SCAFFOLDING A STUDENT TO INSTILL KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a novel knowledge distillation (KD) method to selectively instill teacher knowledge into a student model motivated by situations where the student's capacity is significantly smaller than that of the teachers. In vanilla KD, the teacher primarily sets a predictive target for the student to follow, and we posit that this target is overly optimistic due to the student's lack of capacity. We develop a novel scaffolding scheme where the teacher, in addition to setting a predictive target, also scaffolds the student's prediction by censoring hard-to-learn examples. The student model utilizes the same information as the teacher's soft-max predictions as inputs, and in this sense, our proposal can be viewed as a natural variant of vanilla KD. We show on synthetic examples that censoring hard-examples leads to smoothening the student's loss landscape so that the student encounters fewer local minima. As a result, it has good generalization properties. Against vanilla KD, we achieve improved performance and are comparable to more intrusive techniques that leverage feature matching on benchmark datasets.

1 INTRODUCTION

A fundamental problem in machine learning is to design efficient and compact models with near state-of-the-art (SOTA) performance. Knowledge Distillation (KD) (Zeng & Martinez, 2000; Bucila et al., 2006; Ba & Caruana, 2014; Hinton et al., 2015) is a widely used strategy for solving this problem wherein the knowledge from a large pre-trained teacher model with SOTA performance is distilled onto a small student network.

Vanilla KD. Hinton et al. (2015) proposed the popular variant of KD by matching the student soft predictions, $\mathbf{s}(\mathbf{x})$ with that of the pre-trained teacher, $\mathbf{t}(\mathbf{x})$ on inputs $\mathbf{x}$. Informally, during student training, an additional loss term, $D_{KL}(\mathbf{t}(\mathbf{x}), \mathbf{s}(\mathbf{x}))$ is introduced that penalizes the difference between student and teacher predictive distributions using Kullback-Leibler (KL) divergence. This promotes inter-class knowledge learned by the teacher. We will henceforth refer to Vanilla KD as KD.

Capacity mismatch between student and teacher. One of the primary issues in Vanilla KD is that the loss function is somewhat blind to the student's capacity to interpolate. In particular, when the student's capacity is significantly lower than the teacher's, we expect the student to follow the teacher only on those inputs realizable by the student.

We are led to the following question: *What can the teacher provide by way of predictive hints for each input, so that the student can leverage this information to learn to its full capacity?*

Our Proposal: Scaffolding a Student to Distill Knowledge (DiSK). To address this question, we propose that the teacher, during training, not only set a predictive target, $\mathbf{t}(x)$, but also provide hints on hard to learn inputs. Specifically, the teacher utilizes its model to output a guide function, $g(\mathbf{x})$, such that the student can selectively focus only on those examples that it can learn.

- *if* $g(\mathbf{x}) \approx 1$, teacher discounts loss incurred by the student on the input $\mathbf{x}$.
- *if* $g(\mathbf{x}) \approx 0$, teacher signals the input $\mathbf{x}$ as learnable by student.

With this in mind we modify the KL distance in the Vanilla KD objective and consider, $D_{KL}(\mathbf{t}(\mathbf{x}), \phi(\mathbf{s}(\mathbf{x}), g(\mathbf{x})))$, where $\phi(\mathbf{s}, g)$, which will be defined later, is such that, $\phi(\mathbf{s}, 0) = \mathbf{s}$ if g offers no scaffolding. However, we must impose constraints on the guide function g to ensure that only hard-to-learn examples are scaffolded - in the absence of such constraints, the guide can declare all examples to be hard, and the student would no longer learn. We propose to do so by means of a budget constraint $B(\mathbf{s}, g) \leq \delta$ to ensure that the guide can only help on a small fraction of examples.

While more details are described in Sec. 3, we note that, in summary, our proposed problem is to take the empirical linear combination of the aforementioned KL distance and a cross-entropy term as the objective, and minimize it under the empirical budget constraint.

