Supplementary Material
Distill on the Go: Online knowledge distillation in self supervised learning

A. Nearest Neighbor Evaluation

Table 1 presents the nearest neighbor evaluation of ResNet and Wider ResNet models on Tiny-ImageNet dataset. The observations made under section 4.3 still hold. All participating models except for WRN10-2 show an improvement over baseline. Smaller models benefit more from joint training than their larger counterparts. Also, their performance improves with the increase in modeling capacity of their counterparts.

Baseline	ResNet-50 (33.07)	ResNet34 (32.07)	ResNet-18 (29.90)
ResNet-50 (33.07)	34.4 \ 34.76 +4.0% +5.1%	34.70 \ 34.06 +4.9% +6.2%	34.05 \ 31.49 +3.0% +5.3%
ResNet34 (32.07)	34.06 \ 34.70 +6.2% +4.9%	33.12 \ 33.96 +3.3% +5.9%	32.57 \ 31.51 +1.6% +5.4%
ResNet-18 (29.90)	31.49 \ 34.05 +5.3% +3.0%	31.51 \ 32.57 +5.4% +1.6%	30.82 \ 30.86 +3.0% +3.2%

Baseline	WRN28-2 (22.03)	WRN16-2 (19.30)	WRN10-2 (17.20)
WRN28-2 (22.03)	22.77 \ 22.44 +3.4% +1.9%	22.14 \ 21.15 +0.4% +9.6%	21.82 \ 19.76 -0.9% +14.9%
WRN16-2 (19.30)	21.15 \ 22.14 +9.6% +0.4%	20.97 \ 20.42 +8.7% +5.8%	20.20 \ 19.59 +4.6% +13.9%
WRN10-2 (17.20)	19.76 \ 21.82 +14.9% -0.9%	19.59 \ 20.2 +13.9% +4.6%	19.17 \ 19.52 +11.5% +13.5%

Table 1. Tiny-ImageNet top-1 accuracy(%) under nearest neighbor evaluation for various ResNet and Wider ResNet models. Baseline measures the linear evaluation accuracy when the models are pre-trained using contrastive learning alone. Each box contains DoGo results: left value corresponds to model in the corresponding row, right value corresponds to model in the corresponding column. Percentage change inaccuracy between baseline and our method are highlighted in green.