Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

Avital Oliver*, Augustus Odena*, Colin Raffel*, Ekin D. Cubuk, Ian Goodfellow

Google Brain

Background

Semi-supervised learning (SSL leverages unlabeled data when labels are limited or expensive to obtain.

Recent approaches based on neural networks have been successful on standard benchmark tasks.

However, we argue that these benchmarks fail to address real-world settings.

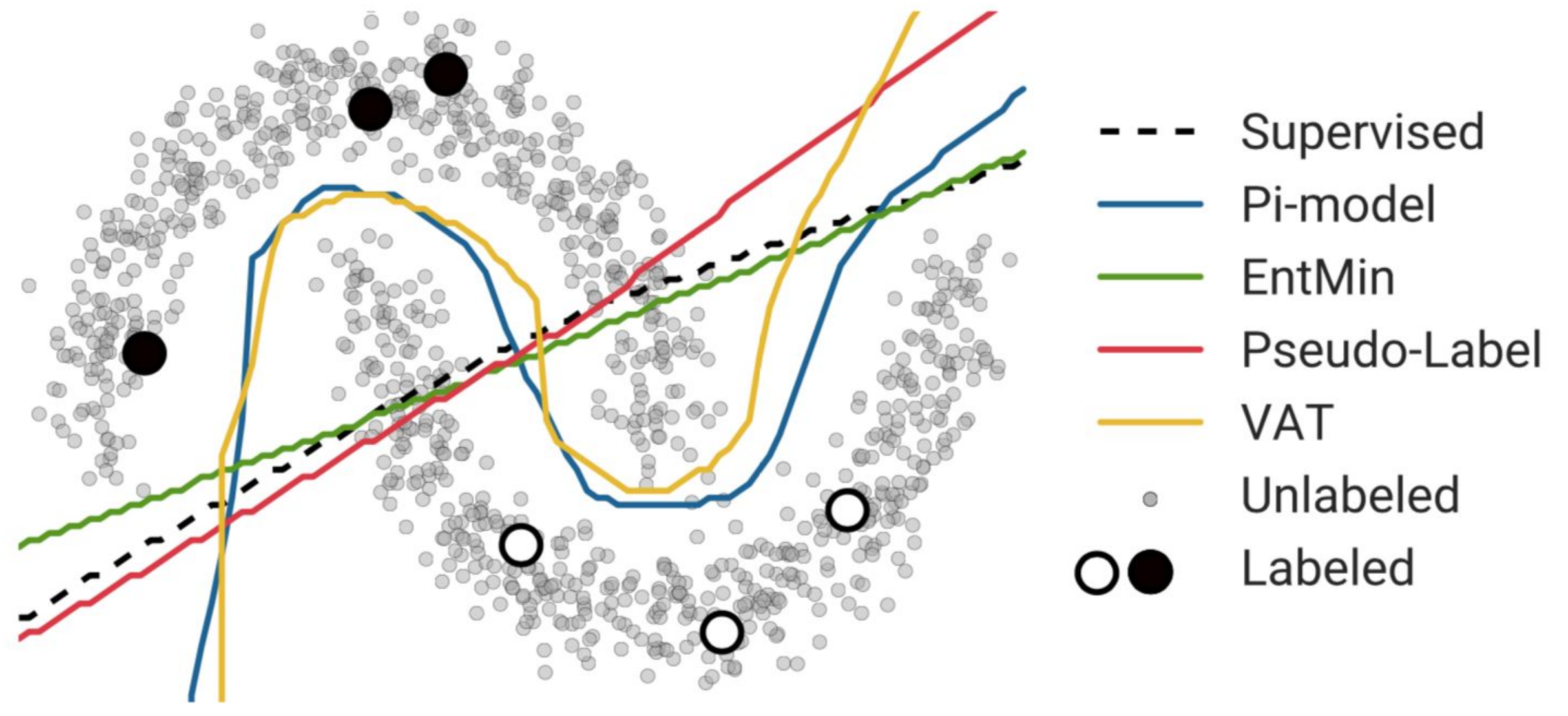


Figure 1. Demonstration of the behavior of different SSL approaches on a simple toy dataset ("two moons"). Training the network on only the labeled data produces a decision boundary which does not follow the contours of the data "manifold", as indicated by additional unlabeled data.

Shared Model

To compare a few widely used SSL techniques, we standardize on a neural network architecture, create a unified reimplementation, and tune all hyperparameters with the same budget.

We test the techniques in a suite of experiments designed to simulate real-world settings.