We emphasize that $g(\mathbf{x})$ is used only during training . The inference logic for the student remains the same as there is no need for $g(\mathbf{x})$ during inference. The guide function supported student training has three principal benefits. The benefits are explored in Sec. 2.

• **Censoring Mechanism.** Our guide function censors examples that are hard to learn for the student. In particular, when there is a large capacity gap, it is obvious that the student cannot fully follow the teacher. For this reason, the teacher must not only *set an expectation* for the student to predict, but also selectively gather examples that the student has the *ability to predict.*
• **Smoothen the Loss landscape.** We also notice in our synthetic experiments that whenever scaffolding is powerful, and can correct student's mistakes, the loss landscape undergoes a dramatic transformation. In particular, we notice fewer local minima in the loss viewed by the guided student.
• **Good Generalization.** The solution to our constrained optimization problem in cases where the guide function is powerful can ensure good student generalization. Specifically, we can bound the statistical error essentially in terms of student complexity, and as such do not suffer additional complexity due to the teacher.

Contributions. We summarize our main results.

• We develop a novel approach to KD that exploits teacher representations to adjust the predictive target of the student by scaffolding hard-to-learn points. This novel scaffolding principle has wider applicability across other KD variants, and is of independent interest.
• We design a novel response-matching KD method (Gou et al., 2021) which is particularly relevant in the challenging regime of large student-teacher capacity mismatch. We propose an efficient constrained optimization approach that produces powerful training scaffolds to learn guide functions.
• Using synthetic experiments, we explicitly illustrate the structural benefits of scaffolding. In particular, we show that under our approach, guides learn to censor difficult input points, thus smoothening the student's loss-landscape and often eliminating suboptimal local minima in it.
• Through extensive empirical evaluation, we demonstrate that the proposed DiSK method
– yields large and consistent accuracy gains over vanilla KD under large student-teacher capacity mismatch (upto 5% and 2% on CIFAR-100 and Tiny-Imagenet).
– produces student models that can get near-teacher accuracy with significantly smaller model complexity (e.g. $8\times$ computation reduction with $\sim 2\%$ accuracy loss on CIFAR-100).
– improves upon vanilla KD even under small student-teacher capacity mismatch, and is even competitive with modern feature matching approaches.

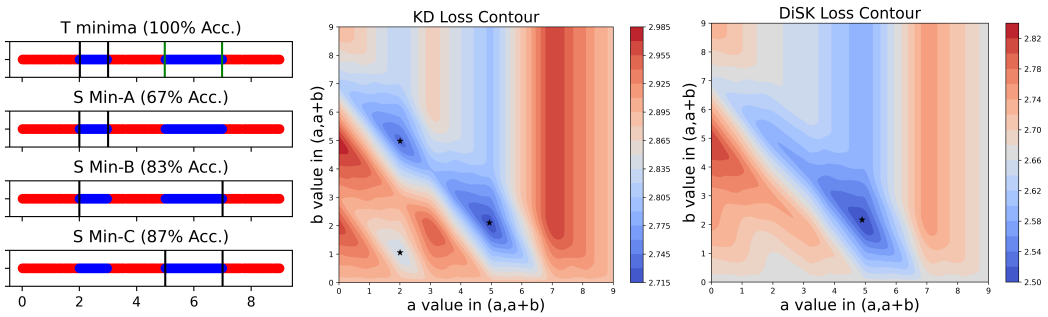
2 Illustrative Examples

We present two synthetic examples to illustrate the structural phenomena of the censoring mechanism and smoothening of student's loss landscape enabled by the scaffolding approach DiSK, which lead to globally optimal test errors. We defer exact specification of the algorithm to Sec.3.

Example 1 (1D Intervals). Consider a toy dataset with one dimensional features $x \in [0, 9]$ and binary class labels $y \in \{\mathrm{Red}, \mathrm{Blue}\}$ as shown in Figure 1. There are two Blue labelled clusters as in $[2, 3]$ and in $[5, 7]$. The remaining points are labelled as Red. We sample 1000 i.i.d. data points as the training set and 100 data points as the test set with balanced data from both classes. We describe the details of the experiment setup such as models and training details in Appendix A.1.

