Twitter Thread by Lucas Beyer

Lucas Beyer
@giffmana

So you think you know distillation; it's easy, right?

We thought so too with @XiaohuaZhai @__kolesnikov__ @_arohan_ and the amazing @royaleerieme and Larisa Markeeva.

Until we didn't. But now we do again. Hop on for a ride (+the best ever ResNet50?)

■■https://t.co/3SlkXVZcG3

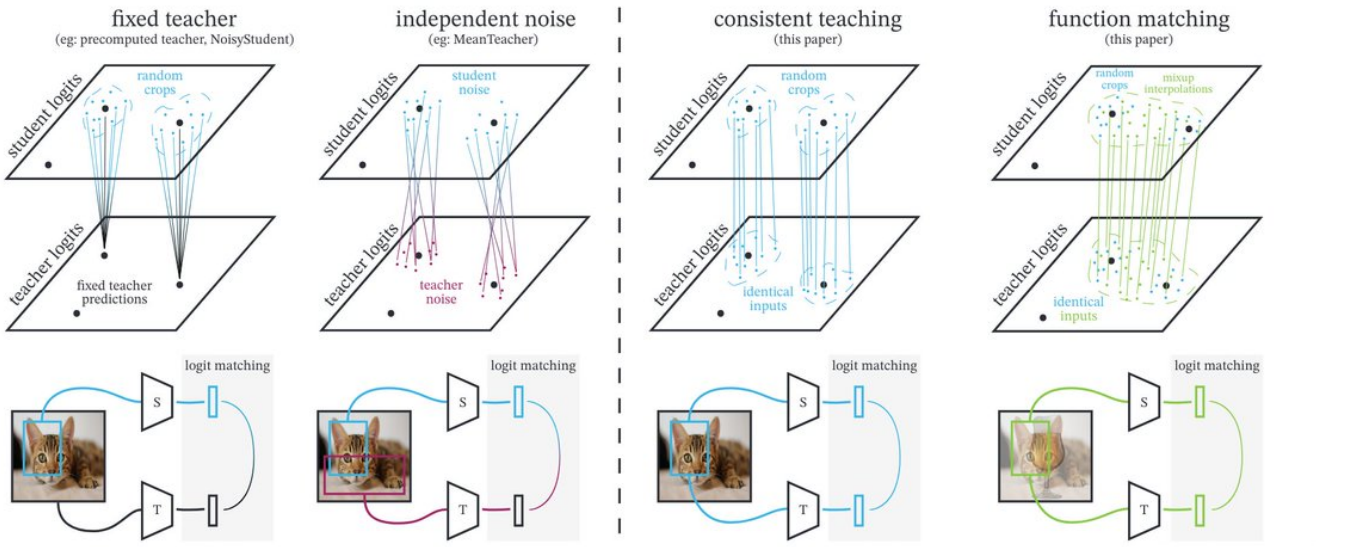


Figure 1: Schematic illustrations of various design choices when doing knowledge distillation. **Left**: *Teacher* receives a fixed image, while *student* receives a random augmentation. **Center-left**: *Teacher* and *student* receive independent image augmentations. **Center-right**: *Teacher* and *student* receive consistent image augmentations. **Right**: *Teacher* and *student* receive consistent image augmentations plus the input image manifold is extended by including linear segments between pairs of images (known as *mixup* [50] augmentation).

This is not a fancy novel method. It's plain old distillation.

But we investigate it thoroughly, for model compression, via the lens of *function matching*.

We highligh two crucial principles that are often missed: Consistency and Patience. Only both jointly give good results!

0. Intuition: Want the student to replicate _the whole function_ represented by the teacher, everywhere that we expect data in input space.

This is a much stronger view than the commonly used "teacher generates better/more informative labels for the data". See pic above.

1. Consistency: to achieve this, teacher and student need to see the same view (crop) of the image. For example, this means no pre-computed teacher logits! We can generate many more views via mixup.

Other approaches may look good early, but eventually fall behind consistency.

