Date: 4th Jan. 2023

Towards understanding knowledge distillation

Presenter: Yong-Min Shin
jordan3414@yonsei.ac.kr / yongminshin.simple.ink

Cho & Hariharan, On the efficacy of knowledge distillation, CVPR 2020
Stanton et al., Does knowledge distillation really work?, NeurIPS 2021

00 Structure of the talk

1 Cho & Hariharan, 2020

- Analysis mainly based on model capacity
- First paper to investigate knowledge distillation itself

2 Stanton et al., 2021

- Differentiation of 'fidelity' and 'generalization'
- Mixed conclusion for the efficacy of knowledge distillation

3 Ojha et al., 2022

- Focus on distillation of teacher's properties other than performance
- Most recent paper, paper is written in a manner that the reader is easy to follow

1. Question regarding distillation + Hypothesis building → **2. Design experiments that can either reject / accept hypothesis** → **3. Observation of results & Discussion to gain insight**

Knowledge distillation: Towards more powerful and smaller models

- *Idea of compressing a larger capacity & high performing model into a smaller one (Bucilă et al., 2006)*
- *"Distilling" knowledge via transferring the output probability of the teacher network was popularized by (Hinton et al., 2015)*

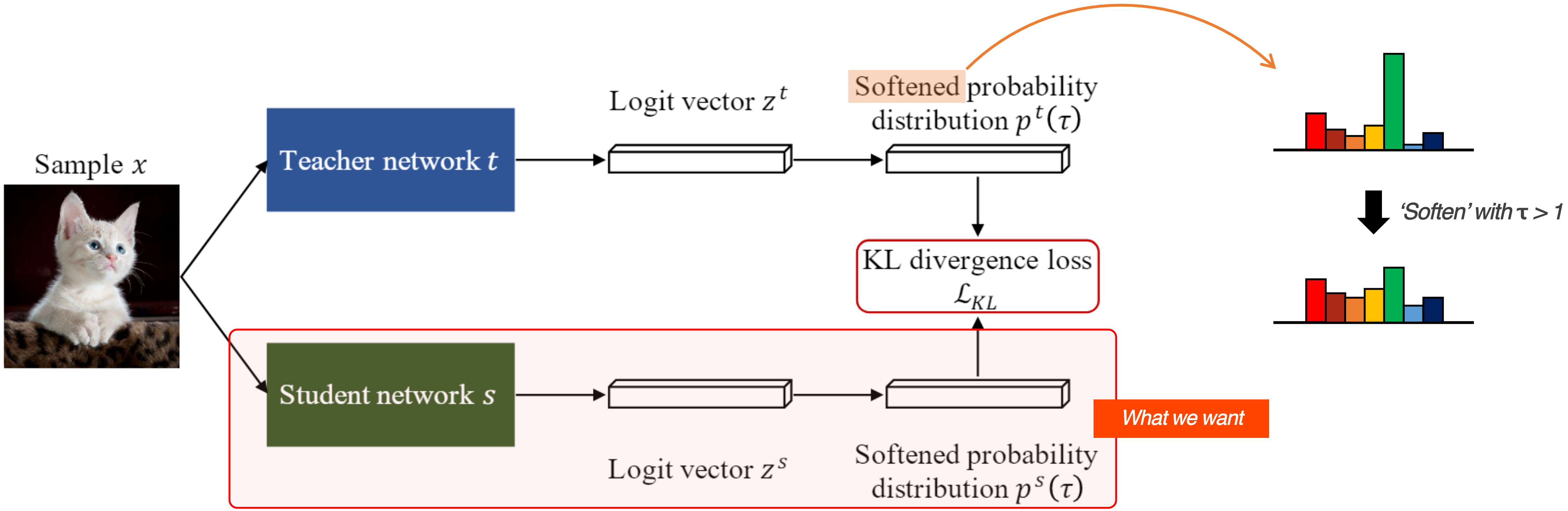


Image from (Kim et al., 2021)

$$\mathcal{L} = \alpha \mathcal{L}_{\mathrm{CE}} + (1-\alpha)\tau^2 \mathcal{L}_{\mathrm{KD}}$$

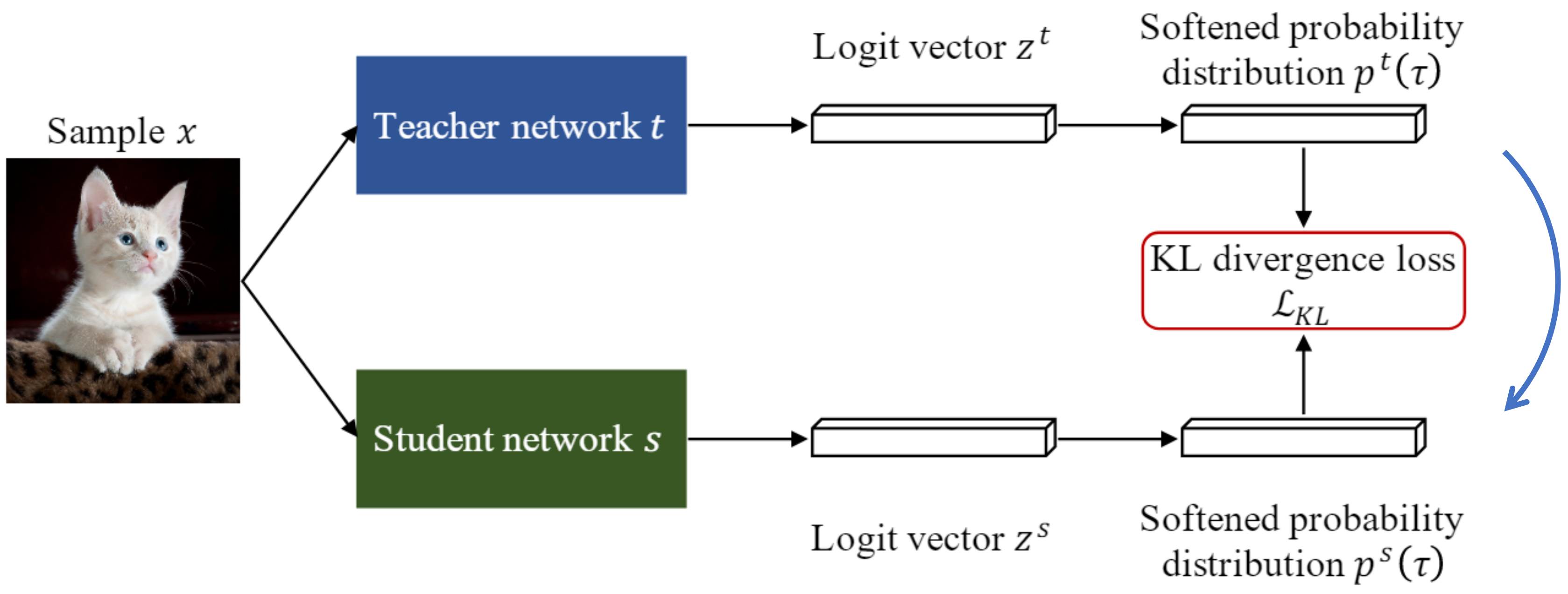
**Popular choices for τ: 3,4,5 / α: 0.9*