Teacher T belongs to the 2-interval function class, and the capacity-constrained student S belongs to the 1-interval function class. Since teacher capacity is sufficient to separate the two classes without error, it learns the correct classifier (see Figure 1). In contrast, the best possible student hypothesis cannot correctly separate the two classes. Hence, the student will have to settle onto one of the many local minima. We show these minima and the contour plot for the student in Figure 1. We present the results of training student models with different initializations in Table 1.

KD suffers from bad local minima. KD loss landscape contains many local minima (see Figure 1). Due to a big gap between student and teacher capacity, it is unable to help the student discern between these minima. Hence, KD fails to distinguish between the different minima (see Table 1).



(a) **(b)** **(c)**

Figure 1: 1a: 1D Intervals. Data distribution on x-axis $[0, 9]$. Teacher T learns the correct decision boundary with 2-intervals and it is the global minima for this binary classification task. Student S has many bad local minima, and one global minima that best describes the decision boundary with 1-interval. 1b: Contour plot for the KD loss shows the various local minima in the loss landscape. 1c: Contour plot for DiSK shows the bad local minima no longer exist.

Table 1: The number of times each method lands on various local minima in two toy problems for 100 runs.

Dataset	**1D Intervals**			**2D Gaussians**			
Minima	**A**	**B**	**C (Global)**	**A**	**B**	**C**	**D (Global)**
Accuracy	**67%**	**83%**	**87%**	**70%**	**80%**	**90%**	**100%**
Cross-Entropy	35	64	1	73	12	9	6
KD	30	67	3	1	11	31	57
DiSK	9	1	90	0	0	3	97

DiSK censors interval $[2, 3]$ *and in addition focuses training on learnable datapoints.* If we analyze the guide function at the end of the training, we see that it covers (censors) the first Blue cluster. Indeed, both clusters are not simultaneously *l*earnable with available student capacity. However, if we were to censor the interval $[2, 3]$, then the problem becomes realizable for the student model. The guide function thus captures the excess capacity of the data.

DiSK smoothens loss landscape. The guide function and the budget constraint enable our method to have a smooth loss landscape thanks to the guide-function censoring points, which eliminate the local minima. Hence, DiSK solution lands in the global minimum with high probability.

Example 2 (2D Gaussians). Consider another toy dataset with two dimensional features $\mathbf{x} \in \mathbb{R}^2$ and three class labels $y \in \{\text{Red}, \text{Green}, \text{Blue}\}$. Here we wish to show that DiSK can allow for *globally optimal solutions reaching 100% accuracy, which appears unachievable with cross-entropy minimization regardless of data size.*

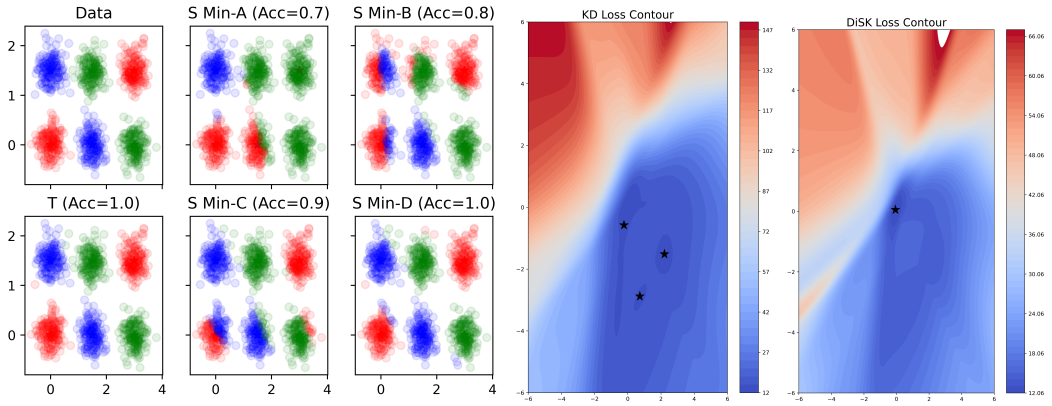
Figure 2a shows the labelled data. There are six cluster centers, two with each class label. Data points are drawn using Gaussian balls around the cluster centers with small radii. We sample 1000 i.i.d. data points as the training set and 1000 data points as the test set with equal representation from all three classes. We provide details (hypothesis classes, learning procedure, etc.) in Appendix. A.2.