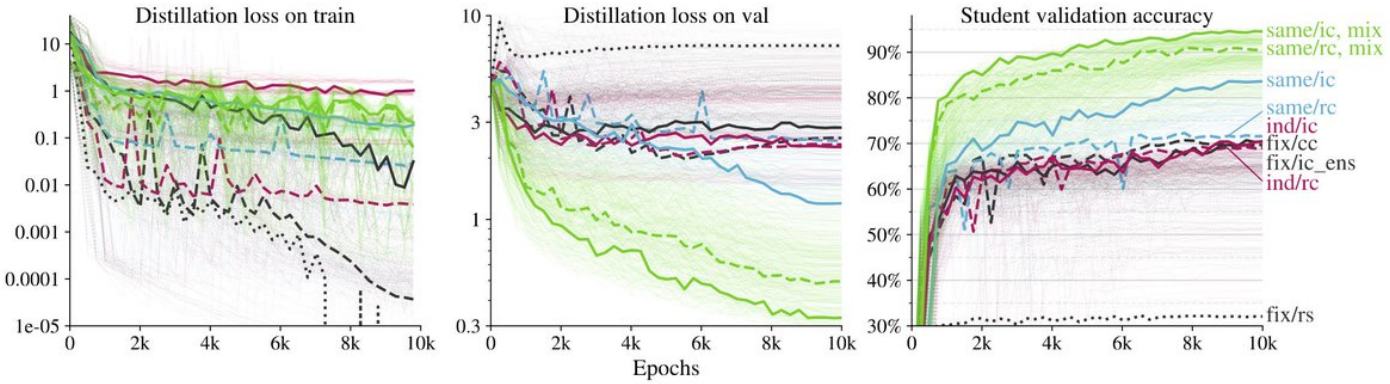
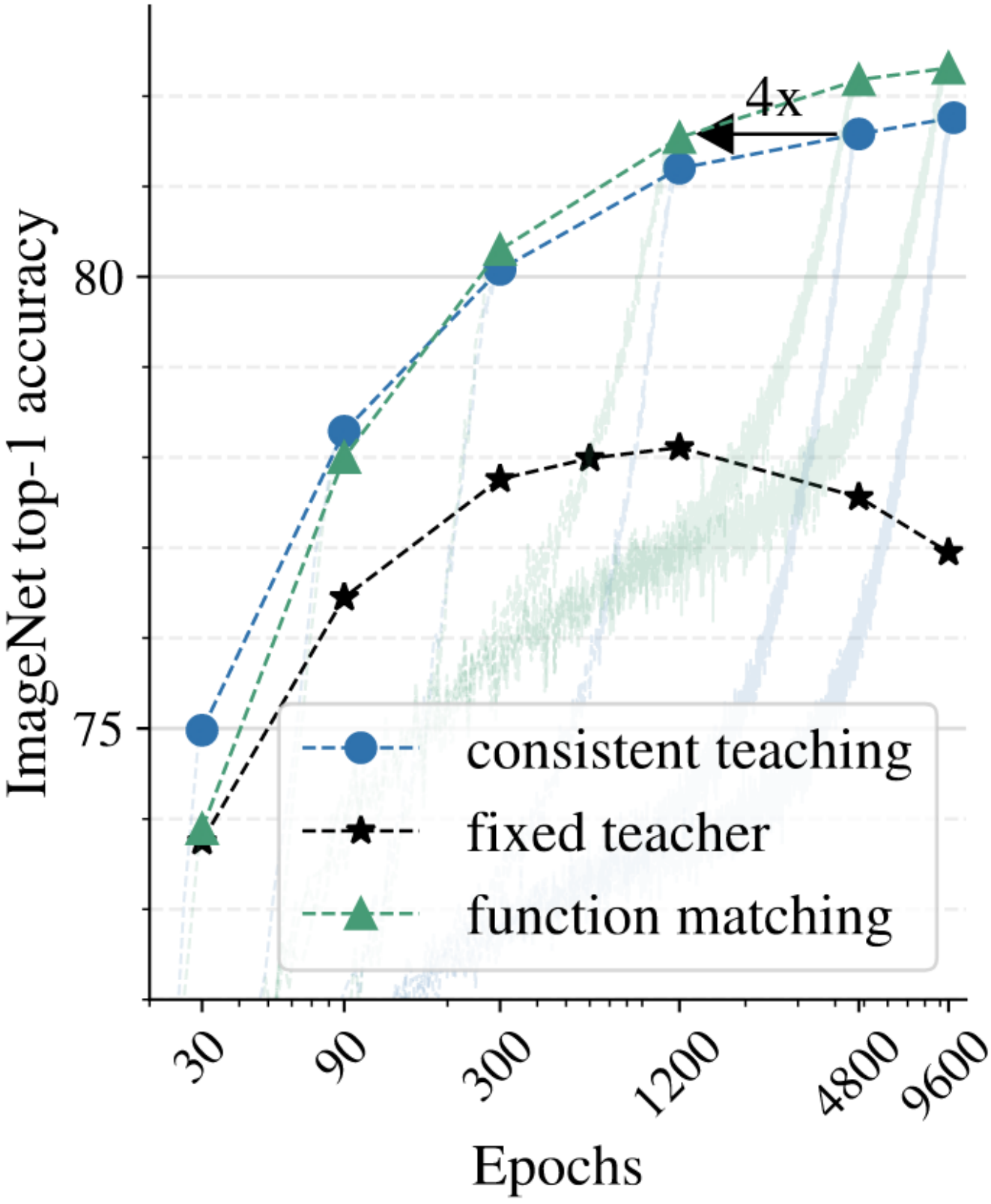


