uniform sampling in large-scale scenarios. SOSampler, in contrast, maintains high performance even for large N and T values.

Dataset	Backbone	Method	N / T 8 / 30	16 / 60	32 / 100
ActivityNet	ResNet50	Uniform	77.1%	79.4%	80.4%
		OCSampler	78.0%	79.1%	80.1%
		SOSampler	**78.7%**	**80.2%**	**81.1%**
	TimeSformer	Uniform	88.5%	89.9%	90.3%
		OCSampler	85.0%	85.3%	84.5%
		SOSampler	**89.5%**	**90.1%**	**90.5%**
Mini-Sports1M	ResNet50	Uniform	46.9%	48.8%	49.1%
		OCSampler	48.6%	49.6%	50.0%
		SOSampler	**50.0%**	**51.1%**	**51.5%**
	TimeSformer	Uniform	53.9%	55.6%	56.8%
		OCSampler	48.9%	49.6%	50.0%
		SOSampler	**55.1%**	**56.9%**	**57.8%**

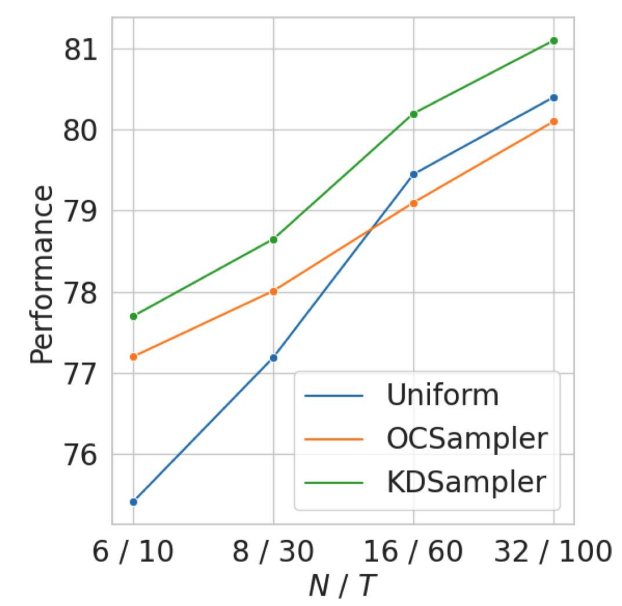
Table 5: **Experiment on long videos for large N and T**. The best performing model is **bold-faced**.

Methods	Backbones	Mini-Sports1M mAP	Mini-Sports1M GFLOPs	COIN Top-1	COIN GFLOPs
LiteEval [43]	ResNet50	44.7%	66.2	-	-
SCSampler [17]		44.3%	42.0	79.8%	42.0
AR-Net [26]		45.0%	37.6	-	-
AdaFuse [27]		44.1%	60.3	-	-
OCSampler [24]		46.7%	25.8	80.1%	25.8
SOSampler		**48.3%**	25.8	**80.7%**	25.8
OCSampler [24]	TimeSformer	45.6%	76.8	81.4%	76.8
SOSampler		**49.1%**	76.8	**87.7%**	76.8

Table 4: **Comparison on Mini-Sports1M and COIN Comparison on for small N and T**. The best performing model is **bold-faced**.

Methods	Backbones	ActivityNet mAP	ActivityNet GFLOPs	Mini-Kinetics Top-1	Mini-Kinetics GFLOPs
LiteEval [43]	ResNet50	72.7%	95.1	61.0%	99.0
SCSampler [17]		72.9%	42.0	70.8%	41.9
AR-Net [26]		73.8%	33.5	71.7%	32.0
videoIQ [33]		74.8%	28.1	72.3%	20.4
AdaFocus [41]		75.0%	26.6	72.9%	38.6
FrameExit [11]		76.1%	26.1	72.8%	19.7
OCSampler [24]		77.2%	25.8	73.0%	21.6
SOSampler		**77.7%**	25.8	**73.5%**	21.6
Ada2D [22]	SlowOnly50	84.0%	701	79.2%	738
OCSampler [24]		87.3%	68.2	82.6%	27.3
SOSampler		**88.0%**	64.0	**83.0%**	27.3
FrameExit [11]	X3D-S	86.0%	9.8	–	–
OCSampler [24]		86.6%	7.9	–	–
SOSampler		**87.2%**	7.6	–	–
OCSampler [24]	TimeSformer	83.2%	76.8	80.7%	76.8
SOSampler		**88.7%**	76.8	80.7%	76.8

Table 3: **Comparison on ActivityNet-v1.3 and Mini-Kinetics for small N and T**. The best performing model is **bold-faced**.



[1] Lin, Jintao, et al. "Ocsampler: Compressing videos to one clip with single-step sampling." CVPR. 2022.