The teacher is a 3-layer neural network with 8, 16, and 3 neurons. The student is a 2-layer neural network with 2 and 3 neurons. We point out that the teacher being an over-parameterized network in this feature space, easily learns the correct decision boundary. While the student being severely constrained network suffers in learning the task. Different training runs lead to different local minima. We show teacher solution and student local minima in Figure 2a. The contour plots for the student models under KD loss and DiSK loss are shown in Figure 2b-2c using Li et al. (2018).

The results are similar to the 1D example - KD converges to a poor local minimum with at least 43% of the initializations, while in contrast, DiSK escapes these by focusing on the learnable part of the input space (Fig. 2c), converging to the global minimum nearly always (Table 1).

To conceptualize our findings in these examples, let us attempt to intuitively infer the example-censoring, landscape-smoothening, and good generalization, by utilizing the following conditions that appear to be satisfied for these synthetic examples.

Realizability. Suppose we are in a situation where the guide function $g \in \mathcal{G}$ is sufficiently powerful that there is a student and guide function capable of interpolation, i.e., predictions supported by the guide function, $\phi(\mathbf{s}, g)$, interpolates to mimic the labels.



(a) **(b)** **(c)**

Figure 2: 2a. 2D Gaussians. Data distribution on $\mathbb{R}^2$. Teacher T learns the correct decision boundary with 3 layer neural network and it is the global minima for this three-way classification task. While student S has many bad local minima, and one global minima that best describes the decision boundary with 2 layer neural network. 2b. Contour plot for the KD loss shows the various local minima in the loss landscape. 2c. Contour plot for DiSK shows the bad local minima no longer exist (wider minima, join two adjust minima, remove bad local minima).

Example: For instance, consider a binary classification problem with the labels $y \in \{-1, 1\}$. Let $\phi(s, g) = y(s + g)$ with $s(\mathbf{x}) \in [-1, 1]$. Our realizability condition is that we always satisfy $y(s(\mathbf{x}) + g(\mathbf{x})) > 0$. As such, this leads to the condition that if $ys(\mathbf{x}) \leq 0$, then $yg(\mathbf{x}) > 0$. Therefore, $\mathbb{E}[\mathbb{1}_{[ys(\mathbf{x})<0]}] \leq \mathbb{E}[\mathbb{1}_{[ys(\mathbf{x})<0, yg(\mathbf{x})>0]}] \leq \mathbb{E}[\mathbb{1}_{[yg(\mathbf{x})>0]}]$.

Small Guide Function Capacity. In addition to realizability suppose the class of guide functions $g \in G$ has a small capacity (for instance, small VC dimension). For our case this condition is satisfied because our guidance function is obtained by using an MLP on teacher's last layer features.
Example: Continuing with the example above, say we now have m training instances, $(\mathbf{x}_i, y_i)$, $i \in [m]$, $\hat{g}(\mathbf{x})$ is guide function output of DiSK. We can infer by standard statistical learning results (Shalev-Shwartz & Ben-David, 2014) that, for the estimated function $\hat{g} \in G$, it follows with probability greater than $1 - \eta$ that $\mathbb{E}[\mathbb{1}_{[y\hat{g}(\mathbf{x})>0]}] \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{[y_i\hat{g}(\mathbf{x}_i)>0]} + \sqrt{\frac{VC(G)}{m} + \frac{\log\frac{1}{\eta}}{m}}$. As a result, we can say that if there is a student, $s(\mathbf{x})$ (not necessarily that output by DiSK), which complements $\hat{g}(\mathbf{x})$ and satisfies realizability, then with probability greater than $1 - \eta$: $\mathbb{E}[\mathbb{1}_{[ys(\mathbf{x})<0]}] \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{[y_i\hat{g}(\mathbf{x}_i)>0]} + \mathcal{O}\left(\sqrt{\frac{VC(G)+\log\frac{1}{\eta}}{m}}\right)$.