Bucilă et al. Model compression, KDD 2006
Hinton et al., Distilling the knowledge in a neural network, arXiv 2015
Kim et al., Comparing Kullback-Leibler divergence and mean squared error loss in knowledge distillation, IJCAI 2021

00 Preliminaries

Knowledge distillation: "Dark knowledge"

- *It is usually thought that aside from the teacher's predictions, it also distills "dark knowledge" to the student.*
- *Common question: What exactly is this "dark knowledge"?*



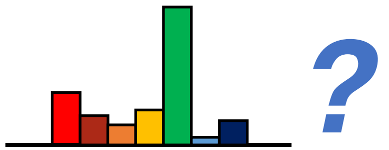


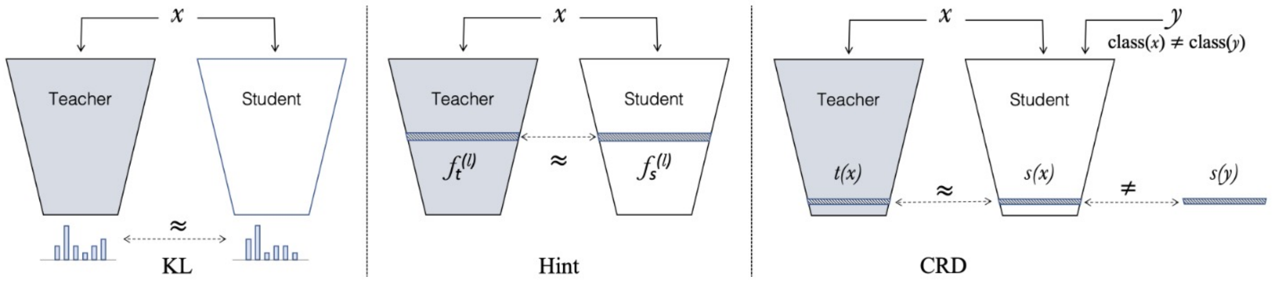
Image from (Kim et al., 2021)

Bucilă et al. Model compression, KDD 2006
Hinton et al., Distilling the knowledge in a neural network, arXiv 2015
Kim et al., Comparing Kullback-Leibler divergence and mean squared error loss in knowledge distillation, IJCAI 2021

00 Preliminaries

Setup: Knowledge distillation in computer vision

- *The papers are within the domain of computer vison*
- *Hence, the discoveries may be confined withing CV, and may not hold in other data types (e.g., graphs)*
- *Widely used datasets & models are investigated (e.g., ResNet + ImageNet)*
- *Usually focused on original KD ('KL', Hinton et al., 2015)*



1 Cho & Hariharan

- Dataset: CIRAR10, ImageNet
- Models: ResNet, WideResNet (WRN), DenseNet
- Methodology: KL

2 Stanton et al.

- Dataset: MNIST, EMNIST, CIFAR100
- Models: LeNet, ResNet, VGG (appendix)
- Methodology: KL

3 Ojha et al.

- Dataset: MNIST, ImageNet, (Geirhos et al., 2021)
- Models: ResNet, VGG, ViT, Swin
- Methodology: KL, *Hint, *CRD (See figure)

Gerihos et al., Partial success in closing the gap between human and machine vision, NeurIPS 2021
Hinton et al., Distilling the knowledge in a neural network, arXiv 2015

01 Cho & Hariharan, 2020

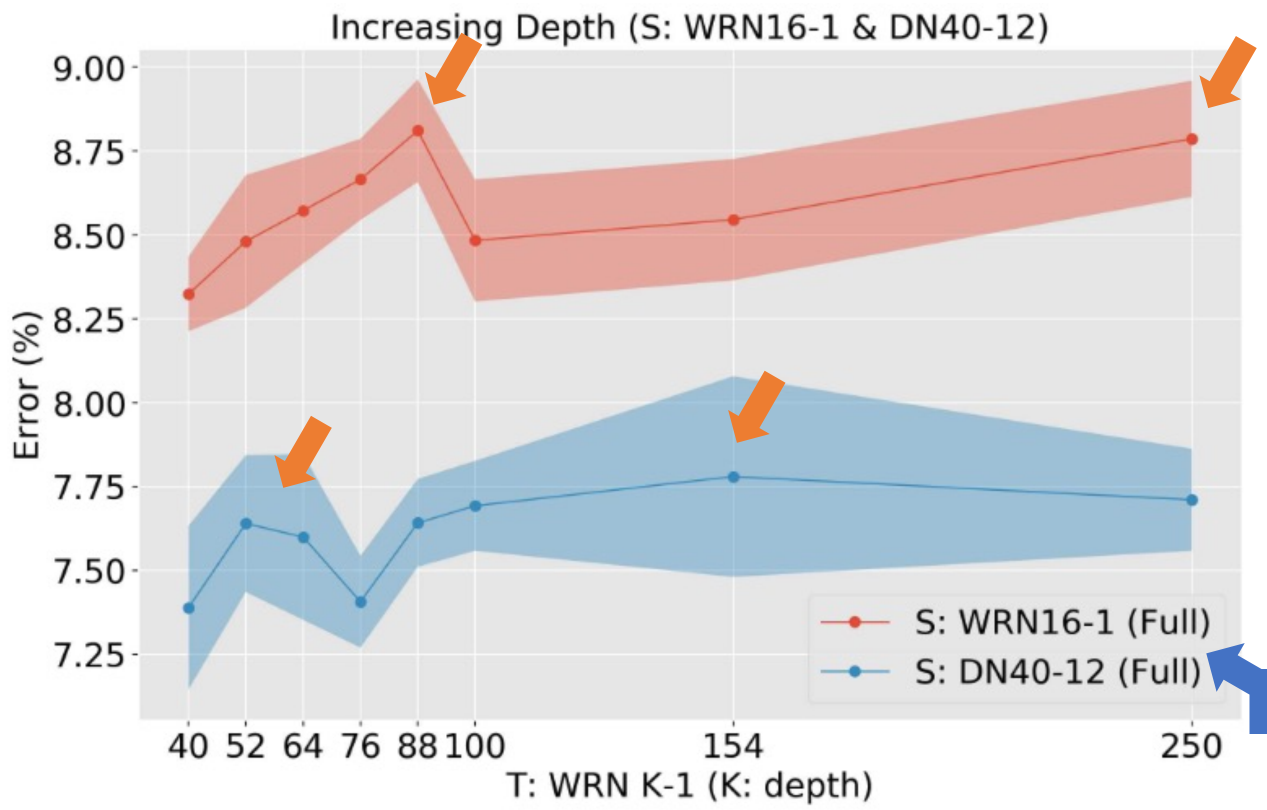
(1) Experiment with bigger teacher models

