Appendix

A. Technical Details

Model configurations. In all our experiments, we use ViT as the teacher model and Adventurer as the student model—both featuring a plain (non-hierarchical) design that maintains consistent spatial resolutions across layers. Their detailed configurations are summarized in Table 7.

Model	Embedding dimension	MLP dimension	Blocks	Parameters
ViT-Base, Patch size 16×16	768	3,072	12	86M
ViT-Large, Patch size 14×14	1,024	4,096	24	307M
Adventurer-Small, Patch size 16×16	512	1,280	12	44M
Adventurer-Base, Patch size 16×16	768	1,920	12	99M
Adventurer-Large, Patch size 14×14	1,024	2,560	24	346M

Table 7. Detailed configuration of the models used in this paper.

Training recipes. In our distillation stage, we did not perform extensive hyperparameter tuning. Instead, we mainly followed the settings adopted in prior ViT-based masked distillation studies [36], but applied stronger data augmentation and higher drop path rates, which previous findings [46, 47] suggest are better suited for Mamba-style models. Detailed hyper=parameters can be found in Table 8 and 9. For semantic segmentation fine-tuning, we simply follow the recipe in [36].

Config	Small/Base	Large
optimizer	AdamW	
peak learning rate	1.5e-3	
minimum learning rate	1e-5	
weight decay	0.05	
epochs	300	
optimizer betas	0.9, 0.999	
batch size	2048	
warmup epochs	10	20
stochastic depth (drop path)	0.1	0.2
layer-wise lr decay	✗	
label smoothing	✗	
random erasing	✗	
Rand Augmentation	✗	
repeated augmentation	✓	
ThreeAugmentation	✓	

Table 8. Configurations of the distillation stage.

Config	Small/Base	Large
optimizer	AdamW	
peak learning rate	5e-4	
minimum learning rate	1e-6	
weight decay	0.05	
epochs	100	50
optimizer betas	0.9, 0.999	
batch size	1024	
warmup epochs	20	5
stochastic depth (drop path)	0.4	0.6
layer-wise lr decay	0.65	0.8
label smoothing	✓	
random erasing	✗	
Rand Augmentation	rand-m9-mstd0.5-inc1	

Table 9. Configurations of the fine-tuning stage.