🔥 LIT: 🔥

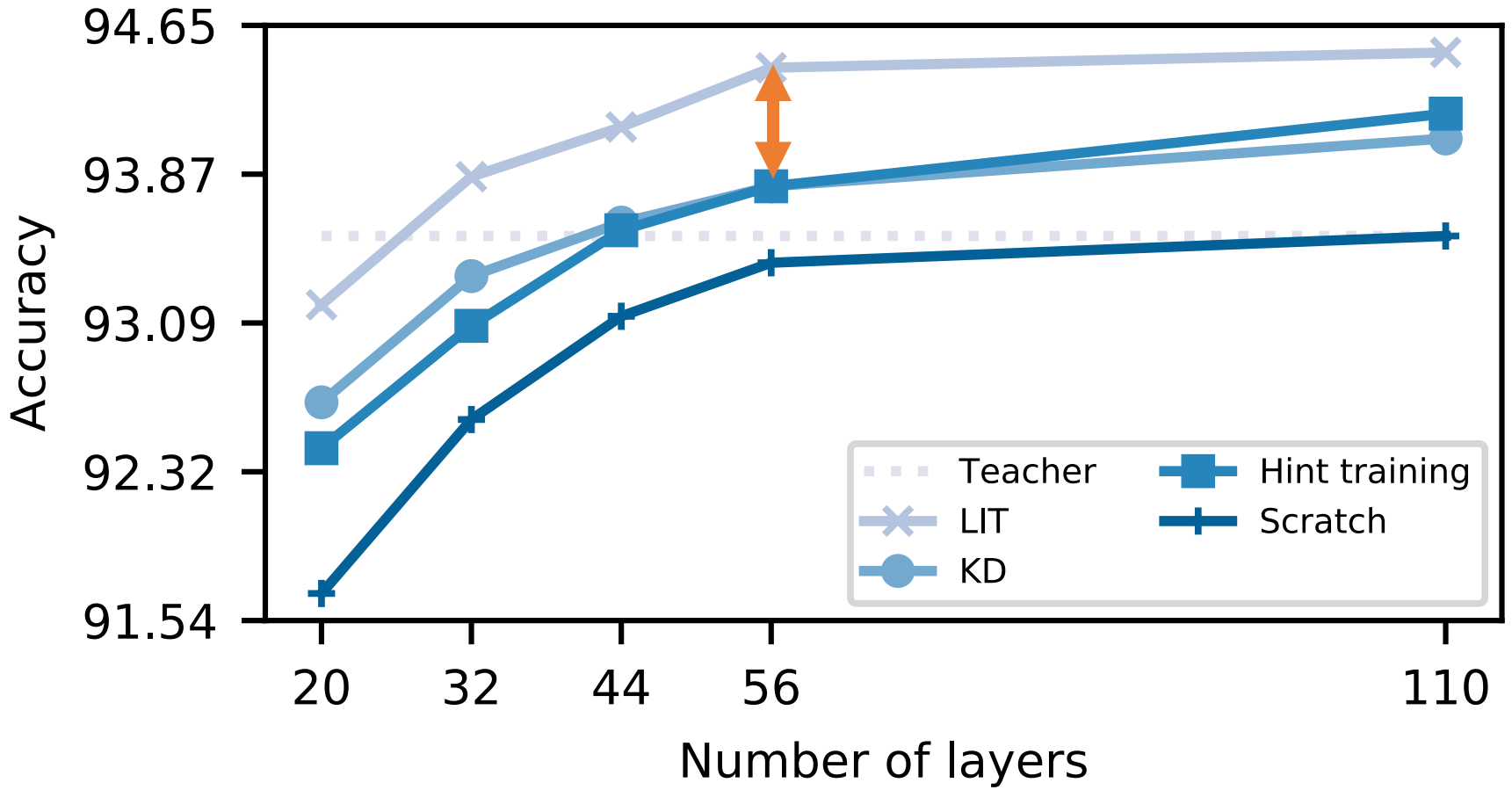
Learned Intermediate representation Training for Model Compression