Common conception (Hypothesis)

Larger models → Better captures underlying class distribution → Provides better supervision during distillation

Experiment design

Observe: Student performance after distillation // Varying factor: Depth or width of teacher model → (Performance vs.



Different teacher architecture

- *Performance (error) vs. Depth plot*
- *The hypothesis is not true, it even gets less accurate*
- *Perhaps of overconfidence of teacher? → Softening does not help*
- *Leads to next experiment…*

01 Cho & Hariharan, 2020

(2) Experimenting capacity discrepancy

Hypothesis from last experiment
(1) Student can mimic teacher but does not translate to accuracy // (2) Student is unable to mimic teacher (capacity

Experiment design
Observe: Agreement ("KD error") between teacher and student // Varying factor: Depth or width of teacher model

Student	Teacher	KD Error (%,Train)	KD Error (%,Test)
WRN28-1	WRN28-3	0.23	4.05
	WRN28-4	0.25	4.53
	WRN28-6	0.23	4.54
	WRN28-8	0.31	4.81
WRN16-1	WRN16-3	1.70	6.32
	WRN16-4	1.69	6.52
	WRN16-6	1.94	6.91
	WRN16-8	1.69	7.01

Increasing width

Increasing width

KD error also increases

- *KD error does increase with bigger teacher model*
- *Therefore, it suggests that there is a capacity gap issue*

01 Cho & Hariharan, 2020

(3) Ineffectiveness of KD in ImageNet

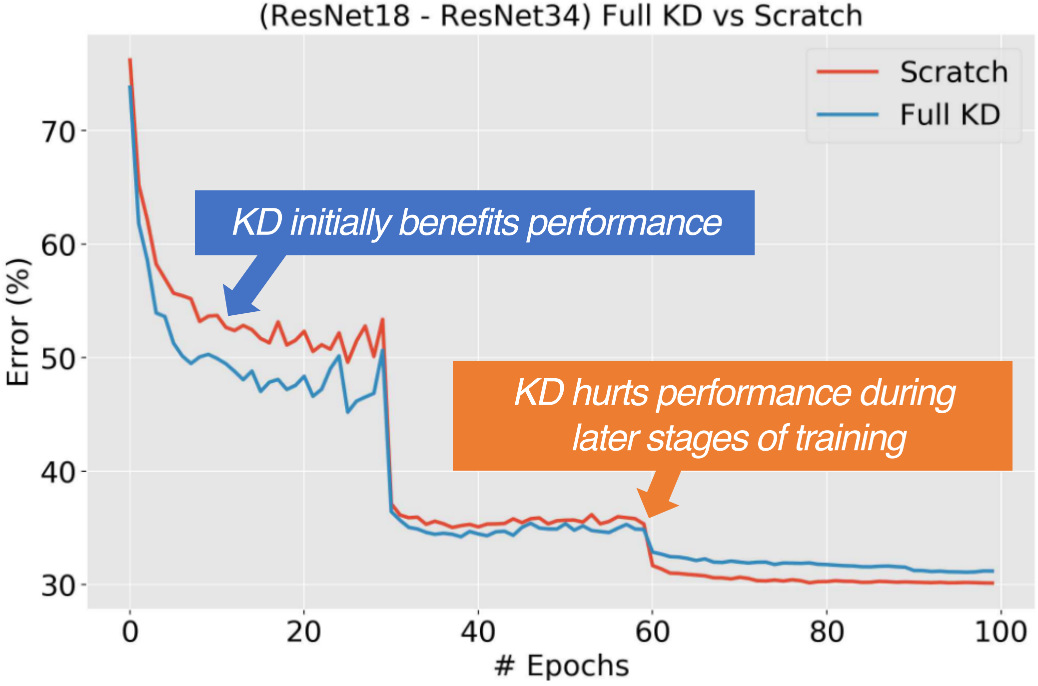
Observation

Teacher	Teacher Error (%)	Student Error (%)
-	-	30.24
ResNet18	30.24	30.57
ResNet34	26.70	30.79
ResNet50	23.85	30.95

Trained from scratch (No KD)

KD performs WORSE!

Further investigation



Conclusion from further investigation

1) Stop distillation early
2) Train with cross entropy loss only for the rest of the epochs

→ "ESKD" (Early-stopped knowledge distillation)