Method	CIFAR-10 4000 Labels	SVHN 1000 Labels
Π-Model [46]	11.29%	–
Π-Model [32]	12.36%	4.82%
Mean Teacher [50]	12.31%	3.95%
VAT [39]	11.36%	5.42%
VAT + EntMin [39]	10.55%	3.86%
Results above this line cannot be directly compared to those below		
Supervised	20.26 ± 0.38%	12.83 ± 0.47%
Π-Model	16.37 ± 0.63%	7.19 ± 0.27%
Mean Teacher	15.87 ± 0.28%	5.65 ± 0.47%
VAT	13.86 ± 0.27%	5.63 ± 0.20%
VAT + EntMin	13.13 ± 0.39%	5.35 ± 0.19%
Pseudo-Label	17.78 ± 0.57%	7.62 ± 0.29%

Table 1. Top: Reported results in the literature; Bottom: Using our proposed unified reimplementation (Wide ResNet WRN-28-2). The model below the line has roughly half as many parameters.

Our Findings

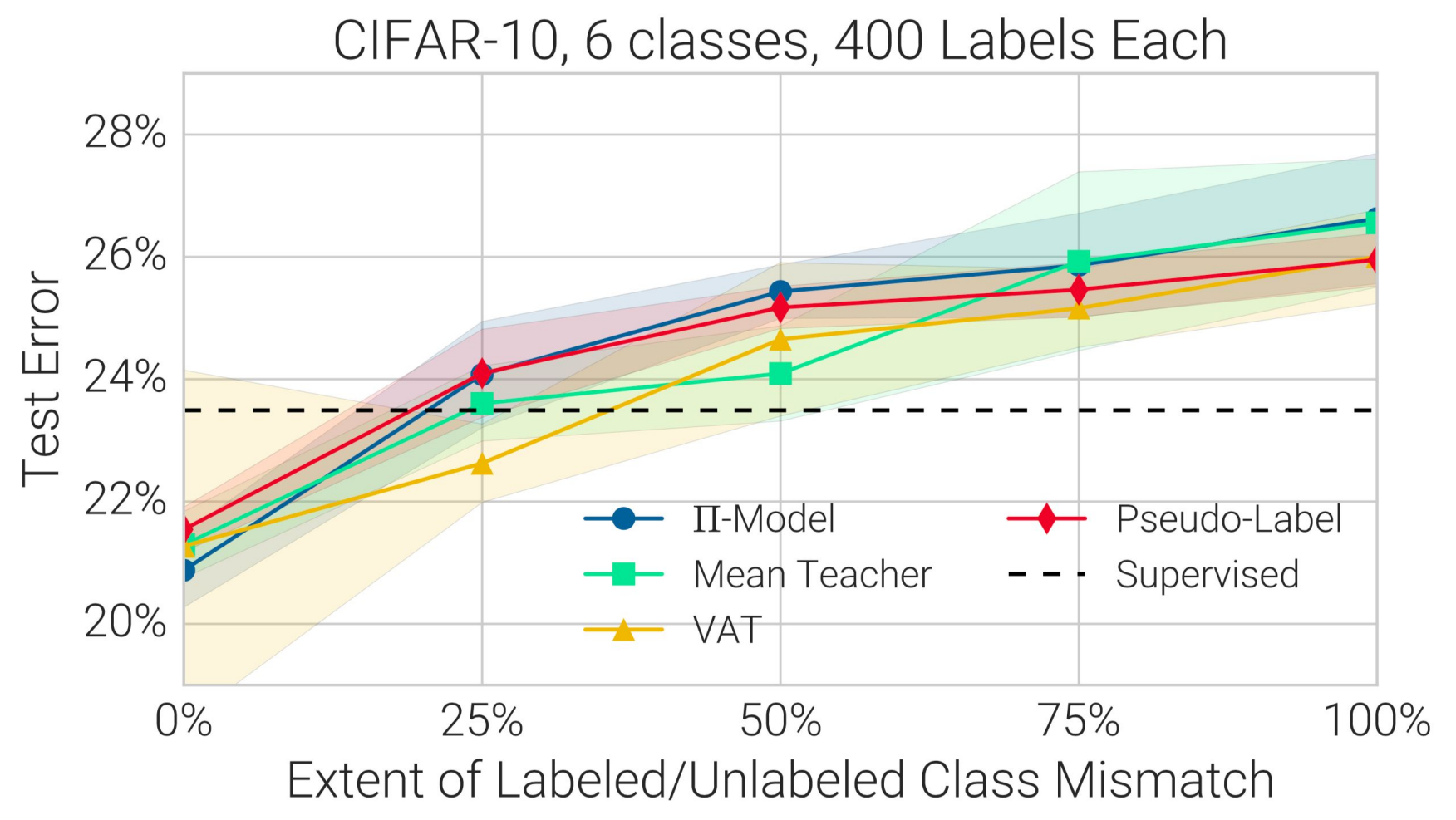
SSL gains smaller than reported

Papers report surprisingly poor supervised-only baselines. **When hyperparameters are tuned with the same budget, the gains from using unlabeled data shrink.**