Remarks. Note that the key point here is that the student capacity is considerably larger since we typically train an entire DNN, and student complexity-based bound can be vacuous. While the guidance function does bound the student generalization error in terms of guide function complexity, there are strong caveats— we require the strong assumption of realizability on the entire domain, and additionally, while the guide function can witness student error, we are not in a position to precisely estimate it without additional training data. Furthermore, the RHS is a relaxed bound on the student training error. This motivates having a budget constraint to ensure that student learns with small training error, a point that will be clearer in the sequel.

3 Definitions and Formulations

Notation. Let $\mathcal{X}$ and $\mathcal{Y} = \{1, \ldots, C\}$ be the feature and label spaces respectively, focusing on a C-class classification task. We assume that we have a training set of N i.i.d. data points $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. We use symbols S and T to denote the student and teacher models respectively. Let $\mathbf{l}_S(\mathbf{x}) \in \mathcal{R}^{|\mathcal{Y}|}$ and $\mathbf{l}_T(\mathbf{x}) \in \mathcal{R}^{|\mathcal{Y}|}$ be the scores (logits) vector predicted by S and T on input $\mathbf{x}$. We use τ as the temperature used to soften the probability distribution. We write the resulting softened student and teacher probabilities as $\mathbf{s}(\mathbf{x})$ and $\mathbf{t}(\mathbf{x})$, i.e.,

$$\mathbf{s}^{\tau}(\mathbf{x}) = \text{softmax}\left(\frac{\mathbf{l}_S(\mathbf{x})}{\tau}\right); \quad \mathbf{t}^{\tau}(\mathbf{x}) = \text{softmax}\left(\frac{\mathbf{l}_T(\mathbf{x})}{\tau}\right)$$

The standard soft prediction probabilities correspond to $\mathbf{s}^1(\mathbf{x})$ and $\mathbf{t}^1(\mathbf{x})$. We will use symbols $\mathbf{s}^\tau_y(\mathbf{x})$ to denote the yth coordinate in $\mathbf{s}^\tau(\mathbf{x})$, and similarly $\mathbf{t}^\tau_y(\mathbf{x})$. The hard prediction of the student is $p_S(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} \mathbf{s}^1(\mathbf{x})$, and similarly $p_T(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} \mathbf{t}^1(\mathbf{x})$ for the teacher.

We use $g(\mathbf{x}) \in [0,1]$ to denote the helper guide function for the student and teacher pair (S,T). Guide takes input $\mathbf{x}$ and any other feature processed by (S,T) pair and decides whether or not the student needs help on the input $\mathbf{x}$. Finally, we define ReLU activation as $(\cdot)_+ = \max(0,\cdot)$.

3.1 Vanilla Knowledge Distillation

Vanilla KD relaxes the $0-1$ error between the student predictions and the true labels y using the cross-entropy loss $\mathcal{L}_{CE}$. Similarly, KD denotes the distance between the student and teacher softened probability distributions using the KL divergence. We summarize the corresponding losses as,

$$\mathcal{L}_{CE}(\mathbf{s}) = -\frac{1}{N}\sum_{i=1}^{N}\log \mathbf{s}^1_{y_i}(\mathbf{x}_i); \quad \mathcal{L}^\tau_{KL}(\mathbf{s}) = -\frac{1}{N}\tau^2\sum_{i=1}^{N}\sum_{y}\mathbf{t}^\tau_y(\mathbf{x}_i)\log\frac{\mathbf{s}^\tau_y(\mathbf{x}_i)}{\mathbf{t}^\tau_y(\mathbf{x}_i)}$$