01 Cho & Hariharan, 2020

(4) Effectiveness of ESKD

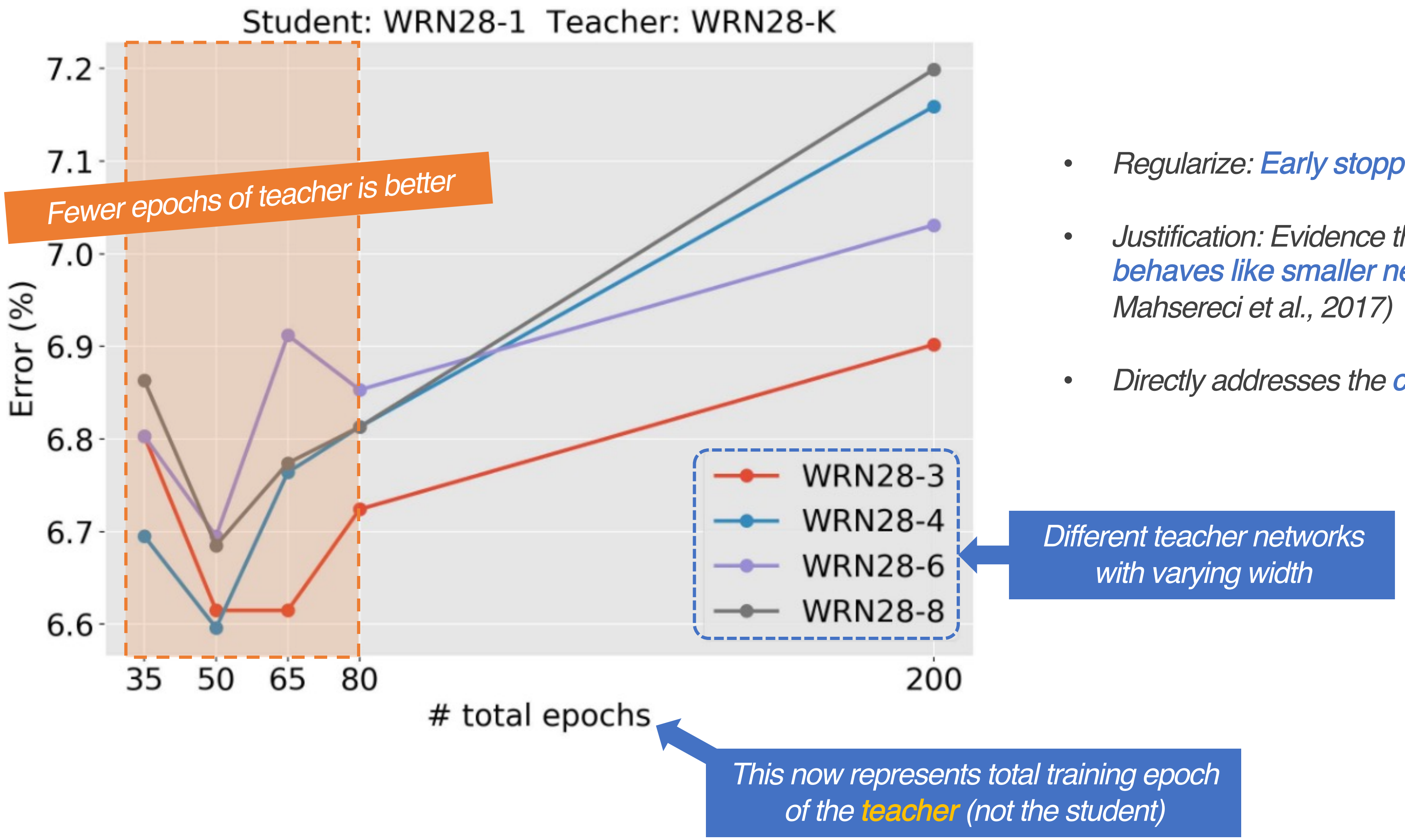
Teacher	Top-1 Error (%, Test)	CE (Train)	KD (Train)	KD (Test)
ResNet18	30.57	0.146	2.916	3.358
ResNet18 (ES KD)	29.01	0.123	2.234	2.491
ResNet34	30.79	0.145	1.357	1.503
ResNet34 (ES KD)	29.16	0.123	2.359	2.582
ResNet50	30.95	0.146	1.553	1.721
ResNet50 (ES KD)	29.35	0.124	2.659	2.940

Suggests: Student model was trading off cross-entropy loss & knowledge distillation loss.

However, this still does not solve the core problem of capacity discrepancy between teacher & student.

01 Cho & Hariharan, 2020

(5) Regularizing the teacher during training

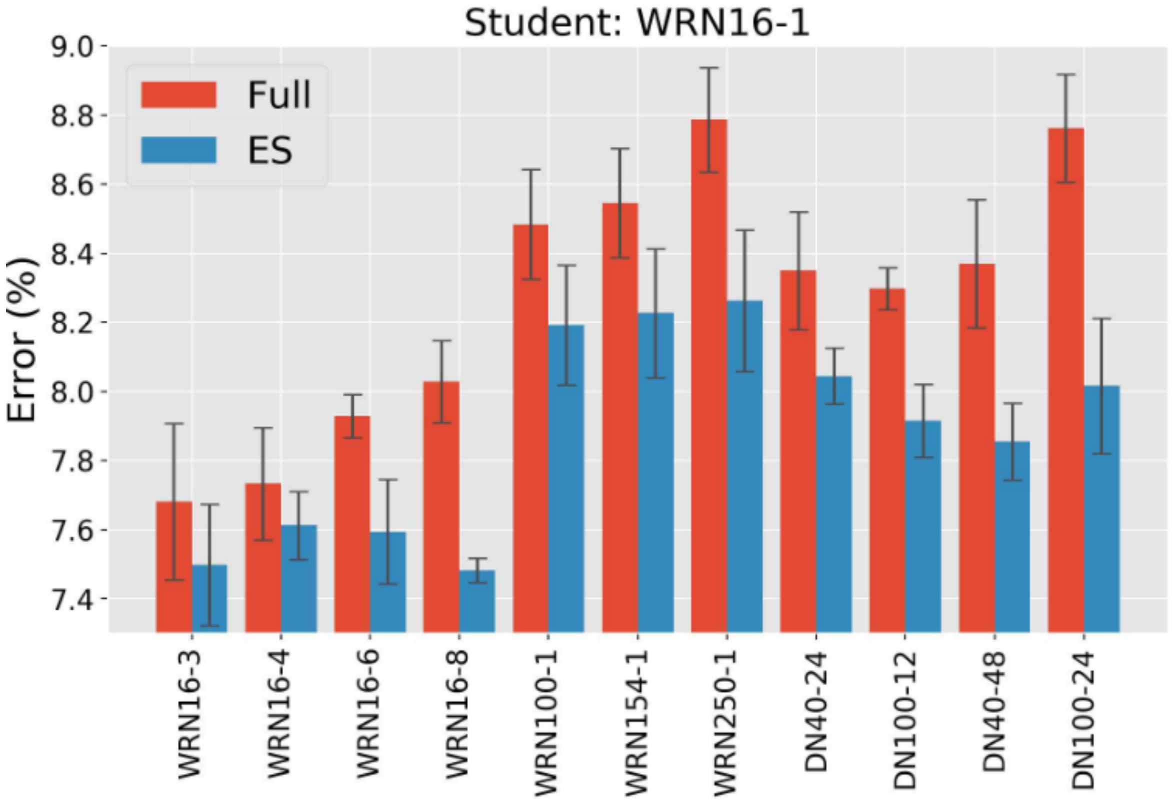


- *Regularize: Early stopping the teacher network also helps*
- *Justification: Evidence that larger model + few epochs behaves like smaller network (Caruana et al., 2001; Mahsereci et al., 2017)*
- *Directly addresses the capacity problem*

