

# New Experimental Results for Rebuttal

Table 1: Updated version of the main results of ScaleKD.  $\dagger$  denotes the model pre-trained on IN-22K [44] and  $\ddagger$  denotes the model pre-trained by EVA [40]. Underlined results are for the experiments performed by us. **Red results are added in the rebuttal phase.**

| Teacher                       | Student                       | Params (M) |              | FLOPs (G) |             | Accuracy (%) |                |
|-------------------------------|-------------------------------|------------|--------------|-----------|-------------|--------------|----------------|
|                               |                               | T          | S            | T         | S           | Top-1        | $\Delta$ Top-1 |
| Swin-L $^\dagger$ (86.24)     | MobileNet-V1 ( <u>72.10</u> ) |            | 4.23         |           | 0.58        | 75.15        | +3.05          |
|                               | ResNet-50 ( <u>78.64</u> )    | 196.53     | 25.56        | 34.04     | 4.12        | 82.03        | +3.39          |
|                               | ConvNeXt-T (82.14)            |            | 28.59        |           | 4.46        | 84.16        | +2.02          |
|                               | Mixer-S/16 (74.02)            |            | 18.53        |           | 3.78        | 78.63        | +4.61          |
|                               | <b>ResMLP-S12 (76.51)</b>     | 196.53     | <b>15.40</b> | 34.04     | <b>3.03</b> | <b>80.54</b> | <b>+4.03</b>   |
|                               | Mixer-B/16 (76.44)            |            | 59.88        |           | 12.61       | 81.96        | +5.52          |
|                               | ViT-S/16 (79.90)              |            | 22.05        |           | 4.61        | 83.93        | +4.03          |
|                               | <b>PVT-S (79.80)</b>          | 196.53     | <b>24.50</b> | 34.04     | <b>3.80</b> | <b>83.72</b> | <b>+3.92</b>   |
|                               | Swin-T (81.18)                |            | 28.29        |           | 4.36        | 83.80        | +2.62          |
|                               | ViT-B/16 (81.80)              |            | 86.57        |           | 17.58       | 85.53        | +3.73          |
| BEiT-L/14 $^\ddagger$ (88.58) | ResNet-50 ( <u>78.64</u> )    |            | 25.56        |           | 4.12        | 82.34        | +3.70          |
|                               | Mixer-B/14 ( <u>76.62</u> )   | 304.14     | 59.88        | 81.06     | 16.45       | 82.89        | +6.27          |
|                               | ViT-B/14 ( <u>82.02</u> )     |            | 86.57        |           | 23.09       | 86.43        | +4.41          |

Table 2: Comparisons between CNN teachers and ViT teachers. The experiments are performed under the advanced training strategy.

| Student           | Teacher             | Method        | Top-1 (%) |
|-------------------|---------------------|---------------|-----------|
| ResNet-50 (79.80) | ConvNeXt-XL (86.97) | VanillaKD [4] | 81.10     |
| ResNet-50 (78.64) | ConvNeXt-XL (86.97) | ScaleKD       | 81.72     |
|                   | Swin-L (86.24)      |               | 82.02     |
|                   | BEiT-L (88.58)      |               | 82.34     |

Table 4: Experiments on applying CAP to OFA distillation framework. All results are reproduced on the OFA codebase.

| Student           | Teacher        | Method    | Top-1 (%) |
|-------------------|----------------|-----------|-----------|
| ResNet-18 (69.75) | DeiT-T (72.17) | OFA [36]  | 71.33     |
|                   |                | OFA + CAP | 71.60     |

Table 3: Ablation study on the compatibility of DFM with other feature projectors.

| Method          | Top-1 (%) | $\Delta$ Top-1 (%) |
|-----------------|-----------|--------------------|
| <i>Baseline</i> | 76.55     | -                  |
| Linear          | 77.43     | +0.88              |
| Linear + DFM    | 77.78     | +1.23              |
| Conv            | 77.52     | +0.97              |
| Conv + DFM      | 77.97     | +1.42              |

Table 5: Ablation study on the necessity of AC components in the first path of DFM.

| Method          | Top-1 (%) | $\Delta$ Top-1 (%) |
|-----------------|-----------|--------------------|
| <i>Baseline</i> | 76.55     | -                  |
| CAP             | 77.87     | +1.32              |
| DFM (Dir + Alt) | 78.23     | +1.68              |
| DFM (All + Alt) | 78.51     | +1.96              |

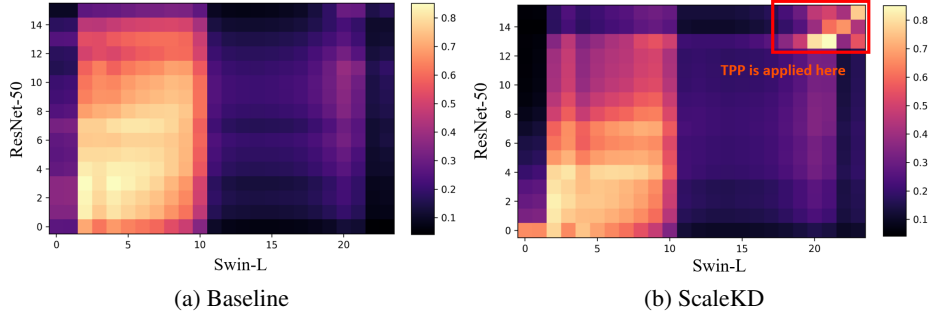


Figure 1: CKA visualization on Swin-L→ResNet-50 teacher-student network pair. We use the output features of each basic block. The calculation of heatmaps is based on 1280 samples in IN-1K [44]. TPP is only applied in the last stage, where it brings obvious improvements to each block.