SUPPLEMENTARY MATERIAL FOR MULTI-TASK HETEROGENEOUS TRAINING WITH MODULAR ADAPTATION

Anonymous authors
Paper under double-blind review

Table A1: We compare three ways of selecting a subset of experts to fine-tune, while freezing the remaining experts. We first learn new routers on the new downstream to determine each expert's frequency of being chosen. Random represents randomly choosing experts. Best represents choosing the experts with the highest frequency. Worse represents choosing the experts with the lowest frequency. We report mean top-1 accuracy on CUB, Cars, and Pets. Other settings are the same as in Table.3 in the paper.

	Random	Best	Worse
Ro. w/1 Ex.	90.6	90.5	90.6
Ro. w/2 Ex.	92.3	92.3	92.2

A1 DIFFERENT WAYS TO SELECT EXPERTS TO BE FINE-TUNED.

Table.A1 compares various methods of selecting experts to fine-tune while freezing the rest. We compare random selecting experts and selecting experts that are more or less likely to be chosen by routers. We find out that the selection method does not significantly affect the fine-tuning performance. Therefore, we use random selection for simplicity.

A2 ABLATION ON TOP-K.

As shown in Table. A2, we explore the effect on Top-K in MoE module. The experiment setting is the same as in Table.1 in the paper with 12 experts per MoE module. We report the mean performance on pre-train and downstream datasets of our MHTL with Davit-T as the backbone. To control the FLOPs to be the same for different Top-K, the hidden dimension of MLP experts is divided by K. All experiments have the same parameter size and the same FLOPs. We find that Top-$K = 4$ has the best performance.

A3 ABLATION ON THE NUMBER OF EXPERTS.

As shown in Table. A3, we explore the effect on number of experts E for MoE MLP layer. The settings are the same as in § A2 with a Top-K as 4.

Table A2: **Ablation study of Top-K on MoE MLP layer.**

	FLOPs(G)	Params(M)	Hidden Dim	Pre-train mean	Downstream mean
K=2	5.1	51.2	768	58.1	80.3
K=4	5.1	51.2	384	58.2	80.4
K=6	5.1	51.2	256	57.9	80.0

Table A3: **Ablation study of expert number E on MoE MLP layer.**

	FLOPs(G)	Params(M)	Pre-train mean	Downstream mean
E=6	5.1	33.4	57.2	78.5
E=9	5.1	42.3	57.9	80.0
E=12	5.1	51.2	58.2	80.4
E=15	5.1	60.1	58.2	80.5

A4 TRAINING DETAILS

Optimization and convergence. Each task in our framework has a dedicated module and its own loss. The losses on datasets D_i are weighted and alternately optimized with predetermined weights w_{l_i}. Gradient conflicts between tasks pose a challenge, slowing convergence. Well-defined loss and sampling weights contribute to training stability, and the large batch optimizer Lamb (**?**) is effective in heterogeneous training. Convergence in this setting typically requires approximately 50% more iterations than single-task training due to the complexity of joint optimization. Loading pre-trained single-task models can significantly accelerate training, as discussed in the next section.

Training details. During pre-training, data sampling weight is set to $\{3, 2, 1\}$, loss weight is set to $\{1.0, 0.6, 0.2\}$, and batch size is set to $\{64, 2, 2\}$ for classification, detection, and segmentation, respectively. Weight decay is set to 0.05 and the maximal gradient norm is clipped to 0.1. We use a simple triangular learning rate schedule with a maximum learning rate of 0.004, as in DaviT.