BiMAE - A Bimodal Masked Autoencoder Architecture for Single-Label Hyperspectral Image Classification

Supplementary Material

Contents:

A . ViT-Small configuration

Parameter	Value
patch_size	24
encoder_dim	384
encoder_num_heads	6
depth	12
mlp_ratio	4
cross_attn_num_heads	3
decoder_dim	192

Table 5. *ViT*-Small configuration

B . Pretraining settings

Hyperparameters	Value
Optimizer	AdamW[38]
Base learning rate	1e-4
Weight decay	0.05
Adam (β_1, β_2)	(0.9, 0.999)
Batch size	512
Learning rate sched.	Cosine decay[37]
Training epochs	300
Warmup learning rate	1e-6
Warmup epochs	30
Non-masked hs tokens	15
Non-masked rgb tokens	16
Target hs mask ratio (s_{hs})	0.2
Target rgb mask ratio (s_{rgb})	0.5
Augmentation	HorizontalFlip, VerticalFlip

Table 6. Pretraining setting.

C . Finetuning settings

Hyperparameters	Value
Optimizer	AdamW [38]
Base learning rate	5e-4
Weight decay	0.05
Adam (β_1, β_2)	(0.9, 0.999)
Batch size	512
Learning rate sched.	Cosine decay[37]
Training epochs	50
Warmup learning rate	1e-6
Warmup epochs	3
Augmentation	HorizontalFlip, VerticalFlip

Table 7. Finetuning settings.

D . Class-wise accuracy evaluation

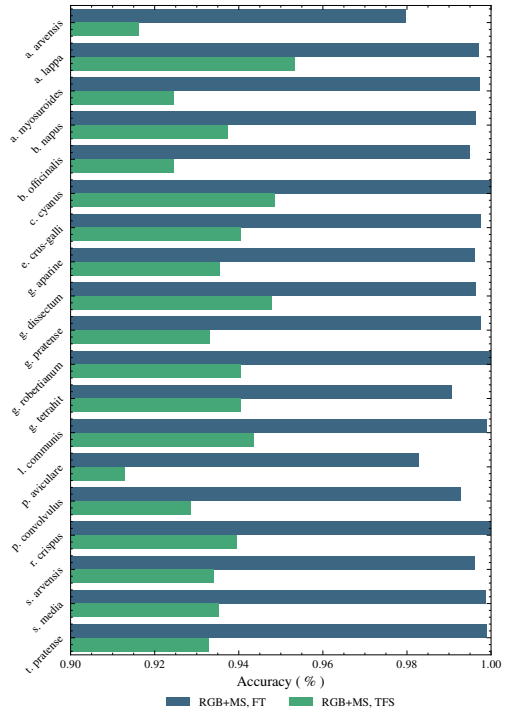

Figure 5. Class-wise accuracy comparison among BiMAE, finetuned, and trained from scratch models using bimodal data.

As there were only slight differences between the finetuned (99.55%) and trained from scratch (99.28%) BiMAE on bimodal data, we decided to evaluate the accuracies for each class in the dataset. Figure 5 clearly illustrates the significant differences between the classes, confirming the benefits of pretraining BiMAE.