For hyperparameters $\alpha \in [0,1], \tau > 0$, KD minimizes a mixture of the above losses, as shown below

$$\mathcal{L}^{\tau,\alpha}_{KD}(\mathbf{s}) = \alpha\mathcal{L}_{CE}(\mathbf{s}) + (1-\alpha)\mathcal{L}^\tau_{KL}(\mathbf{s}). \tag{1}$$

3.2 Selective Knowledge Distillation.

KD attempts to transfer the knowledge from the teacher to the student on all training data points, which is a sub-optimal objective when there is a capacity mismatch between the student and the teacher. Instead, we propose distilling selective knowledge (DiSK) to allow the student to selectively ignore some hard-to-learn data points during training, transferring the teacher's knowledge only on easy-to-learn inputs, and matching the learning to student capacity. Our objective is to minimize

$$\min_{\mathbf{s},g,\delta} \underbrace{\frac{1}{N}\sum_{i=1}^{N}\text{distance}(\mathbf{t}(\mathbf{x}_i);\phi(\mathbf{s},g)(\mathbf{x}_i))}_{\text{Distance between } T \text{ and } S \text{ with help of } g} \quad \text{subject to.} \quad \underbrace{\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{x}_i)\mathbb{1}_{\{y_i \neq \arg\max_y \mathbf{s}_y(\mathbf{x}_i)\}} \leq \delta}_{\text{Support budget constraint on } g} \tag{2}$$

where ϕ interpolates student predictions based on the guide's help. The divergence term helps in minimizing the distributional distance between the teacher and student probabilities after the guide g is included. While the budget term in the optimization constrains the helper g to provide help only when necessary, the amount of help given to the student should be within the budget $\delta \in [0,1]$.

Function g Construction. As previously stated, we use the teacher's last layer features and soft predictions as input to the guide g. The guide is structured as a light-weight three-layer neural network with these inputs, with a sigmoid activation at the last layer. We re-emphasise that g is not used at inference time, and only aids training. More details are left to Appx.A.4.

Relaxed Losses, Lagrangian & Optimization Algorithm. We relax Eq. 2 and construct a Lagrangian by integrating the constraint into the minimization.

Budget constraint relaxation. We relax the indicator loss in the budget to a cross-entropy, and treat δ as a hyperparameter to get

$$\mathcal{L}^\delta_{budget}(\mathbf{s},g) = \left[-\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{x}_i)\log s^1_{y_i}(\mathbf{x}_i) - \delta\right]_+ \tag{3}$$

We view the scaffold as a way to allow the student to interpolate the uncensored data. This suggests that a good initialization for the budget is the error of cross-entropy trained model, when the student does not have the benefit of teacher supervision. Using this insight, we scan the budget in a small interval around this initialization.

Distillation objective. Motivated from KL loss, we construct a distance loss with guide function as,

$$\mathcal{L}^{\tau,K}_{dist}(\mathbf{s},g) = -\frac{1}{N}\tau\tau_{\mathbf{t},\mathbf{s},\mathcal{D}}\sum_{i=1}^{N}\sum_{y} t^\tau_y(\mathbf{x}_i)\log\left(s^{\tau_{\mathbf{t},\mathbf{s},\mathcal{D}}}_y(\mathbf{x}_i) + 1_{y\in\text{top}_K(t^\tau(\mathbf{x}_i))}g(\mathbf{x}_i)\right) \tag{4}$$

Algorithm 1 DiSK: Distilling Selective Knowledge.

Input: Training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, Teacher $\mathbf{t}$,
Parameters: τ, K, α, $\lambda_{\min}$, $\lambda_{\max}$, Number of iterations R, λ_T cosine period, Budget δ,
Initialize: $\mathbf{s}$, randomly initialize g, $\lambda = \lambda_{\min}$,
for $r = 1$ **to** R **do**
  Randomly Shuffle Dataset $\mathcal{D}$
  $g \leftarrow \arg\min_g \mathcal{L}_{DiSK}^{\tau,K,\delta,\alpha}(\mathbf{s}, g, \lambda)$
  $\mathbf{s} \leftarrow \arg\min_{\mathbf{s}} \alpha\mathcal{L}_{CE}(\mathbf{s}) + (1-\alpha)\mathcal{L}_{dist}^{\tau,K}(\mathbf{s}, g)$
  $\lambda \leftarrow \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) * \frac{(1-\cos\frac{r \mod \lambda_T}{\lambda_T}\pi)}{2}$
Return : $\mathbf{s}$

