Supplementary Materials

1. Additional baselines for efficient video object detection methods

Table 1 compares our approach to efficient video object detection baselines on ImageNet-VID. LSTS [1] introduces additional modules to enhance per-frame features through temporal correlation, PatchNet [2] exploits patchwise correlation to skip computation. Although PatchNet significantly reduces computation, it suffers mAP reduction by $>7\%$.

Method	**Backbone**	**mAP-50**	**Sparsity**
LSTS	ResNet-101+DCN	0.801	
PatchNet	ResNet-101	0.731	0.80
MaskVD	ViT-B	**0.805**	0.80

Table 1. Efficient video object detection on ImageNet-VID.

Keep Rate	P	k_s	**mAP-50**	**GFLOPs**	**Latency (ms)**
1.0	-	-	81.46	467.4	134.42
0.44	4	0.1	81.01	217.14	111.03

Table 2. MaskVD on ImageNetVID with resolution 1024×1024.

2. Scalability with increasing video resolution or longer video sequences

Our method can mask $\sim 56\%$ regions with 0.4% mAP drop on a subset of ImageNet-VID with input size 1024×1024 (see Table 2). This demonstrates the scalability of our method to increasing video resolution. Results on BDD100K shown in the paper containing video sequences up to 40s in length also demonstrate the scalability to longer video sequences.

References

[1] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. Learning where to focus for efficient video object detection. In *European Conference on Computer Vision*, 2020.

[2] Huizi Mao, Sibo Zhu, Song Han, and William J. Dally. Patchnet – short-range template matching for efficient video processing. *arXiv preprint arXiv:2103.07371*, 2021.