# OCSampler: Compressing Videos to One Clip with Single-step Sampling (Supplementary Materials)

## 1. Introduction of Prior Works

OCSampler is compared with several competitive works that focus on efficient video recognition, including AdaFrame [16], LiteEval [15], SCSampler [7], AR-Net [11], VideoIQ [12], AdaFocus [13], Ada2D [8], ListenToLook [2], MARL [14], and FrameExit [3].

- AdaFrame [16] learns to dynamically select informative frames with reinforcement learning and performs adaptive inference.
- LiteEval [15] combines a coarse LSTM and a fine LSTM to adaptively allocate computation based on the importance of frames.
- SCSampler [7] introduces a light-weighted framework to efficiently identify the most salient temporal clips within a long video. We follow the implementation of [11].
- AR-Net [11] dynamically identifies the importance of video frames, and processes them with different resolutions accordingly.
- VideoIQ [12] learns to dynamically select optimal quantization precision conditioned on input clips.
- AdaFocus [13] dynamically processes video frames with different patches accordingly.
- Ada2D [8] learns instance-specific 3D usage policies to determine frames and convolution layers to be used in a 3D network.
- ListenToLook [2] fuses image and audio information to select the key clips within a video
- MARL [14] proposes to learn to select important frames with multi-agent reinforcement learning.
- FrameExit [3] adopts a deterministic policy function and gating modules to determine the earliest exiting point for inference.

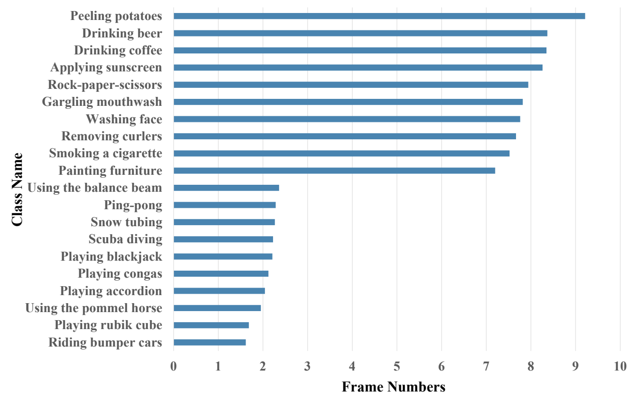


Figure 1. **The Top-10 classes that require the most and the least number of frames in average.** Specifically, videos whose backgrounds contribute a lot demand less computational cost, while videos containing continuous and subtle actions require more frame number budgets. We visualize some cases in Figure 4.

## 2. Implementation Details

In our implementation, we train $f_S$ and $f_C$ using an SGD optimizer with cosine learning rate annealing and a Nesterov momentum of 0.9 [4, 9, 11, 13]. The size of the mini-batch is set to 64, while the weight decay is set to 1e-4. For ImageNet pretrained settings, we initialize $f_S$ and $f_C$ with ImageNet pretrained MobileNetV2-TSM [9] and ResNet-50 [4]. For Kinetics pretrained settings, we initialize models with Kinetics-400 pretrained weight and fine-tune them on the target dataset. In stage I, we warm up $f_S$ and $f_C$ using uniformly sampled frames for 50 epochs with an initial learning rate of 0.01 and 0.005, respectively. In stage II, we train $\pi$ with an SGD optimizer with cosine learning rate annealing for 50 epochs and an initial learning rate of 0.001. We conduct all experiments on 8 TITAN XPs and will release our codes public to facilitate future works.

## 3. The Ability of Adaptive Selection

We statistically analyze the number of frames used in different categories. Figure 1 shows the Top-10 classes that require the most and the least number of frames. The number of frames required by different video classes varies significantly, affected by the complexity of video content.

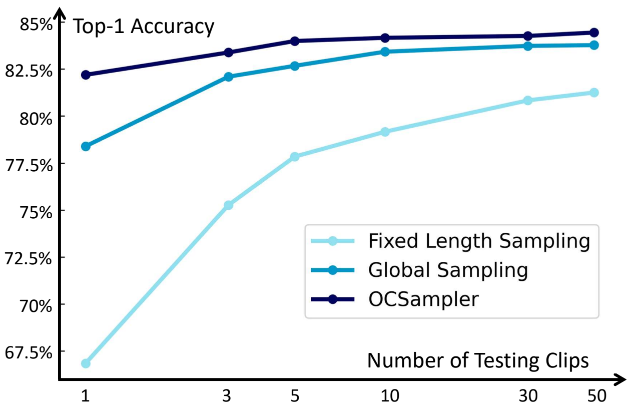


Figure 2. **Different sampling strategies with multi-clips on ActivityNet-v1.3.** OCSampler achieves more competitive recognition performance with only one-clip testing over other strategies with multi-clip testing.

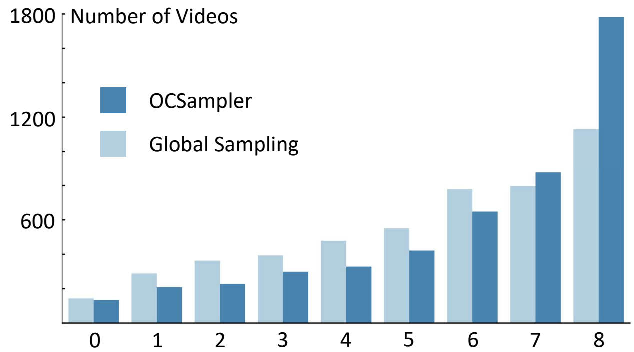


Figure 3. **Validation with instance-level annotations.** We demonstrate how many videos have $M(0 \leq M \leq 8)$ sampled frames in the annotated segments of ActivityNet-v1.3 validation set. OCSampler can gather more significant frames (which fall into the ground-truth segments).