Caruana et al., Backpropagation, conjugate gradient, and early stopping, NeurIPS 2001
Mahsereci et al., Early stopping without a validation set, arXiv 2017

01 Cho & Hariharan, 2020

(6) Final conclusions



- *Overall, short distillation from early stopped teacher is recommended*
- *Early stopping acts as a strong regularization tool during distillation*

02 Stanton et al., 2021

(1) Fidelity & Generalization

- *Fidelity: Ability of a student to match the teacher's predictions*

1. *Average Top-1 Agreement*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\text{Teacher prediction of input } i = \text{Student prediction of input } i\}$$

2. *Average Predictive KL*

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}(\hat{p}_{\text{teacher}}(\mathbf{y}|\mathbf{x}_i)||\hat{p}_{\text{student}}(\mathbf{y}|\mathbf{x}_i))$$

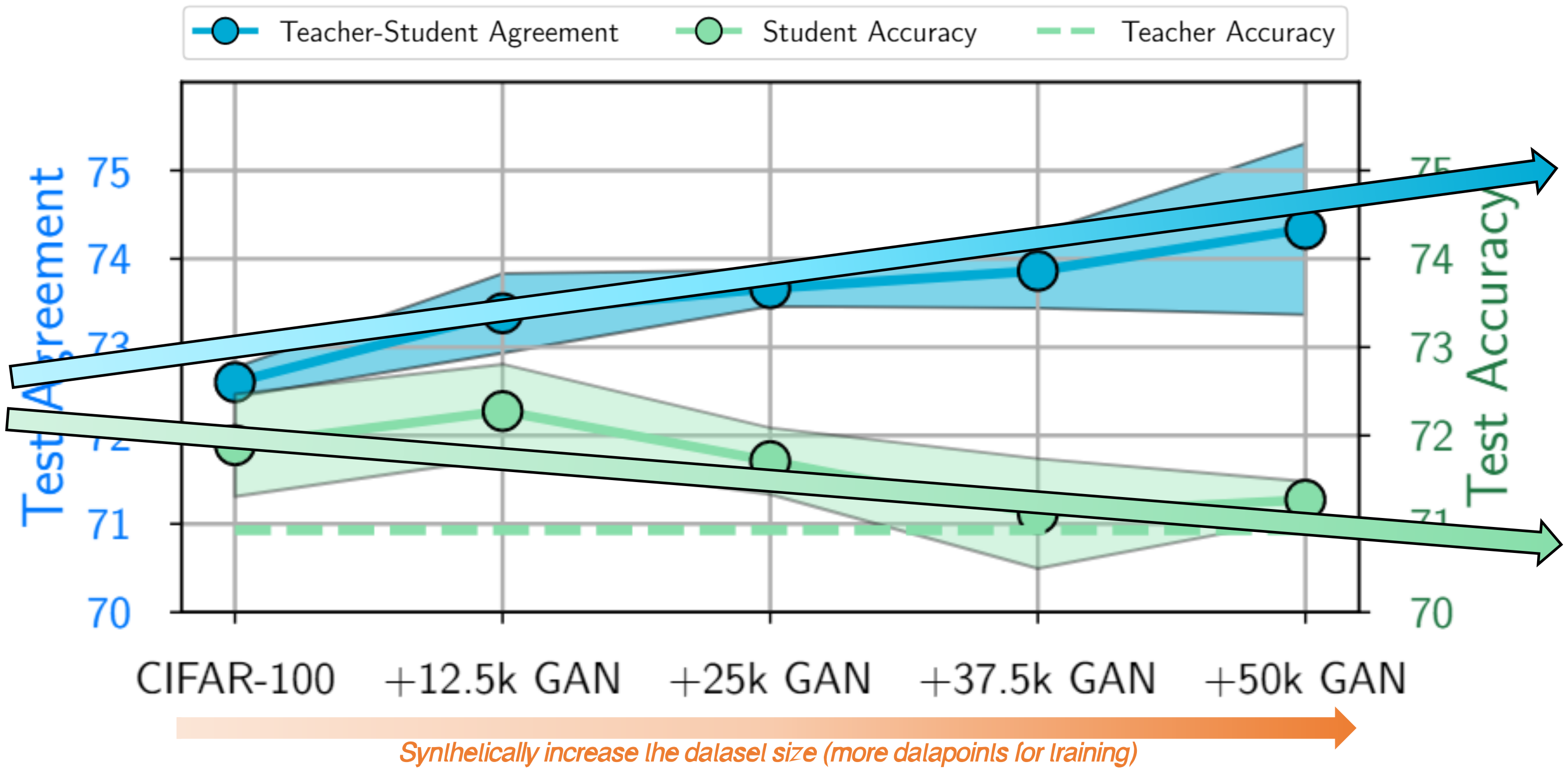
- *Generalization: Student's performance in unseen data*

02 Stanton et al., 2021

(2) Fidelity and generalization needs to be carefully addressed : Self-distillation

Common conception (Hypothesis)
Making the student to better mimic the teacher is desirable (Beyer et al., 2022)

Experiment design
Observe: Fidelity & Performance // Varying factor: Amount of dataset (Larger datasets will benefit fidelity)



