

# Appendix for Refine Myself by Teaching Myself : Feature Refinement via Self-Knowledge Distillation

## 1. Implementation Details

**Classification** WRN-16-2 consist of three blocks with channel dimensions of 32, 64, and 128 respectively, while ResNet18 consists of four blocks with channel dimensions of 64, 128, 256 and 512, respectively. We set the width hyperparameter,  $w$ , as two for all experiments on classification tasks in the paper. Therefore, channel dimensions of the self-teacher networks are 64, 128, 256 for WRN-16-2 and 128, 256, 512, 1028 for ResNet18.

For Mixup and Cutmix data augmentation, we need to set hyperparameter to model beta distribution that determine the mixing weights [2, 3]. For Mixup, we set hyperparameter as 0.2 for CIFAR-100 and TinyImageNet, and 0.3 for FGVR. For Cutmix, we set hyperparameter as 1.0 for all datasets.

**Semantic segmentation** For semantic segmentation, we set width hyperparameter as one, and we set other settings for network architecture following [1]. We attach the self-teacher network with two repeated BiFPN layers and we do not change the channel dimension of BiFPN layers on semantic segmentation task. Therefore, classifier network with three BiFPN layer including backbone network and the self-teacher network are trained by cross entropy loss from ground truth label. Additionally, the self-teacher network perform distillation for classifier network by soft-label and features.

Since the labeled dataset for semantic segmentation is insufficient, it is common to use a pretrained backbone network. We use pretrained efficientnet-b0 and efficientnet-b1 on ImageNet. Therefore, classifier network consists of pretrained backbone network and BiFPN layers. We apply warm-up and annealing technique to the hyperparameter of FRSKD, because distillation from the beginning of learning could rather be a hindrance to train the backbone network. We set warm-up epoch as 40 and after warm-up we adaptively increase the hyperparameters. We set hyperparameter  $\alpha$  as one and  $\beta$  as 50.

## 2. Sensitivity Analysis

We evaluate FRSKD with varying hyperparameter values to investigate the effect of hyperparameters,  $\alpha$  and  $\beta$ .

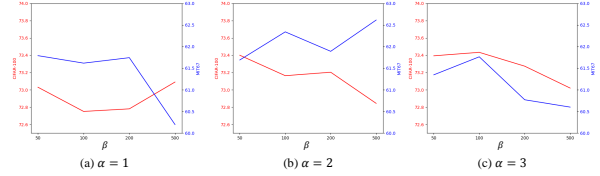


Figure 1: Sensitivity analysis for hyperparameter  $\alpha$  and  $\beta$ . Red line indicates the accuracy of WRN-16-2 on CIFAR-100; and blue line indicates the accuracy of ResNet18 on MIT67. The accuracy is average of three repeated experiments.

We conduct experiments with  $\alpha \in \{1, 2, 3\}$  and  $\beta \in \{50, 100, 200, 500\}$ . Also, we set classifier network as WRN-16-2 on CIFAR-100 and ResNet18 on MIT67. Figure 1 shows the accuracy of each dataset with varying hyperparameters. We keep all-settings except hyperparameters as same as Section 4.1. We find that FRSKD is robust on hyperparameter  $\alpha$  and  $\beta$ , but different hyperparameters perform well for different datasets.

## 3. Qualitative Attention Map Comparison

This section provides additional result of qualitative attention map comparison. We use a channel-wise pooling to the feature map as same as analysis of Section *Further Analyses on FRSKD* in the paper. Figure 2 shows the attention map changes as the learning progresses. As the training progresses, Both the classifier network and the self-teacher network concentrate pm the main object. Additionally, the difference between concentration on the main object is bigger at the early epoch of the training than at the latter part of training.

## References

- [1] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020. 1
- [2] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019. 1

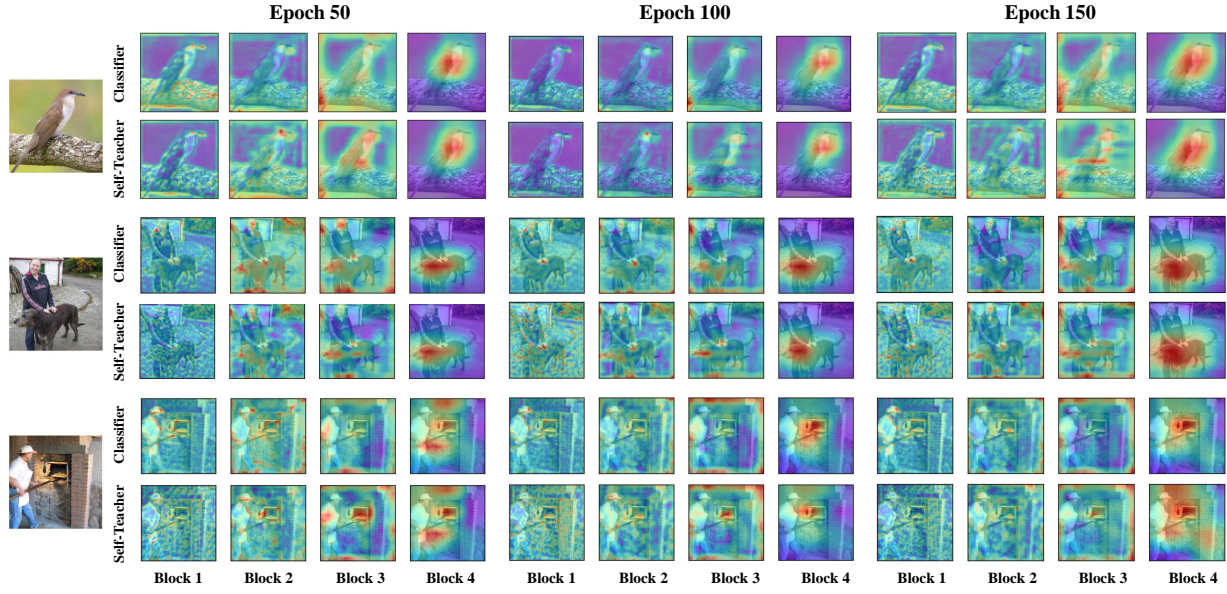


Figure 2: The block-wise attention map comparison between the classifier network and self-teacher network with varying epochs. From above each data is taken from CUB200, Dogs, and MIT67.

- [3] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1