Animesh Koratana*, **Daniel Kang***, Peter Bailis, Matei Zaharia

DAWN Project, Stanford InfoLab

http://dawn.cs.stanford.edu/

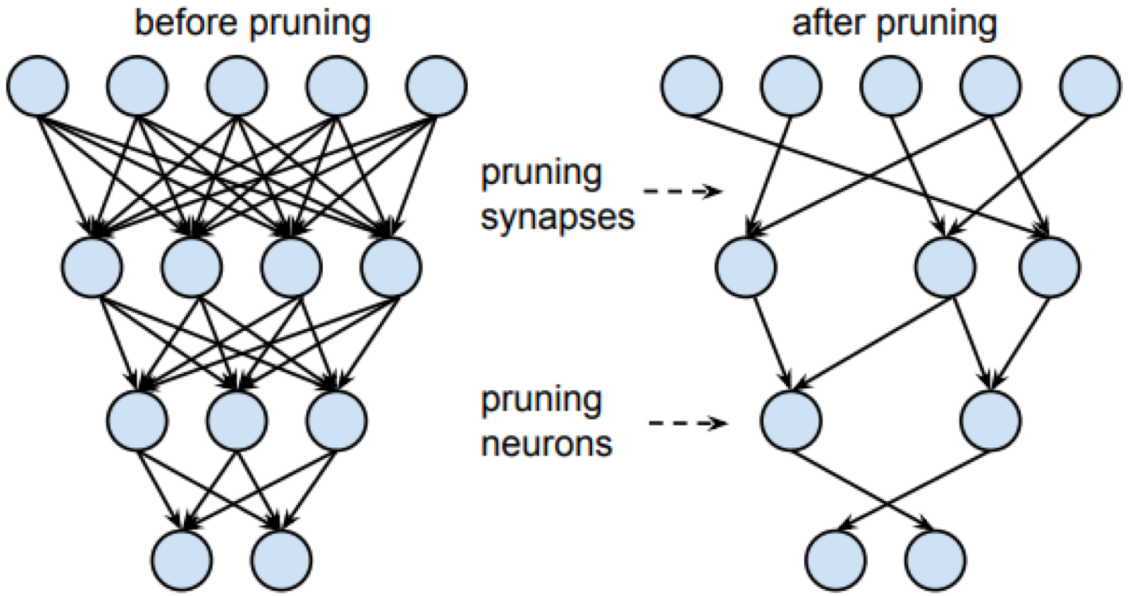
LIT can compress models up to 4x on CIFAR10: ResNet -> ResNet



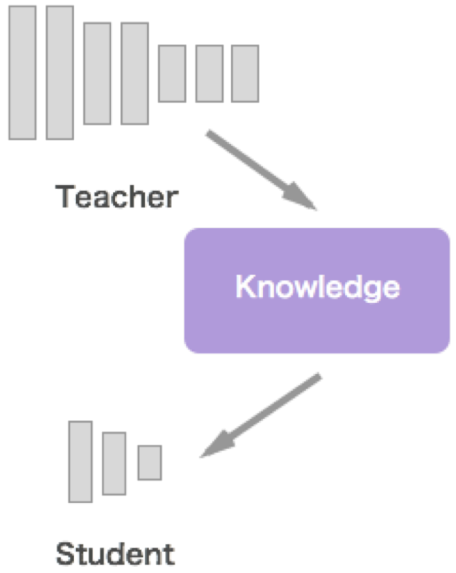
This talk: achieving higher compression on modern deep networks