Student better matches teacher (high fidelity) on more data

However, the performance decreases

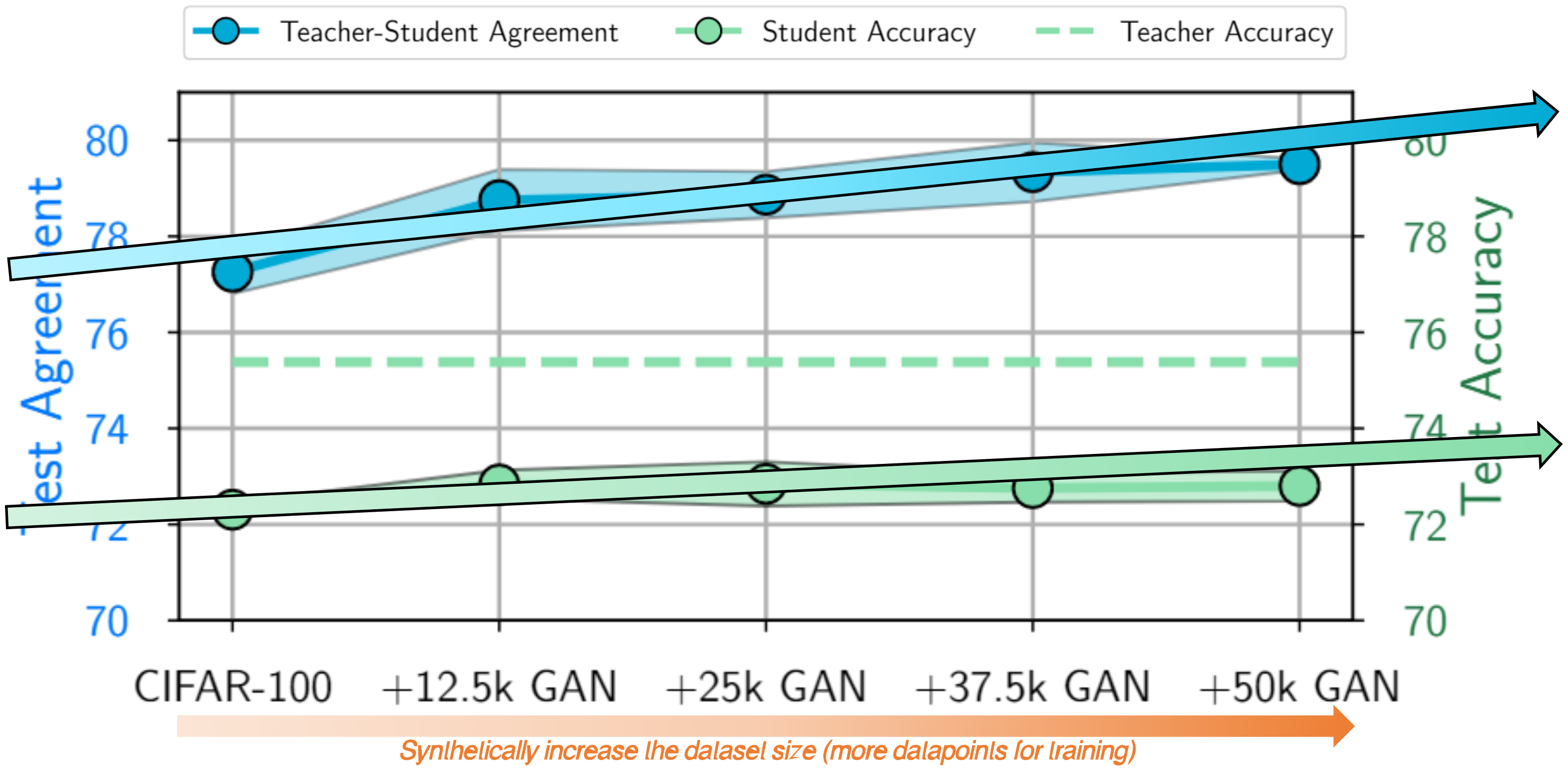
Beyer et al., Knowledge distillation: A good teacher is patient and consistent, CVPR 2022

02 Stanton et al., 2021

(2) Fidelity and generalization needs to be carefully addressed : Non-self-distillation

Common conception (Hypothesis)
Making the student to better mimic the teacher is desirable (Beyer et al., 2022)

Experiment design
Observe: Fidelity & Performance // Varying factor: Amount of dataset (Larger datasets will benefit fidelity)



Student better matches teacher (high fidelity) on more data

The performance slightly increases

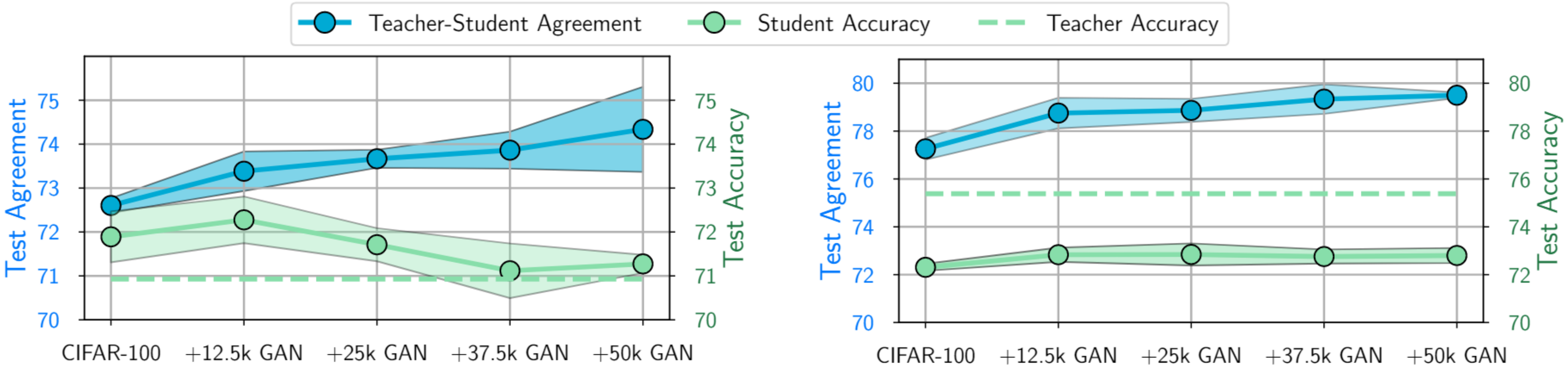
Beyer et al., Knowledge distillation: A good teacher is patient and consistent, CVPR 2022

02 Stanton et al., 2021

(2) Fidelity and generalization needs to be carefully addressed

Common conception (Hypothesis)
Making the student to better mimic the teacher is desirable (Beyer et al., 2022)

Experiment design
Observe: Fidelity & Performance // Varying factor: Amount of dataset (Larger datasets will benefit fidelity)



Despite mixed results, since we cannot in general measure generalization, fidelity is still the key consideration outside self-distillation.