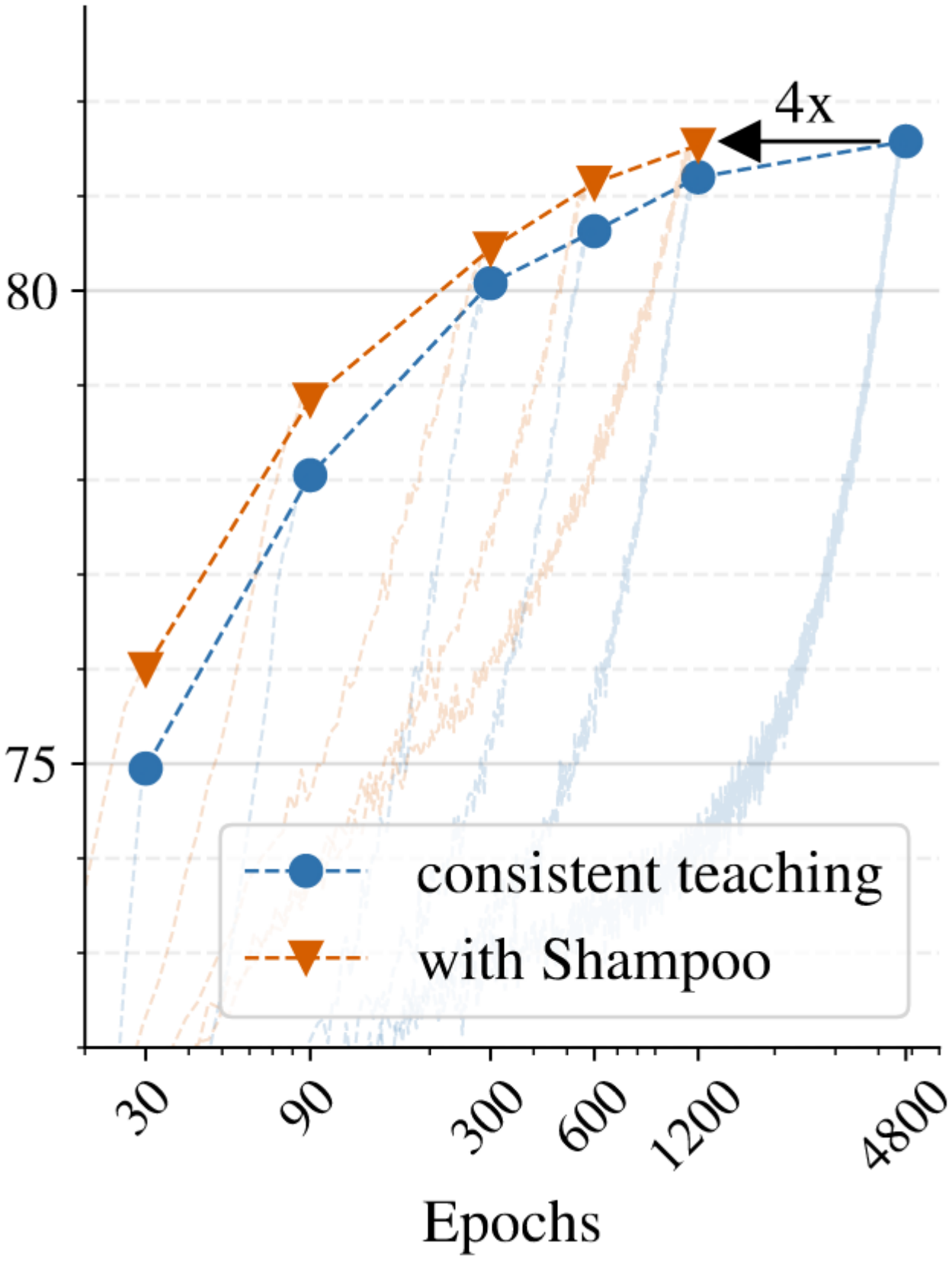
Figure 2: Experimental validation of the "consistency" requirement on the *Flowers102* dataset. Colors match different knowledge distillation design choices as introduced in Figure 1 and Section 3.1.1. Note that while the fixed teacher settings achieve significantly lower distillation loss, they lead to students which do not generalize well. In contrast, consistent teaching and function matching approaches lead to significantly higher student performance. Similar results on more datasets are reported in Appendix C.

