# Perspective-Aware Teaching: Adapting Knowledge for Heterogeneous Distillation

## Supplementary Material

| Method | T. | S. | KD | OFA | AT | CRD | PAT |
|---|---|---|---|---|---|---|---|
| Acc. | 73.31 | 69.75 | 70.66 | **72.10** | 69.56 | 71.17 | <u>71.18</u> |

Table 1. Result on ImageNet with ResNet34 - ResNet18 homogeneous teacher-student model pair. Our results are the average over 3 trials. The highest results are indicated in bold, while the second-best results are underlined.

## 1. Implementation Details

We follow the training procedure of OFA-KD[1] and employ different training settings for different student models based on their architecture. Specifically, we utilize the SGD optimizer for CNN-based models, whereas we opt for the AdamW optimizer for ViT-based and MLP-based models and additionally incorporate augmentations such as Mixup [3], and CutMix [2]. All models are trained for 300 epochs on CIFAR-100. In the case of ImageNet, CNN-based models are trained for only 100 epochs, while ViT-based and MLP-based student models are trained for 300 epochs.

## 2. Lower Effectiveness in CNN-based students

As shown in the experimental results on CIFAR-100 and ImageNet, PAT yields relatively limited gains on CNN-based student models. We attribute this to limitations of the RAA module. The effectiveness of RAA relies on the overlap between the student's effective receptive field (ERF) and the global context needed for teacher-student feature alignment. CNNs typically exhibit a Gaussian-shaped ERF, limiting global context coverage and leading to suboptimal feature fusion. Consequently, RAA becomes less effective, as CNN-based students lack sufficient global information in intermediate features for attention-based reblending. While there is no architectural incompatibility as attention is a superset of convolution, CNN-based students may require longer training to benefit from RAA, and future work could explore solutions specifically tailored to this scenario.

## 3. Distillation in homogeneous architectures

We further assess our approach through the distillation of homogenous architectures to validate its generalizability. The results are shown in Table 1. When distilling from a homogeneous teacher, our method experiences a 1.43% performance improvement to the student model but has a slight performance drop compared to state-of-the-art OFA. Our proposed module, RAA, is meticulously crafted to reconcile the disparity between the perspectives of the student and teacher models. However, when distilling from models that possess similar perspectives, the efficacy of RAA diminishes, and the incorporation of additional parameters may complicate the training process.

|  | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| Student | 40.74 | 55.98 | 65.28 | 68.00 |
| PAT | 52.23 | 68.21 | 74.58 | 79.59 |

Table 2. Date efficiency experiment result on CIFAR-100 with ConvNeXt-T - DeiT-T teacher-student model pair.

|  |  | ConvNeXt-T | ConvNeXt-T |
|---|---|---|---|
| RAA | AFP | DeiT-T | ResMLP-S12 |
| Baseline (FitNet) |  | 60.71 | 45.47 |
| ✓ |  | 70.12 | 75.04 |
|  | ✓ | 73.07 | 80.12 |
| ✓ | ✓ | 79.59 | 83.50 |

Table 3. The effectiveness of RAA, AFP on CIFAR-100.

## 4. Data efficiency on ViTs in heterogeneous KD

Vision transformers (ViTs) are known for their data-hungry issue, where they need substantial amounts of data to achieve satisfactory performance. On the other hand, CNNs are capable of attaining commendable performance levels with a comparatively modest volume of data. Consequently, we have devised an experiment aimed at investigating whether a ViT-based student model can attain comparable performance levels using a reduced amount of data, particularly when under the distillation of a data-efficient CNN-based teacher model.

Table 2 illustrates the comparative performance of the naive student model alongside our PAT methodology utilizing varying proportions of the training dataset, specifically 25%, 50%, 75%, and 100%. It is evident that our PAT consistently surpasses the naive student model. The absence of inductive bias in the ViT-based model makes it challenging to attain favorable performance levels with restricted data, as evidenced by its modest achievement of 40.74% with merely 25% of the available data. Nevertheless, the ViT-based model demonstrates a significant enhancement in performance when subjected to distillation from a data-efficient CNN-based teacher, thereby highlighting an additional advantage of heterogeneous knowledge distillation.

## 5. Ablation of RAA in Table 4

Table 3 shows that removing RAA results in a significant performance drop, confirming its importance. Without RAA, student must rely on single-stage alignment, losing the ability to integrate multi-layer features to bridge the gap with the teacher's representations.

## References

[1] Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In *NeurIPS*, 2023.

[2] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019.

[3] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.