Deep networks can be compressed to reduce inference costs

e.g., deep compression, knowledge distillation, FitNets, …



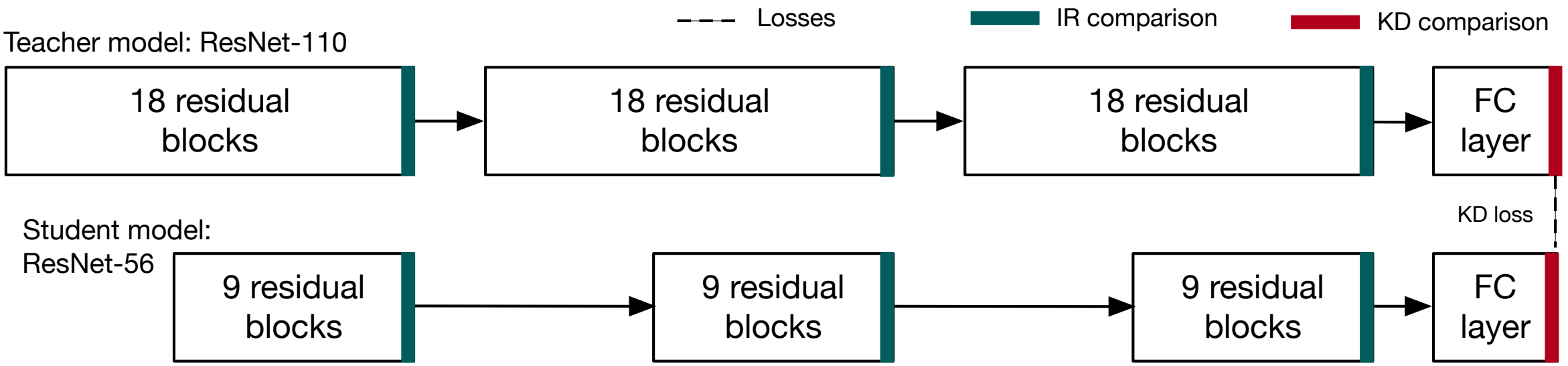
Deep compression



Knowledge distillation

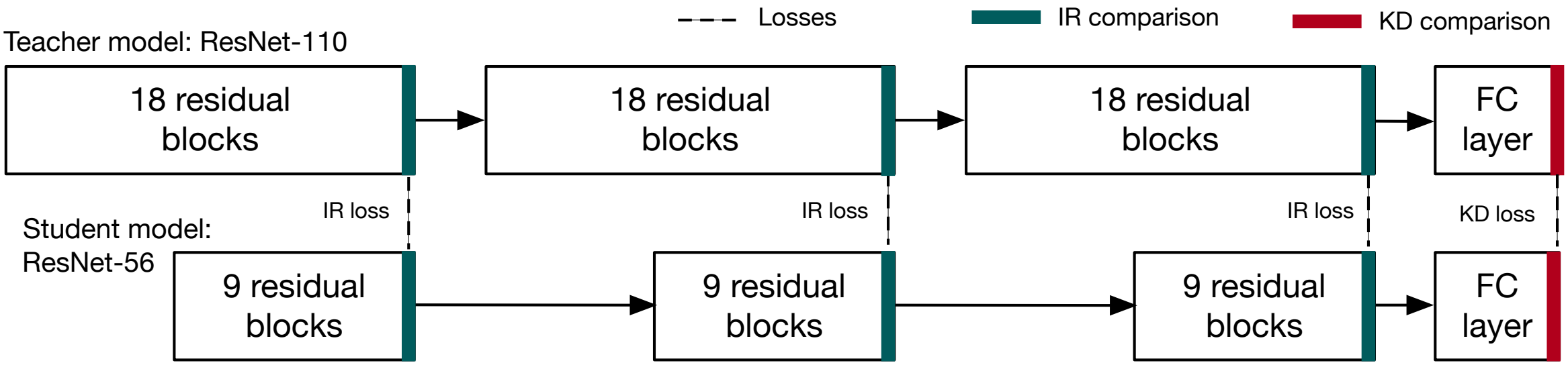
These methods are largely architecture agnostic

LIT: Learned Intermediate-representation Training for modern, very deep networks



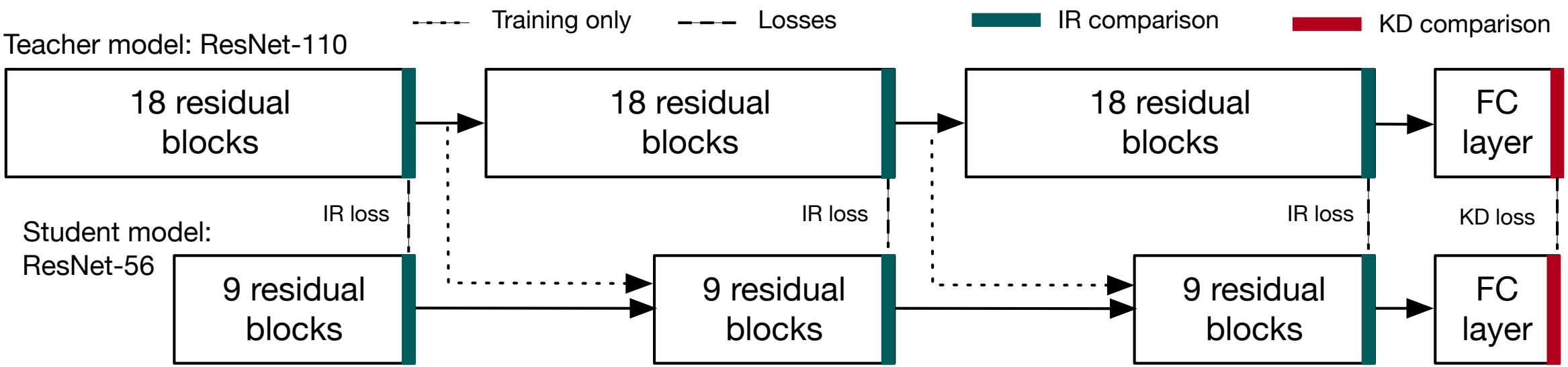
Modern networks have highly repetitive sections – can we compress them?