We point out two modifications in the distillation loss. First, $\mathcal{L}_{dist}$ explicitly adds guide value to softened student probabilities in selected class indices. The class indices guide function adds value are picked as top K classes based on the teacher probabilities for any input $\mathbf{x}_i$ where K is a hyperparameter of our method. The rest of the class indices do not get any value from g. Second, we use different temperature parameters for teacher and student. Temperature parameter for teacher, τ, is a hyperparameter. The student temperature is found by minimizing the KL loss between teacher softened probabilities and the student softened probabilities over the training dataset, .i.e $\tau_{\mathbf{t},\mathbf{s},\mathcal{D}} = \arg\min_{\tau'} \sum_i KL(\mathbf{t}^{\tau}(\mathbf{x}_i), \mathbf{s}^{\tau'}(\mathbf{x}_i))$.

Similar to KD, we incorporate standard cross entropy loss between student model predictions and the ground truth labels for stability. We construct our Lagrangian by combining Eq. 3 and 4 as,

$$\mathcal{L}_{DiSK}^{\tau,K,\delta,\alpha}(\mathbf{s}, g, \lambda) = \alpha\mathcal{L}_{CE}(\mathbf{s}) + (1-\alpha)\mathcal{L}_{dist}^{\tau,K}(\mathbf{s}, g) + \lambda\mathcal{L}_{budget}^{\delta}(\mathbf{s}, g) \tag{5}$$

where α is a hyperparameter and λ is the dual parameter of DiSK.

We optimize Obj. 5 using a primal dual update scheme as explained in Algorithm 1.

Primal Parameter Updates $(\mathbf{s}, g)$*.* We learn the student $\mathbf{s}$ and the guide function g using alternating minimization. We approximate $\arg\min$ with running SGD for a small number of epochs on the full dataset. In each iteration, we first learn the guide function g to select the data partition from which the knowledge needs to be distilled. Next, given the function g, we learn the student using the help g. We empirically found that not optimizing the student model on budget loss gives more stable results. Hence, we minimize the student model only on the distillation and cross-entropy losses.

Dual Parameter Update (λ) *Intuition.* Although it is tempting to optimise the above via a dual ascent and primal descent scheme (wherein the dual parameter λ is increased by residual term in the constraint until constraint satisfaction), recent work (Sun & Sun, 2021) has proposed to decrease the λ in the non-convex regime. Inspired by this, we update the dual parameter by a fixed schedule between $[\lambda_{\min}, \lambda_{\max}]$, and update the bounds $\lambda_{\min}, \lambda_{\max}$ during training to enforce the constraint. $\lambda_{\min}$ encourages exploration and allows student model to distill knowledge from all points. On the other hand, $\lambda_{\max}$ enforces the constraint and forces the student model to learn on uncensored inputs. We choose $R \approx 4\lambda_T$, so that the algorithm is exposed to a few exploratory periods. For the final period, we increase λ monotonically so that budget is more strictly enforced at termination.

Computational Efficiency. Algorithm 1 trains both student and guide networks. The guide network being small (three-layer MLP) relative to the student (CNN model), the additional cost in training the guide is relatively insignificant, and as such DiSK efficiency is similar to vanilla KD.

4 Experiments

We evaluate DiSK under various capacity mismatch scenarios on benchmark vision datasets.

Datasets. We use publicly available CIFAR-100 (Krizhevsky, 2009), Tiny-Imagenet (Le & Yang, 2015) datasets. CIFAR-100 contains 50K training and 10K test images from 100 classes with size $32 \times 32 \times 3$. While Tiny-Imagenet contains 100K training and 10K test images from 200 classes with size $64 \times 64 \times 3$. We provide the dataset setup and data augmentations used in detail in Appendix A.3.