Method	CIFAR-10 4000 Labels	SVHN 1000 Labels
Π-Model [32]	34.85% → 12.36%	19.30% → 4.80%
Π-Model [46]	13.60% → 11.29%	–
Π-Model (ours)	20.26% → 16.37%	12.83% → 7.19%
Mean Teacher [50]	20.66% → 12.31%	12.32% → 3.95%
Mean Teacher (ours)	20.26% → 15.87%	12.83% → 5.65%

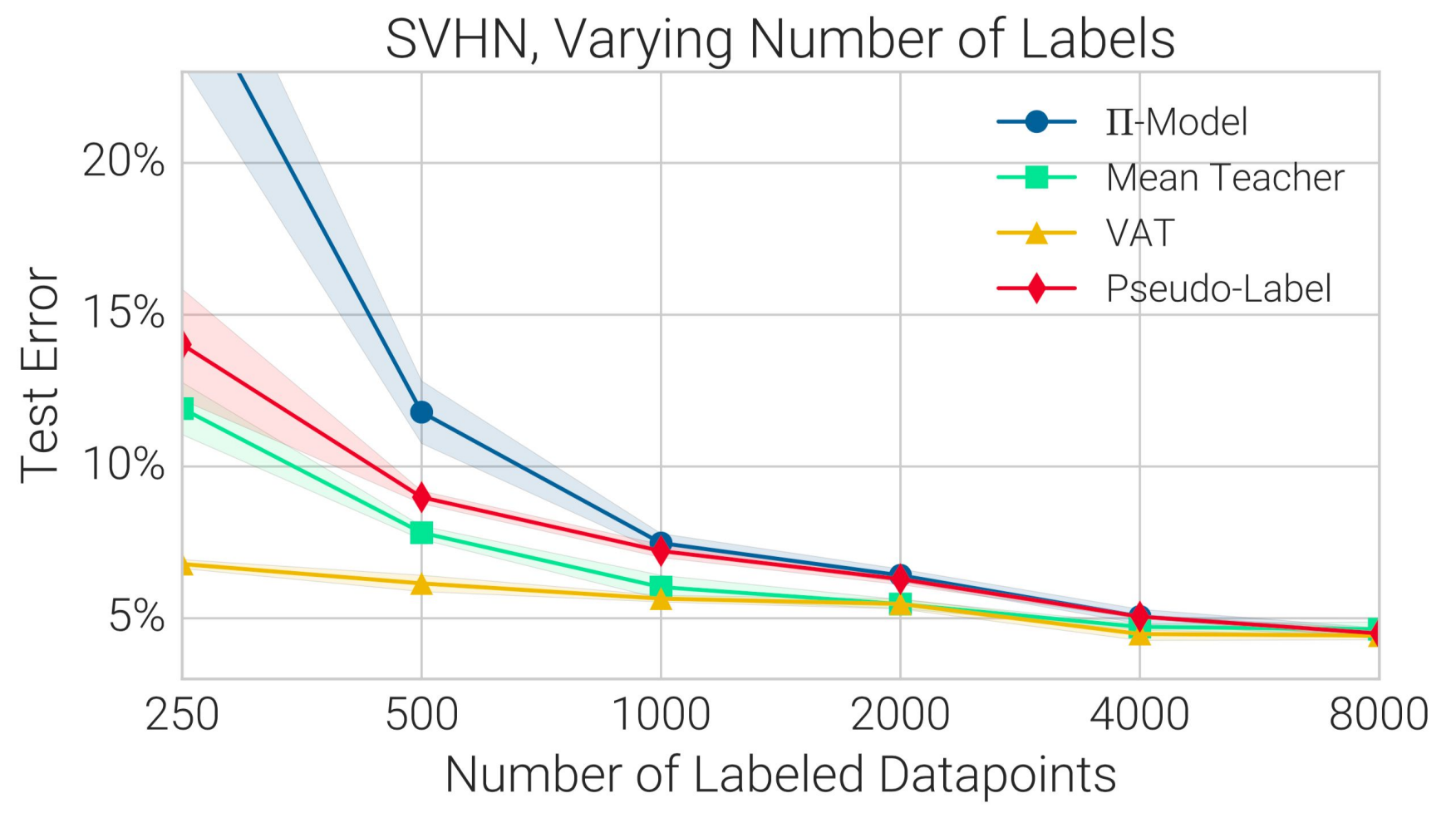
Unlabeled data can hurt the model

We train CIFAR-10 with labeled images of only animal classes, and unlabeled images of varying combinations of animal and non-animal classes. **Adding unlabeled from mismatched classes can hurt a model, compared to using only labeled data.**



Varying behavior with few labels

As the labeled dataset shrinks, different SSL techniques show varying ability to learn from unlabeled data.



