Noisy student paper: https://arxiv.org/abs/1911.04252

Main idea - Train the network with labeled images (large number - complete imagenet), use this trained model for inference on unlabeled dataset (JFT dataset, they discarded the labels). Now we have model labeled JFT dataset, use this (samples with confidence of label higher than 0.3 for each class) **AND** the original dataset (Imagenet) as the training data for a new network (similar or larger than the initial network) **AND** add noise in the images in one of the 3 forms (Data augmentation, dropout, Stochastic depth). Get the inference with the new network and repeat the process till we get good results. *Repeated the process for 3 times for the results mentioned in the paper.

Network Arch: EfficientNet-B0...B7 various experiments.
EffNet B0 with 130M images, 130K per class. There are duplications in the data so there are only 81M original images.
Batch size: 2048 (!!!), 512, 1024, 2048 provided similar performance.
Epochs: 350 for student model and network larger than EffNet-B4, smaller nets with 750 epochs.
***Large batch size for unlabeled images.

***Largest model EffNet-L2 (something larger than B7, damn!) - 6 days, Cloud TPU, 2048 cores - unlabeled BS is 14x labeled BS.

Importance of noising the student
'We dont want the student to just learn the teacher's knowledge'. This makes sense, if trained enough, the cross entropy loss would limit to zero at a point.
Gradually remove the noise for the unlabeled dataset (psuedo labels) while keeping it for the labeled dataset. Mainly preventing the overfitting for the labeled dataset.
The ratio of the psuedo labeled to labeled data is important for the performance here, as the case of iterative training - teacher -> student : student as the new teacher and repeat.

Though the whole approach is very insightful, some particular thoughts w.r.t the data sample size are very important considering the current work,

TL;DR

- Large teacher model : performs better.
- Large amt of unlabeled data
- Soft psuedo labels are better than hard encoded **in certain cases**. Later in the Appendix, they mention this is dependent on the task.
- Large student model is also important for better learning process.
- Data balancing is important for smaller models but not so much for the larger models.

- Joint training shows better performance than training on labeled data and then finetuning on the unlabeled data. Appendix - Just training on the psuedo labels shows lower accuracy than supervised learning on the labeled data.
- Large ratio between unlabeled data and labeled data batch size helps in longer training and thus better results. - Appendix: Mainly for larger models.
- Train the student from scratch: in certain cases.

Prev important work:
Self-training is the main idea, inspired this work. Prev work provided amazing results on the Imagenet dataset, but this work is mainly focused on using the noise in the data and the iterative training with joint dataset.
Data distillation (another interesting work which I'd like to get into) - ensemble the prediction for the image with diff transformations to strengthen the teacher.
Co-training (need to look into this) - divide the features into 2 sets and train 2 nets separately using labeled data. The feature partitioning here is the source of noise as the 2 nets don't agree on unlabeled data.
Semi-supervised learning - need more details.
Knowledge distillation - Train the teacher, transfer knowledge to student net with much smaller size while maintaining the performance. This doesn't use unlabeled data and no aim to improve the student, just make it enough with the teacher.