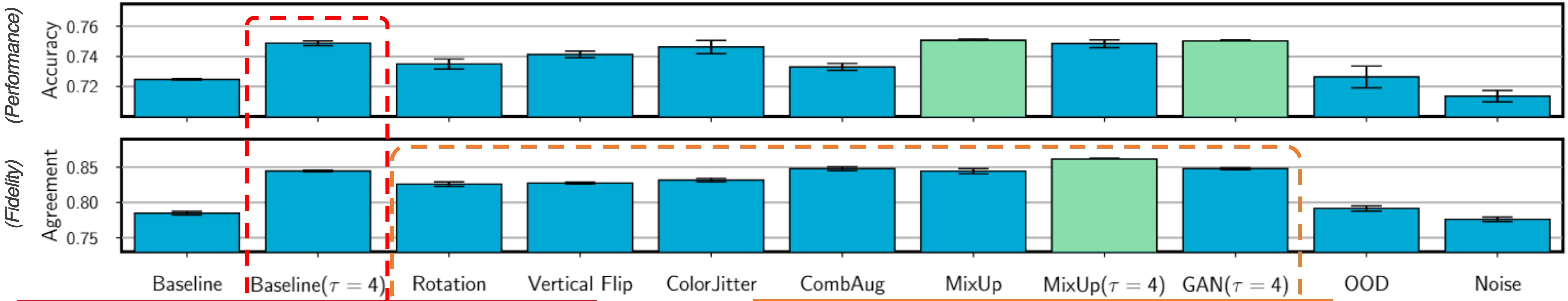
Beyer et al., Knowledge distillation: A good teacher is patient and consistent, CVPR 2022

02 Stanton et al., 2021

(3) Identifiability problem: Have we shown enough teacher outputs to the student?

Question (Hypothesis)
Should we do more data augmentation?

Experiment design
Observe: Fidelity & Performance // Varying factor: Data augmentation strategies // ResNet56 ensemble → ResNet56



1. *Temperature tempering is a strong baseline*
2. *Since this is not an augmentation, insufficient data is not the primary obstacle to high fidelity*

1. *Almost all augmentations increase fidelity*
2. *Mixed results for translating to performance*

02 Stanton et al., 2021

(3) Identifiability problem: Perhaps we are not showing the right teacher outputs

Hypothesis

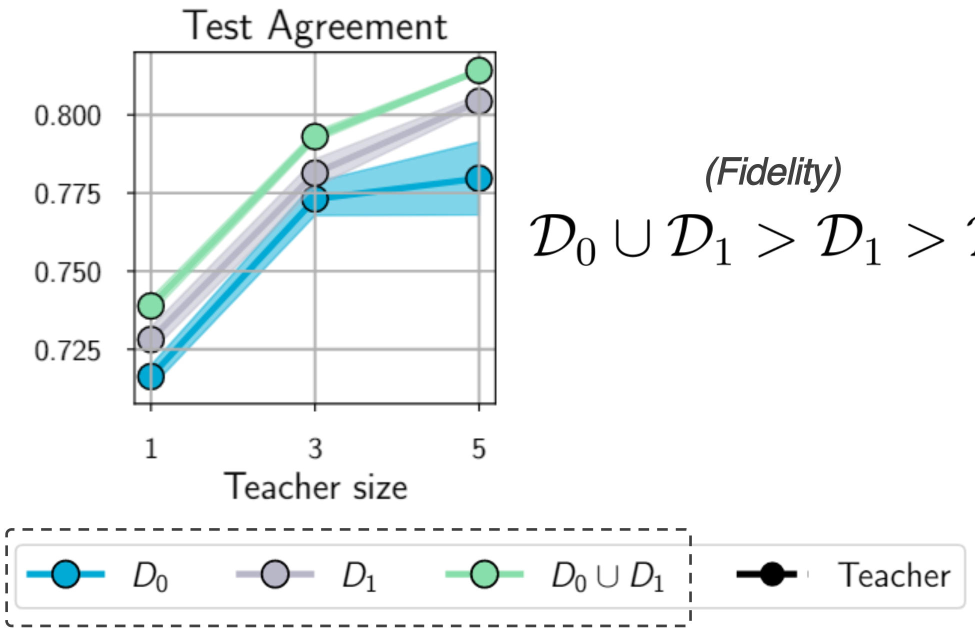
Perhaps we can blame data augmentation (distribution shift) and only using the dataset itself?

Experiment design

Split the dataset into two groups and compare distillation results

CIFAR100 (train) → Split → $\mathcal{D}_0$, $\mathcal{D}_1$

$\mathcal{D}_0$ → Train teacher model



The student is distilled from…

(Fidelity)

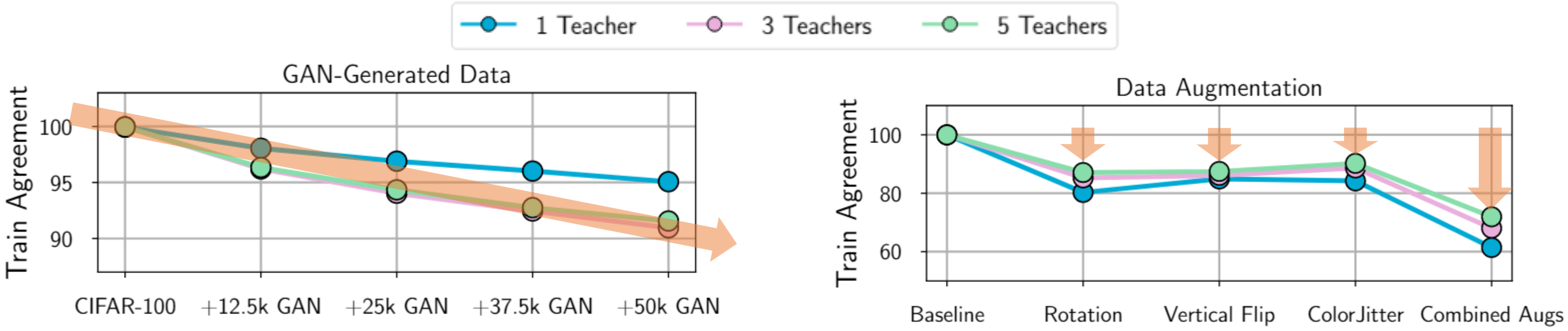
$$\mathcal{D}_0 \cup \mathcal{D}_1 > \mathcal{D}_1 > \mathcal{D}_0$$

- *At all scenarios, best fidelity (~80%) is still lower than the previous analysis (~85%)*
- *Therefore, the distillation data is still not the primary reason for poor fidelity*

02 Stanton et al., 2021

(3) Identifiability problem: Observation on the training dataset (rather than test dataset)

Hypothesis
Perhaps there are simpler answers in the training dataset (distillation dataset).