Transfer learning can outperform SSL

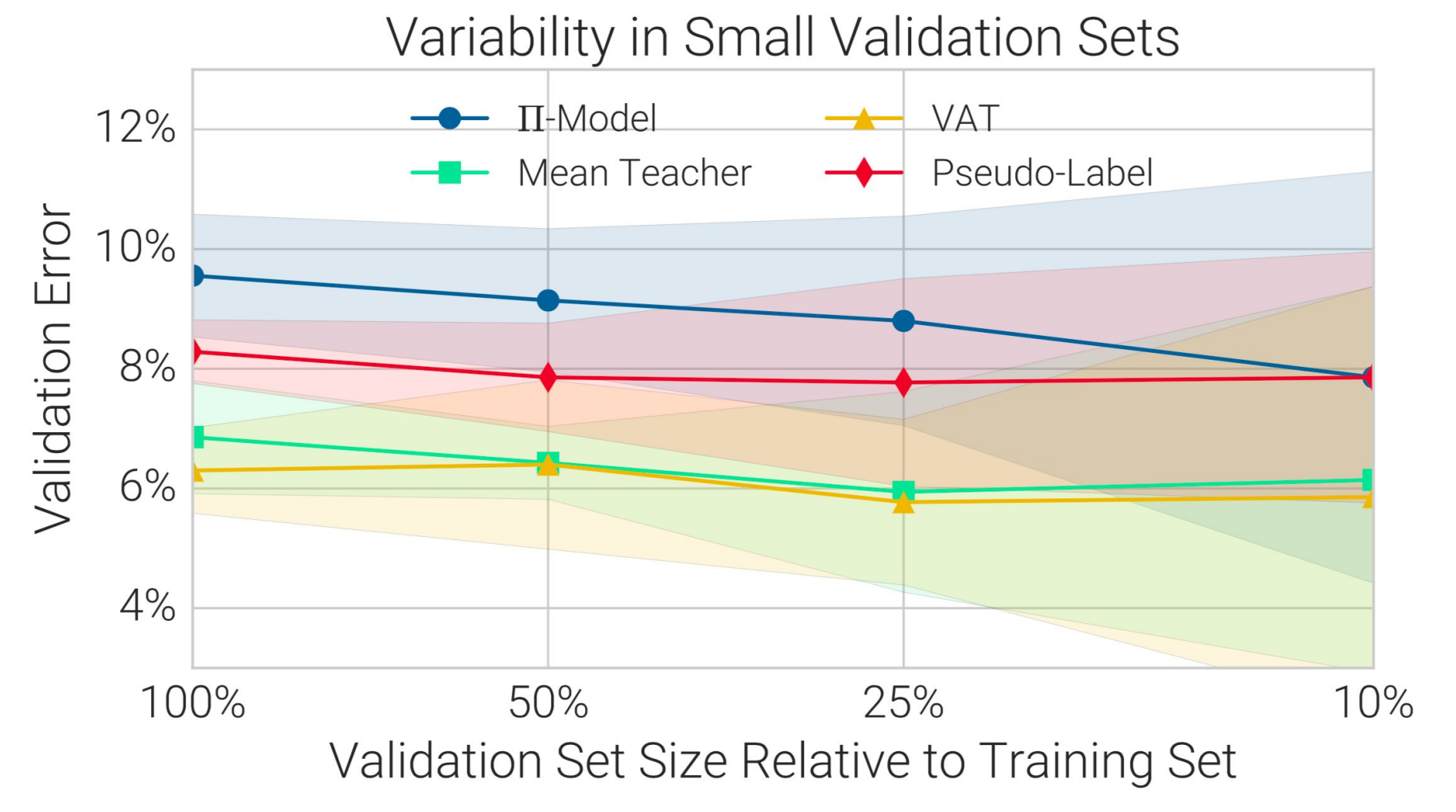
Training on 32x32 ImageNet then **fine-tuning** on CIFAR-10 with 4k labels **outperforms all of our SSL models.**

Method	CIFAR-10 4000 Labels
VAT with Entropy Minimization	13.13%
ImageNet → CIFAR-10	12.09%
ImageNet → CIFAR-10 (no overlap)	12.91%

Models are tuned on unrealistically large validation sets

Commonly, SSL researchers tune hyperparameters on a validation set larger than the labeled training set.

With a realistically sized validation set, error bars on accuracy are larger than differences, hence **good model selection may be infeasible.**



Code available!

```
https://github.com/brain-research/realistic-ssl-evaluation
```