# VideoMAE reading notes

## 1. Method

(1) temporal downsampling

original video $V$ —randomly sample→ one video clip with $t$ consecutive frames —temporal sampling→

$T$ frames, each $H \times W \times 3$, $T = \frac{t}{\tau}$, stride $\tau = 2$ or $4$

(2) cube embedding

joint space-time cube embedding

$2 \times 16 \times 16 \times 3$ size = one token embedding $\longrightarrow \frac{T}{2} \times \frac{H}{16} \times \frac{W}{16}$ 3D tokens

each token —map→ channel dimension $D$

(3) tube masking with high ratios 90%~95%

masking map is the same for all frames

(4) backbone:

vanilla ViT backbone with joint space-time attention

## 2. Datasets

HMDB51 3.5k / 1.5k train/val videos

**HMDB51.** Our VideoMAE is pre-trained with a masking ratio of 75% for 4800 epochs. The batch size and base learning rate are set to 192 and 3e-4, respectively. Here, 16 frames with a temporal stride of 2 are sampled. For fine-tuning, the model is trained with repeated augmentation [32] and a batch size of 128 for 50 epochs. The base learning rate, layer decay and drop path are set to 1e-3, 0.7 and 0.2, respectively. For evaluation, we adopt the inference protocol of 10 clips × 3 crops.