Appendix for “Adaptive Focus for Efficient Video Recognition”

A. Introduction of Baselines

AdaFocus is compared with several competitive baselines that focus on facilitating efficient video recognition, including MultiAgent [9], SCSampler [4], LiteEval [11], AdaFrame [10], Listen-to-look [1] and AR-Net [6].

- MultiAgent [9] proposes to learn to select important frames with multi-agent reinforcement learning.
- SCSampler [4] introduces a light-weighted framework to efficiently identify the most salient temporal clips within a long video. We follow the implementation of [6].
- LiteEval [11] combines a coarse LSTM and a fine LSTM to adaptively allocate computation based on the importance of frames.
- AdaFrame [10] learns to dynamically select informative frames with reinforcement learning and performs adaptive inference.
- Listen-to-look [1] fuses image and audio information to select the key clips within a video. As we do not leverage the audio of videos, for a fair comparison, we adopt its image-based version introduced in their paper.
- AR-Net [6] dynamically identifies the importance of video frames, and processes them with different resolutions accordingly.

B. Implementation Details

B.1. Training Hyper-parameters for Section 4.1

In our implementation, we always train f_{G}, f_{L} and f_{C} using a SGD optimizer with cosine learning rate annealing and a Nesterov momentum of 0.9. The size of the mini-batch is set to 64, while the L2 regularization coefficient is set to 1e-4. We initialize f_{G} and f_{L} by fine-tuning the ImageNet pre-trained MobileNet-V2 [7] and ResNet-50 [2][1] using full inputs for 15 epochs with an initial learning rate of 0.01. In stage I, we train f_{L} and f_{C} using randomly sampled patches for 50 epochs with an initial learning rate of 5e-4 and 0.05, respectively. Here we do not train f_{G} as we find this does not significantly improve the performance, but increases the training time. In stage II, we train π/π' with an Adam optimizer [3] for 50/10 epochs. The same training hyper-parameters as [8] are adopted. In stage III, we only fine-tune f_{C} with the learned policy for 10 epochs, since we find further fine-tuning f_{L} leads to trivial improvements but prolongs the training time. The initial learning rates are set to 5e-4 and 5e-3 for Mini-Kinetics and ActivityNet/FCVID, respectively.

[1] We use the official models provided by PyTorch.

B.2. Training Hyper-parameters for Section 4.2

Here we initialize f_{G} and f_{L} by training them using the same configuration as [5]. The training procedure of AdaFocus is the same as Section 4.1 except for the following changes. In stage I, we use the initial learning rate of 1e-5 and 0.01 for f_{L} and f_{C}, respectively, and train them for 10 epochs. In stage III, we use an initial learning rate of 5e-4 for f_{C}. Note that TSM+ follows exactly the same training procedure as our method. The only difference is that TSM+ does not train the policy network π, since it adopts full frames as inputs.

References

[1] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, pages 10457–10467, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *ICCV*, pages 6232–6242, 2019. 1

[5] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019. 1

[6] 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 *ECCV*, pages 86–104. Springer, 2020. 1

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 1

[8] Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, and Gao Huang. Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. In *NeurIPS*, 2020. 1

[9] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *ICCV*, pages 6222–6231, 2019a. 1

[10] Zuxuan Wu, Hengduo Li, Caiming Xiong, Yu-Gang Jiang, and Larry Steven Davis. A dynamic frame selection framework for fast video recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020b. 1

[11] Zuxuan Wu, Caiming Xiong, Yu-Gang Jiang, and Larry S Davis. Liteeval: A coarse-to-fine framework for resource efficient video recognition. In *NeurIPS*, 2019b. 1