2. Patience: The function matching task is HARD! We need to train *a lot* longer than typical, and actually we were not able to reach saturation yet. Overfitting does not happen, as when function-matching, an "overfit" student is great! (Note: w/ pre-computed teacher, we overfit)



2b. Excessively long training may mean optim struggle. We try advanced optimization via Shampoo, and get 4x faster convergence.

We believe this setting is a great test-bed for optimizer research: No concern of overfitting, and reducing training error means generalizing better!



3. By distilling a couple large BiT R152x2 models into a ResNet-50, we get a ResNet-50 on ImageNet that gets 82.8% at 224px resolution, and 80.5% at 160px! ■

No "tricks" just plain distillation, patiently matching functions.

Table 2: Comparison of our best and literature ResNet-50 models. The metric is accuracy on ImageNet test split (officially *val* split).

Model	Res.	Accuracy
Vanilla R50 [11]	224	77.2%
BiT-M-R50 [22]	224	78.4%
Meal-v2 [37]	224	80.7%
FunMatch (T384+224)	224	**82.8%**
"Revisiting" R50 [2]	160	78.8%
FunMatch (T224)	160	**80.5%**

4. Importantly, this simple strategy works on many datasets of various sizes, down to only 1020 training images, where anything else we tried overfit horribly.

Be patient, be consistent, that's it. Eventually, you'll reach or outperform your teacher!

