Supplementary material

A. ImageNet100

In this paper, we utilized ImageNet100 to train the ViT and other Vision transformer backbones and to measure the accuracy of trained models. ImageNet100 is the subset of the ImageNet dataset, which contains images of 100 classes from the original dataset. It consists of 130,000 images (1,300 images per class) for the training set and 5,000 images (50 images per class) for the validation set. This condensed dataset is available for download on Kaggle.

B. Training details

1) T2T-ViT

T2T-ViT-7 The embed size in T2T-ViT-7 is 256, the patch size is 16, the depth is 7, the number of heads is 4, the kernel ratio in "performer" is 2.0, the learning rate schedule is "cosine", the initial learning rate is 1e-3, the momentum is 0.9, the weight decay is 0.03 and batch size is 128.

T2T-ViT-14 The embed size in T2T-ViT-14 is 384, the depth is 14, the number of heads is 6, the kernel ratio in "performer" is 3.0, and other parameters are the same as in T2T-ViT-7.

SMWP in T2T-ViT Our SMWP is embedded in the tokens to token module of T2T. Before the "soft split" unit, we take the mixed features obtained from different frequency domains through SMWP as the split object.

2) TNT

TNT-S The patch size in TNT-S is 16, the inner stride is 4, the outer dim is 384, the inner dim is 24, the depth is 12, the number of outer heads is 6, the number of inner heads is 4, the learning schedule is "step", the initial learning rate is 1e-3, the momentum is 0.9, the weight decay is 1e-4 and the batch size is 128.

TNT-B The inner stride in TNT-B is 4, the outer dim is 640, the inner dim is 40, the depth is 12, the number of outer heads is 10, the batch size is 64, and other parameters are the same as in TNT-S.

SMWP in TNT Our SMWP is embedded in the "PatchEmbed" module of TNT. Before the "proj" unit, we take the mixed features obtained from different frequency domains through SMWP as the projection object.

3) PyramidTNT

PyramidTNT-Ti The outer dim in PyramidTNT-Ti is 80, the inner dim is 5, the depths are [2, 6, 3, 2], the number of outer heads is 2, the number of inner heads is 1, the learning schedule is "step", the initial learning rate is 1e-2, the momentum is 0.9, the weight decay is 1e-4 and the batch size is 128.

PyramidTNT-S The outer dim in PyramidTNT-S is 128, the inner dim is 8, the depths are [2, 8, 4, 2], the number of outer heads is 4, the number of inner heads is 2, the batch size is 64, and other parameters are the same as in PyramidTNT-Ti.

SMWP in PyramidTNT Our SMWP is embedded in the "Stem" module of PyramidTNT. Before the "inner convs" unit, we take the mixed features obtained from different frequency domains through SMWP as the object.

4) Swin

Swin-Ti The embed dim in Swin-Ti is 96, the depths are [2, 2, 6, 2], the numbers of heads are [3, 6, 12, 24], the size of the window is 7, the drop path rate is 0.2, the learning schedule is "cosine", the initial learning rate is 5e-4, the momentum is 0.9, the weight decay is 0.05 and the batch size is 128.

Swin-S The embed dim in Swin-S is 96, the depths are [2, 2, 18, 2], the numbers of heads are [3, 6, 12, 24], the drop path rate is 0.3, and other parameters are the same as in Swin-Ti.

SMWP in Swin Our SMWP is embedded in Swin's "PatchEmbed" module. Before the "proj" unit, we take the mixed features obtained from different frequency domains through SMWP as the projection object.

5) Wave-ViT

Wave-ViT-S The stem hidden dim in Wave-ViT-S is 32, the embed dims are [64, 128, 320, 448], the numbers of heads are [2, 4, 10,14], the mlp ratios are [8, 8, 4, 4], the depths are [3, 4, 6, 3], the learning schedule is "cosine", the initial learning rate is 5e-4, the momentum is 0.9, the weight decay is 0.05 and the batch size is 128.

Wave-ViT-B The stem hidden dim in Wave-ViT-B is 64, the embed dims are [64, 128, 320, 512], the numbers of heads are [2, 4, 10, 16], the depths are [3, 4, 12, 3], and other parameters are the same as in Wave-ViT-S.

SMWP in Wave-ViT Our SMWP is embedded in the "Stem" module of Wave-ViT. Before the "proj" unit, we take the mixed features obtained from different frequency domains through SMWP as the projection object.

6) Dual-ViT

Dual-ViT-S The stem hidden dim in Dual-ViT-S is 32, the embed dims are [64, 128, 320, 448], the numbers of heads are [2, 4, 10,14], the mlp ratios are [8, 8, 4, 3, 2], the depths are [3, 4, 6, 3], the learning schedule is "cosine", the initial learning rate is 5e-4, the momentum is 0.9, the weight decay is 0.05, and the batch size is 128.

Dual-ViT-B The stem hidden dim in Dual-ViT-B is 64, the embed dims are [64, 128, 320, 512], the numbers of heads are [2,4, 10, 16], the depths are [3, 4, 15, 3], the batch size is 64, and other parameters are the same as in Dual-ViT-S.

SMWP in Dual-ViT Our SMWP is embedded in the "Stem" module of Dual-ViT. Before the "proj" unit, we take the mixed features obtained from different frequency domains through SMWP is the projection object.