*Increasing the distillation dataset **decreases** fidelity*

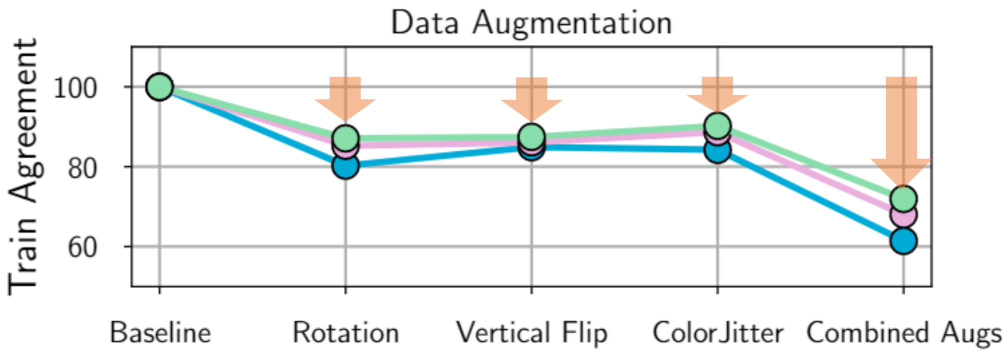
*Heavier augmentations **decreases** fidelity*

*Investigations shows that the **student cannot even match the teacher on the distillation dataset**.*

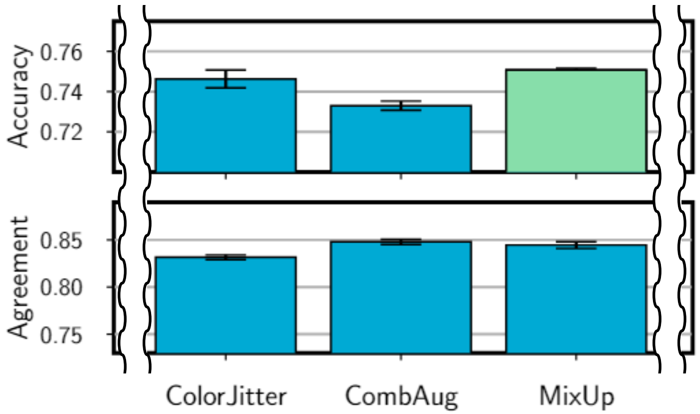
02 Stanton et al., 2021

(3) Identifiability problem: Observation on the training dataset (rather than test dataset)

Trade-off in KD (Hypothesis)
The student needs many data, which increases fidelity in test data but decreases fidelity in training data.



*Heavier augmentations **decreases** fidelity*



*However, it has the best **test fidelity***

Hypothesis
Then the root cause may be in the optimization, rather than the dataset.

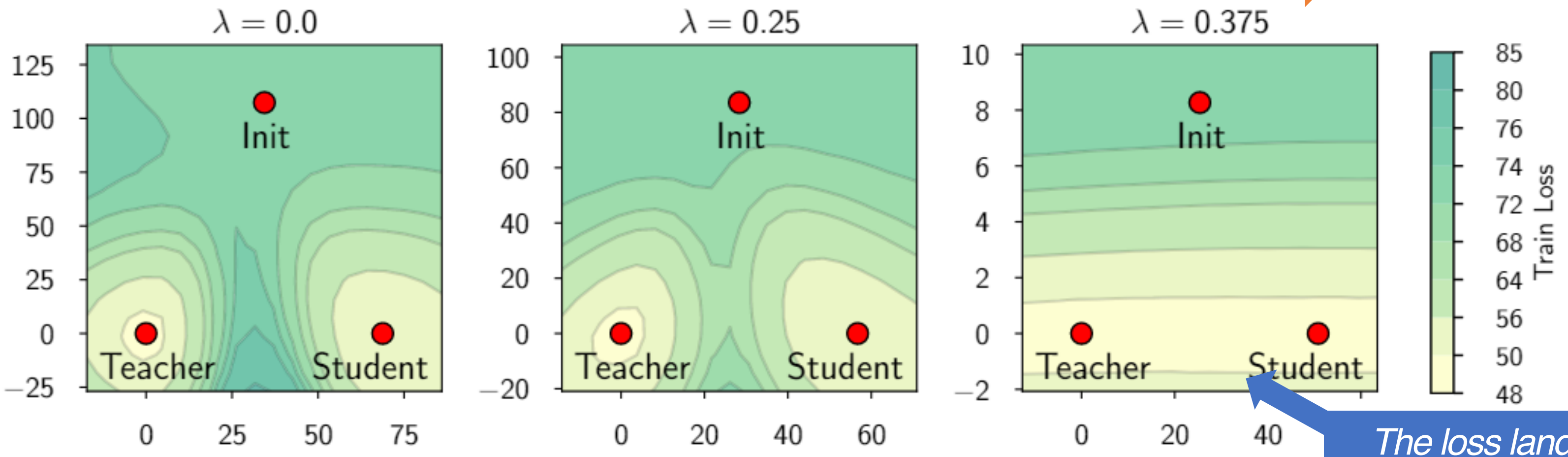
02 Stanton et al., 2021

(3) Identifiability problem: Optimization

Hypothesis

Then the root cause may be in the optimization, rather than the dataset.

Initial student weights are closer to teacher's weights



The loss landscape changes, and the student is now in the same local minima

However, further investigation shows that it is still difficult to match the teacher outputs even when we have access to teacher's weights and use that advantage.

The problem of fidelity is likely to be the results of the optimization dynamics.

03 Summary & Discussions

- *Several investigations on knowledge distillation has been made*
    1. *It seems that the teacher outputs are generally hard to fit for a smaller student model in general*
    2. *Both papers agree that optimization can play a vital role in knowledge distillation*

- *Compared to GLNN (Zhang et al., ICLR 2022)*
    - *With a grain of salt: CV vs. Graph*
        1. *Generally, image datasets have larger classes (~100 classes) compared to graphs (~10 classes).*
           → Increases the chances that class distributions contain complex data
        2. *Different data complexity: # of pixels > # of node attributes, but image has no relational information*
        3. *Different capacity: ResNet, VGG etc. have massive parameters, but GNNs have graph structure as part of the model*
    - *With a graph of salt: CV vs. GLNN*
        1. *Distillation in CV does not worry about input discrepancy as the model has exactly one input (i.e., a single / batch of images).*
        2. *Limited augmentation: Not straightforward for GLNN to discuss edge augmentation as graph topology is not part of the input anyway*