LIT: Learned Intermediate-representation Training for modern, very deep networks



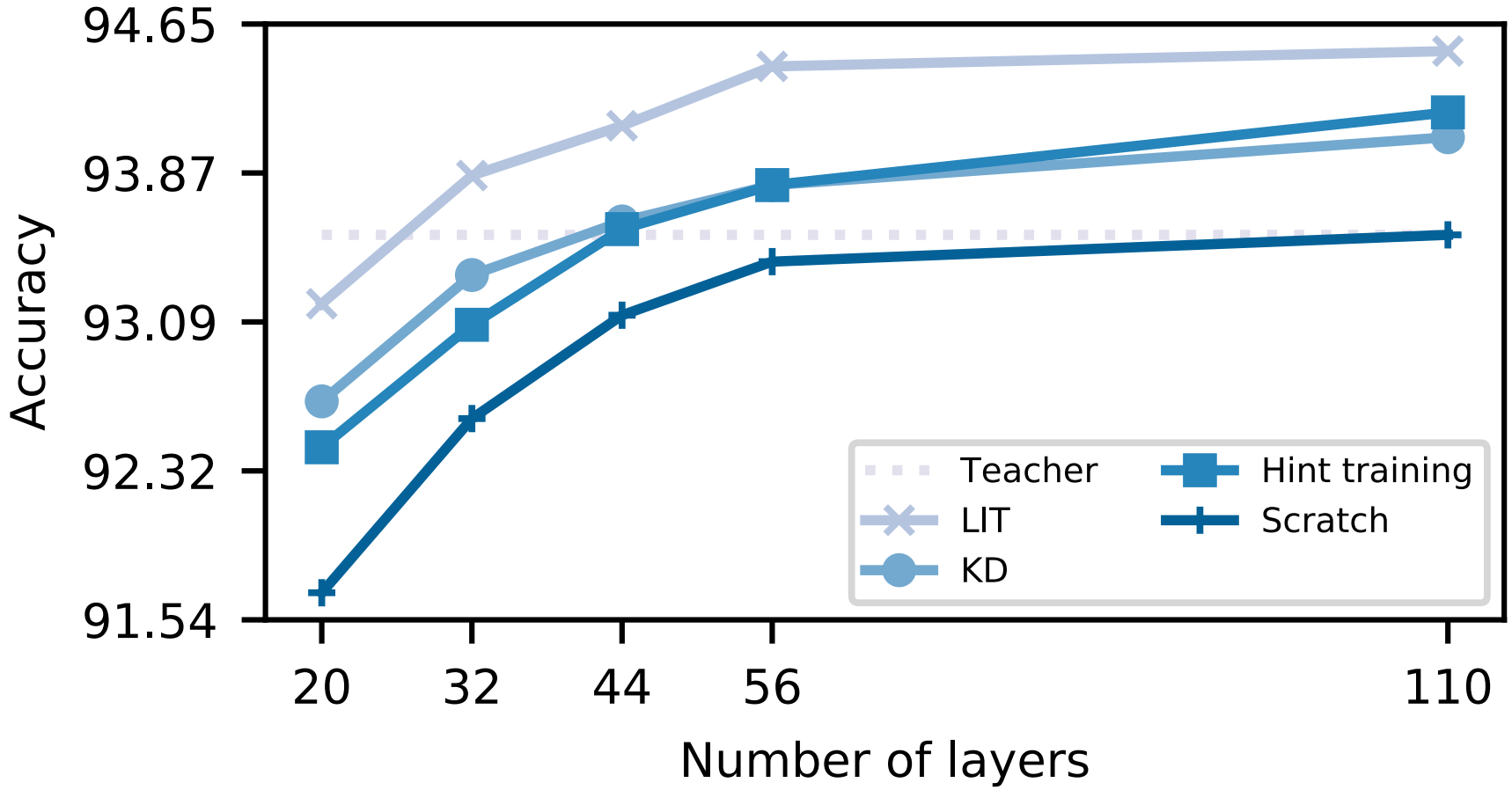
LIT penalizes **deviations in intermediate representations** of architectures with the same width

LIT: Learned Intermediate-representation Training for modern, very deep networks



LIT uses the **output** of the teacher model's **previous section** as input to the student model's **current section**