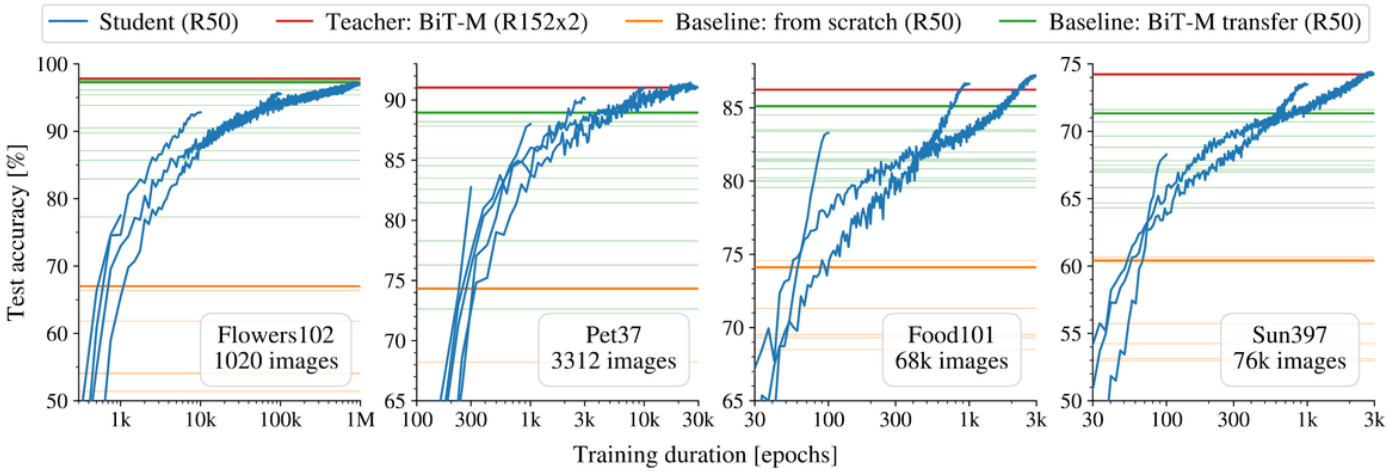
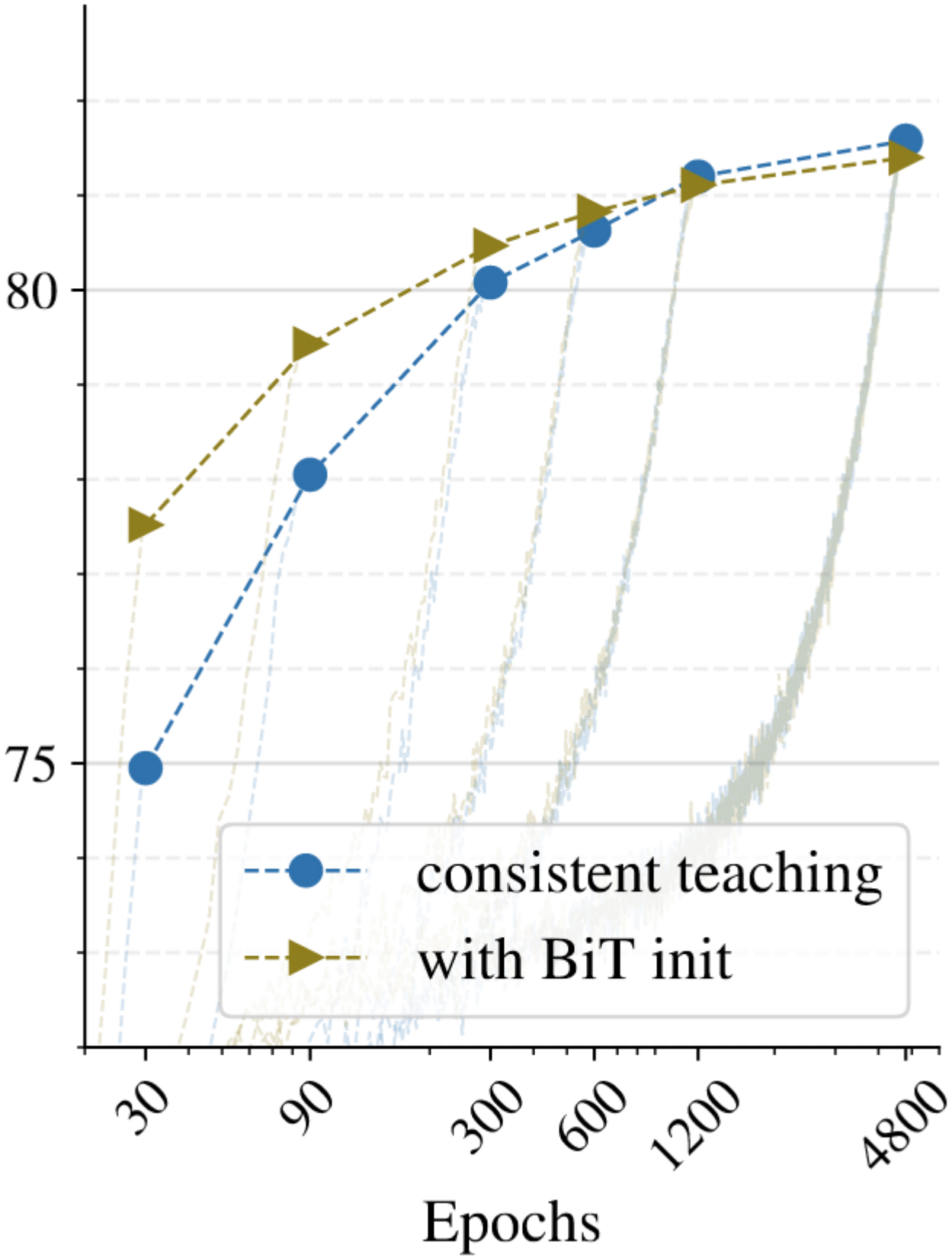


