A Appendix

A .1 Training Strategies

In Table 10, we present the training strategies of all the models in this paper. These strategies are the same as those in their original papers for ImageNet-1k training. Note that we don't employ 'EMA' for small-scale image recognition studies (Section 4 .1), since it hurts the performance of all models by a large margin. For the ImageNet-1k classification task (Section 4 .3), we set the 'EMA' to 0.99996, which is identical to that of sMLPNet so as to enable a fair comparison.

Table 10: Training strategies of different models

Configs	ResNet 18, 34, 50 [33]	ConvMixer 768_32 [21]	DeiT T, S [22]	Swin T [23]	CCT 7-3×1 [16]	NesT T [17]	ResMLP S12, S24 [2]
Training epochs	300	300	300	300	300	300	400
Batch size	2048	640	1024	1024	1024	512	1024
Optimizer	LAMB	AdamW	AdamW	AdamW	AdamW	AdamW	LAMB
LR	5e-3	1e-2	1e-3	1e-3	5e-4	5e-4	5e-3
LR decay	cosine	onecycle	cosine	cosine	cosine	cosine	cosine
Min LR	1e-6	1e-6	1e-5	5e-6	1e-5	0	1e-5
Weight_decay	0.02	0.00002	0.05	0.05	0.05	0.05	0.2
Warmup epochs	5	0	5	20	10	20	5
Warmup LR	1e-4	–	1e-6	5e-7	1e-6	1e-6	1e-6
Rand Augment	7/0.5	9/0.5	9/0.5	9/0.5	9/0.5	9/0.5	9/0.5
Mixup	0.1	0.5	0.8	0.8	0.8	0.8	0.8
Cutmix	1.0	0.5	1.0	1.0	1.0	1.0	1.0
Stoch. Depth	0.05	0	0.1	0.2	0	0.2	0.1
Repeated Aug	✓	✗	✓	✗	✗	✗	✓
Erasing prob.	0	0.25	0.25	0.25	0.25	0.25	0.25
Label smoothing	0	0.1	0.1	0.1	0.1	0.1	0.1
EMA	–	–	–	–	–	–	–

Configs	CycleMLP B1, B2 [3]	HireMLP Ti, S [4]	Wave-MLP T, S [5]	ViP S7 [6]	DynaMixer S [7]	sMLPNet T, (S, B) [19]	Caterpillar Mi, Tx, T, S, B
Training epochs	300	300	300	300	300	300	300
Batch size	1024	2048, 1024	1024	2048	2048	1024	1024
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
LR	1e-3	1e-3	1e-3	2e-3	2e-3	1e-3	1e-3
LR decay	cosine	cosine	cosine	cosine	cosine	cosine	cosine
Min LR	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5
Weight_decay	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Warmup epochs	5	20	5	20	20	20	20
Warmup LR	1e-6	1e-6	1e-6	1e-6	1e-6	1e-6	1e-6
Rand Augment	9/0.5	9/0.5	9/0.5	9/0.5	9/0.5	9/0.5	9/0.5
Mixup	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Cutmix	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Stoch. Depth	0.1	0	0.1	0.1	0.1	0, (0.2, 0.3)	0, 0, 0, 0.2, 0.3
Repeated Aug	✓	✓	✓	✗	✗	✓	✓
Erasing prob.	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Label smoothing	0.1	0.1	0.1	0.1	0.1	0.1	0.1
EMA	*0.99996*	–	*0.99996*	–	*0.99996*	*0.99996*	*0.99996*

A .2 sparse-MLP Module

The sMLP module is proposed in [19] and also adopted in the Caterpillar block for aggregating global information. To have a comprehensive understanding of the proposed Caterpillar, we also depict the sMLP module in Figure 3. As we can see, the sMLP module consists of three branches: two of them is used to mix information along horizontal and vertical directions, respectively, which is implemented by two H (W) × H (W) linear projections, and the other path is an identity mapping. The output of the three branches are concat and then processed by a 3C × C linear projection to obtain the final output.

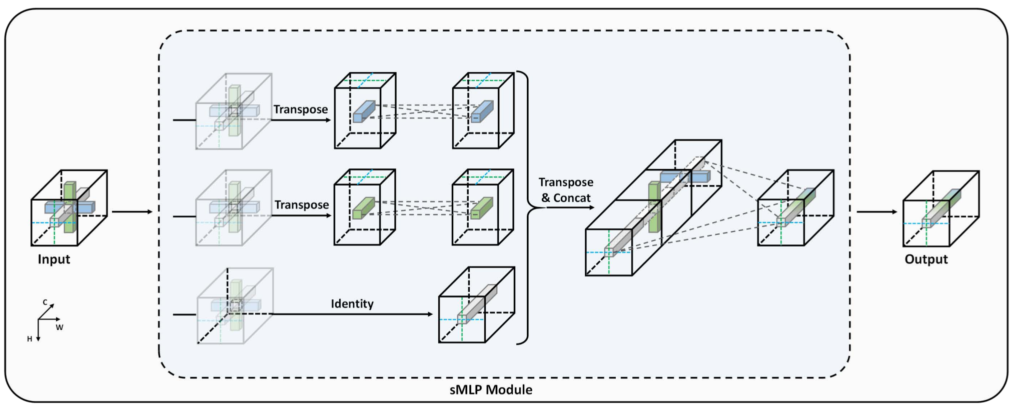

Figure 3: The sparse-MLP module proposed in sMLPNet [19]