LIT can compress models up to 4x on CIFAR10: ResNet -> ResNet

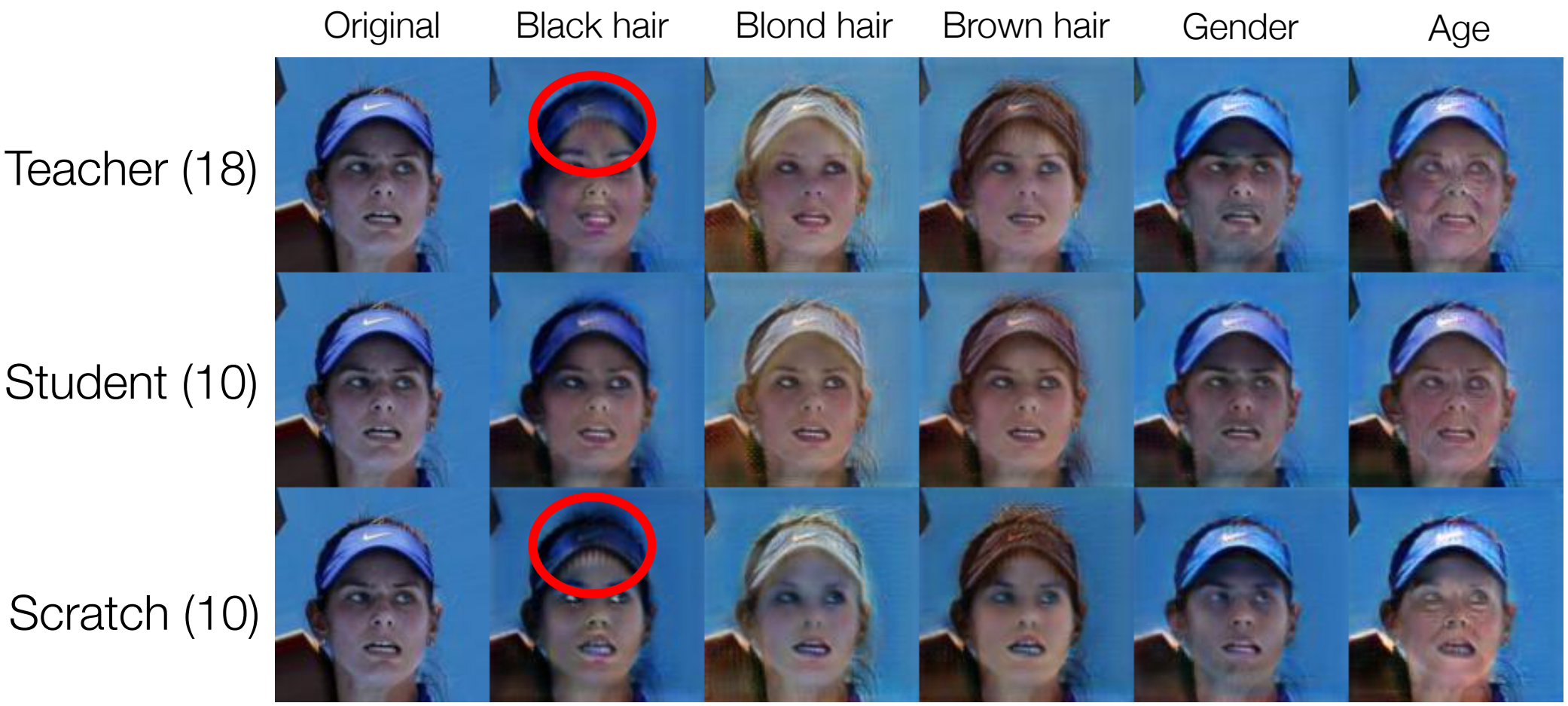


LIT can compress StarGAN up to 1.8x

Model	Inception score (higher is better)	FID score (lower is better)
Teacher (18)	3.49	6.43
LIT student (10)	**3.56**	**5.84**
L2 student (10)	3.46	6.47
From scratch (10)	3.37	6.56
Rand init (10)	2.63	94.00
Rand init (18)	2.45	151.43

Student model outperforms teacher in Inception/FID score

LIT can compress GANs up to 1.8x



Student model also outperforms teacher in qualitative evaluation

Conclusions

Neural networks are becoming more expensive to deploy

LIT is a novel technique that combines both:

1. Intermediate representations and
2. matching outputs

that improves training to give 3-5x compression for many tasks

ddkang@stanford.edu
koratana@stanford.edu

Find our poster at
Pacific Ballroom, #17!