Appendix

In this appendix, we provide details of our training regime and provide additional evaluation on ImageNet1K. First, in Sec. 6.1, we detail hyperparameters and other details about our training regime for ViTTM. Second, in Sec. 6.2, we compare ViTTMs with Token Merging [1], and show results for a small ViTTM model (ViTTM-S). Our code can be found at https://github.com/pjjajal/EfficientTTMs.

6.1. Training Configurations

In Sec. 4.1, we briefly described our training recipe for ViTTM. Here, Tab. 7 provides details of both our pre-training and fine-tuning regimes to enhance reproducibility of our work. ”RRC” indicates the use of random resize crop. The ”CE” Loss refers to Cross Entropy, and ”BCE” refers to Binary Cross Entropy.

	Pre-training	**Fine-Tuning**
Eff. Batch size	4096	2048
Optimizer	AdamW	AdamW
LR	1.5×10^{-4}	0.25^{-4}
Warmup LR	1.0×10^{-6}	1.0×10^{-6}
Min. LR	1.0×10^{-7}	1.0×10^{-7}
LR decay	cosine	cosine
Weight decay	0.03	0.1
Warmup epochs	3	5
Stoch. Depth	0.1	0.1
Gradient Clip.	1.0	1.0
Image Size	224	224
Horiz. flip	✓	✓
RRC	✓	✓
RandAug Ops	×	2
RandAug Mag.	×	20
Mixup alpha	×	0.8
CutMix alpha	×	0.8
Erasing prob.	×	0.25
Loss	CE	BCE

Table 7. Training configurations for ViTTM-B models. All training was performed on NVIDIA A100 80GB GPU’s.

6.2. Extra Results

We present extra ImageNet-1k results in Tab. 8. Specifically, we trained a ViTTM-S model, and include comparisons against ViT-S and ViT-B augmented with Token Merging at various pruning rates (without fine-tunign). ViTTMs consistently have lower latency than state-of-the-art methods while matching their accuracy. Compared with Token Merging, ViTTMs achieve higher accuracy (as expected), while having lower latency across a range of pruning ratios (r).

Model Class	**Model**	**Params** (M)	**GFLOPs** ↓	**Latency** (ms)↓	**Top-1**(%)↑
ViT/DeiT	ViT-S/16	22	4.25	149.5	74.2
	DeiT-S/16	22	4.25	152.0	79.8
	ViT-B/32	88	4.37	138.3	72.2
	ViT-B/16	87	16.87	529.5	81.0
	DeiT-B/16	87	16.87	529.7	81.8
Two-Stream	CrossViT-S	27	5.63	235.7	81.0
	CrossViT-15	28	5.81	249.1	82.3
	CrossViT-15†	28	6.13	252.3	81.5
	CrossViT-B	105	21.22	728.1	82.2
	CrossViT-18	43	9.05	374.1	82.5
	CrossViT-18†	44	9.50	378.2	82.8
	Rev-ViT-B	86	17.49	556.5	81.8
	LookupViT$_{3\times3}$	90	5.26	230.5	77.9
	LookupViT$_{5\times5}$	90	6.94	297.2	81.6
	LookupViT$_{7\times7}$	90	9.45	379.5	83.0
	LookupViT$_{10\times10}$	90	14.80	563.4	83.9
Token Merging [1]	ViT-S/16$_{(r=2)}$	22	4.31	172.7	74.0
	ViT-S/16$_{(r=4)}$	22	4.02	161.6	73.8
	ViT-S/16$_{(r=8)}$	22	3.41	138.2	73.1
	ViT-S/16$_{(r=10)}$	22	3.14	125.9	72.5
	ViT-S/16$_{(r=12)}$	22	2.85	115.2	71.6
	ViT-S/16$_{(r=14)}$	22	2.57	103.3	70.4
	ViT-S/16$_{(r=16)}$	22	2.30	94.0	68.1
	ViT-B/16$_{(r=2)}$	86	16.46	551.1	81.0
	ViT-B/16$_{(r=4)}$	86	15.34	515.7	80.9
	ViT-B/16$_{(r=8)}$	86	13.12	440.6	80.4
	ViT-B/16$_{(r=10)}$	86	12.02	402.0	80.1
	ViT-B/16$_{(r=12)}$	86	10.93	367.0	79.6
	ViT-B/16$_{(r=14)}$	86	9.84	330.2	78.9
	ViT-B/16$_{(r=16)}$	86	8.78	296.4	77.6
Ours	ViTTM-S$_{(28,28)}$	33	1.84	77.7	79.2
	ViTTM-B$_{(28,28)}$	127	7.08	234.1	82.9
	ViTTM-B$_{(32,16)}$	125	7.10	251.5	80.9

Table 8. Comparison of ViTTM with state-of-the-art methods on image classification (ImageNet-1K). Latency was measured on a 80GB A100 with batch size 256. *Notes*: The ViT baseline model is the 224 resolution fine-tuned model from [8], available from `timm` [40]. LookupViT does not have a public implementation, as such we implement a version following the paper. Token Merging [1] is applied to ViT-S/16 and ViT-B/16 models at various pruning rates (r) *without* fine-tuning.