Figure 3: One needs patience along with consistency when doing distillation. Eventually, the teacher will be matched; this is true across various datasets of different scale.

2c. We can't stress patience enough. Multiple strategies, for example initializing the student with a pre-trained model shown here, look promising at first, but eventually plateau and are outperformed by patient, consistent function matching.



5. We have a lot more content. MobileNet students, distilling on on "random other" data (shown below), very thorough baselines, a teacher ensemble, and.... BiT download statistics!

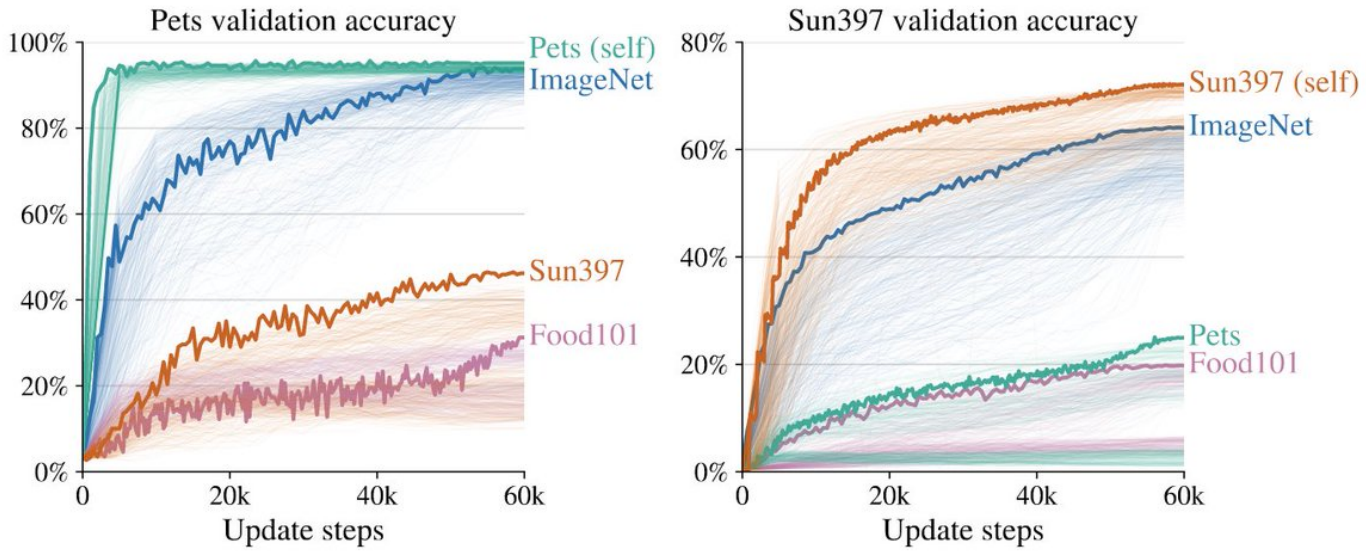


Figure 5: Distilling *pet* and *sun397* datasets on different data sources. Results indicate that distilling on completely unrelated images works to some extent, even though final results are relatively low. Distilling on "in-domain" data is the best and distilling on related/overlapping images can work reasonably well, but may require extra long training schedule.

PS: we are working on releasing a bunch of the models, including the best ones, ... but we're also on vacation. Watch https://t.co/Age8NXgS1D and stay tuned, we're aiming for next week!