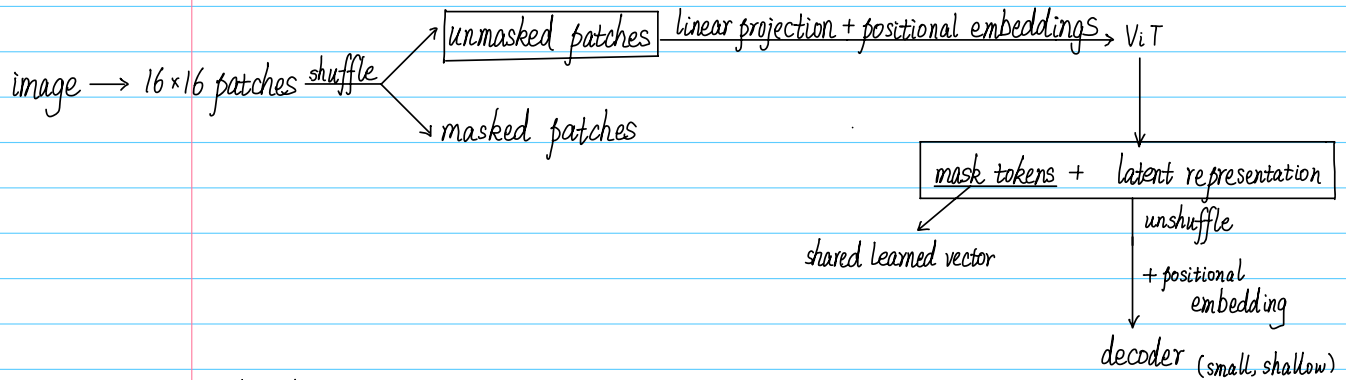


## MAE reading notes

### 1. Methods



\* decoder 只在 pretrain 中用到

\* skip mask token in encoder and apply it later in lightweight decoder → asymmetric

### 2. Tips

supervised ViT-L should train with strong regularization.

config	value
optimizer	AdamW
base learning rate	1e-4
weight decay	0.3
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
batch size	4096
learning rate schedule	cosine decay
warmup epochs	20
training epochs	300 (B), 200 (L/H)
augmentation	RandAug (9, 0.5) [12]
label smoothing [52]	0.1
mixup [69]	0.8
cutmix [68]	1.0
drop path [30]	0.1 (B), 0.2 (L/H)
exp. moving average (EMA)	0.9999

Table 11. Supervised training ViT from scratch.