We provide additional visualization examples to illustrate the learned policy by OCSampler+ in Figure 4. Videos are uniformly sampled in 10 frames. OCSampler+ compresses videos into one clip with informative frames, and dynamically adjusts frame number budgets for different content of videos to further reduce computational costs. Specifically, Videos whose backgrounds contribute a lot (*e.g.*, "Ping Pong" and "Riding Bumper Cars" in the top 2 examples of Figure 4) require less computational overhead, while videos containing continuous and subtle actions (*e.g.*, "Gargling Mouthwash" and "Peeling Potatoes" in the bottom 2 examples of Figure 4) take more frame number budgets for classification.

# 4. Temporal Localization Results

We further extend OCSampler to the temporal localization task. Specifically, we first use BMN [10] to extract action proposals and then use SlowOnly-R50 (which takes 8 frames as input) equipped with OCSampler to assign action labels to each proposal. For comparison, we also report the localization performance of using SlowOnly-8x8 trained with fix-length sampling to assign action labels (with 10-clip testing). Table 1 shows that OCSampler can achieve better localization results with far less computation consumed.

| Methods | GFLOPs | mAP | AP@0.5 | AP@0.6 | AP@0.7 | AP@0.8 | AP@0.9 |
|---|---|---|---|---|---|---|---|
| SlowOnly | 549 | 26.9 | 37.0 | 33.5 | 30.0 | 25.2 | 17.0 |
| OCSampler | **68** | **28.2** | **38.8** | **35.1** | **31.4** | **26.5** | **17.8** |

Table 1. **Localization Results.** We compare the action localization performance of OCSampler and SlowOnly (fix-length sampling, 10-clip testing). OCSampler achieves superior localization performance with far less computation.

# 5. Multi-Clip Results

In this section, we compare our OCSampler using multi-clip testing with two standard sampling strategies: *Fixed-Length* and *Global*. *Fixed-Length* samples frames only in a short temporal window to form a clip, while *Global* selects frames uniformly over the entire videos. Here, we use SlowOnly-R50 with Kinetics pretrained weight on ActivityNet, and each clip is built with 8 frames. Figure 2 demonstrates that OCSampler outperforms other strategies with only one clip by a large margin in recognition accuracy and efficiency.

# 6. Validation with Instance-level Annotations

Besides the improved recognition performance, we find that more frames sampled by OCSampler fall into the annotated action segments compared to *Global* Sampling (Figure 3), which validates OCSampler's capability to sample informative frames from another angle. Here we set $T = 32$ and $N = 8$.

# 7. Training approaches.

Our REINFORCE technique is not hard to train since we adopt the uniformly sampled result as baseline in our reward function in Eq.9 to stabilize the training process. We retrain OCSampler 5 times and obtain $77.25 \pm 0.07\%$ mAP. We also use gumbel-softmax to train OCSampler 5 times and obtain $76.32 \pm 0.41\%$ mAP. By comparison, our training scheme is more stable and achieves higher accuracy.

# 8. Transfer learned policies.

For samplers trained on different datasets, we directly adopt them for frame sampling on other target datasets with off-the-shelf video classifiers. Table 2 shows that training and testing a sampler on the same dataset provides the best performance. However, there is only a negligible drop for cross-dataset training-testing, showing the good transferability of our method.

| Train \ Test | ActivityNet | FCVID | Mini-Sports1M | Mini-Kinetics |
|---|---|---|---|---|
| ActivityNet | **77.2%** | 82.6% | 46.6% | 73.5% |
| FCVID | 77.1% | **82.7%** | 46.5% | 73.4% |
| Mini-Sports1M | 76.7% | 82.2% | **46.7%** | 73.1% |
| Mini-Kinetics | 77.1% | 82.4% | 46.4% | **73.7%** |

Table 2. **Transferring learned policies.** Diagnoal numbers refer to training and testing a sampler on the same dataset while non-diagnoal numbers refer to cross-dataset training-testing.

## 9. Dataset License

ActivityNet-v1.3 [1] dataset is licensed under an MIT license and Kinetics [6] dataset is licensed by Google Inc. under a Creative Commons Attribution 4.0 International License. The Sports-1M [5] dataset is made available under a Creative Commons License.

## References


[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 3

[2] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 1

[3] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15608–15618, 2021. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 3

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[7] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019. 1

[8] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. 2d or not 2d? adaptive 3d convolution selection for efficient video recognition. *arXiv preprint arXiv:2012.14950*, 2020. 1

[9] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 1

[10] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2019. 2

[11] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *European Conference on Computer Vision*, pages 86–104. Springer, 2020. 1

[12] Ximeng Sun, Rameswar Panda, Chun-Fu Richard Chen, Aude Oliva, Rogerio Feris, and Kate Saenko. Dynamic network quantization for efficient video inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7375–7385, 2021. 1

[13] Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16249–16258, October 2021. 1

[14] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6222–6231, 2019. 1

[15] Zuxuan Wu, Caiming Xiong, Yu-Gang Jiang, and Larry S Davis. Liteeval: A coarse-to-fine framework for resource efficient video recognition. *arXiv preprint arXiv:1912.01601*, 2019. 1

[16] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019. 1


Figure 4. **Qualitative examples.** Our proposed approach **OCSampler+** processes more informative frames to form a clip for more complex videos, and takes fewer frames for simpler ones to avoid temporal redundancy and further save computational costs. Best viewed in color.