Models. We evaluate standard convolutional models on these datasets including ResNet(He et al., 2016), Wide-ResNet(Zagoruyko & Komodakis, 2016), MobileNet(Sandler et al., 2018), and Shuf-

Table 2: Model Statistics: We compute the storage (number of parameters) and computational requirements (number of multiply-addition operations) of the models used in this work.

Architecture		CIFAR-100		Tiny-Imagenet		Architecture	CIFAR-100	
		MACs	Params	MACs	Params		MACs	Params
Teacher	Resnet10-ℓ	64M	1.25M	255M	1.28M			
	Resnet10	253M	4.92M	1013M	5M	Resnet32x4	1083M	7.4M
	Resnet18	555M	11.22M	2221M	11.27M	Wide-Resnet-40-2	327M	2.25M
	Resnet34	1159M	21.32M	4637M	21.38M			
Student	Resnet10-xxs	2M	13K	8M	15K	Resnet8x4	177M	1.2M
	Resnet10-xs	3M	28K	12M	31K	ShufflenetV2	44.5M	1.4M
	Resnet10-s	4M	84K	16M	90K	Wide-Resnet-16-2	101M	700K
	Resnet10-m	16M	320K	64M	333K	Wide-Resnet-40-1	83M	570K
						MobileNetV2x2	22M	2.4M

fleNet(Ma et al., 2018). Table 2 shows the storage and computational requirements of all the models used in this work. We provide explicit model configurations in Appendix A.4, including the tiny models we generate from the Resnet architectures.

Methods. We study performance against standard cross-entropy (CE) based learning and the vanilla KD methods. For each method, we train models for 200 epochs using SGD as the optimizer with 0.9 momentum and 0.1 learning rate. See Appendix A.5, for training details. We have recorded the mean in our results as the variance of 3 trials in our experiments is not larger than 0.1 in most of the cases.

We perform evaluations in different settings. Below, we explain individual setups.[1]

Large Capacity Mismatch Setting. We distill knowledge from a teacher model into a student model where the student has much less capacity compared to teacher model. We use four large capacity Resnet teachers and five tiny Resnet students and train these students using CE, KD, and DiSK methods. Performances, and the gains of DiSK are reported in Table 3.

Small Capacity Mismatch Setting. While DiSK has been designed for the scenario when student capacity is very low, we further evaluate it in the setting where teacher and student capacities are similar, to probe how far the power of the method extends. The model classes used are the standard choice for this scenario (Chen et al., 2022; Tung & Mori, 2019). Performance is reported in Table 4.

Table 4 further reports the results of the feature matching distillation methods: FitNets (Romero et al., 2015), SemCKD (Chen et al., 2021), and SimKD (Chen et al., 2022). Such methods can often outperform response matching KD on large students, due to student representations that are more aligned with the teacher, but typically at an increased training cost. While feature matching methods are not the main focus of our work (and in principle scaffolding idea can be extended to them), we observe that DiSK often improves upon their performance without any direct feature matching.

Experiment results. Below, we highlight salient features of DiSK based on empirical data.

DiSK outperforms the baselines uniformly across all datasets and student sizes. As shown by Table 3, DiSK significantly improves the student performance in CIFAR-100 and the (more challenging) Tiny-Imagenet dataset, respectively showing accuracy gains of up to 5% and 2% compared to KD. These gains are consistent across a wide range of student and teacher capacities.

DiSK achieves better performance with worse teachers than KD does with even the best teachers. In Table 3, we point out that the student performance increases for KD as the teacher complexity is increased for a given student. But note that for the same student, DiSK achieves much better performance with even worse teacher. For instance, for the 'Resnet10-m' student, KD accuracy increases from 66.96% to 68.09% by using high capacity teachers. But 'Resnet10-m' trained with even the worst teacher ('Resnet10-ℓ') achieves 70.03% accuracy. This saves a lot of resources in any application as large teacher requires more training time, and larger compute resources.

DiSK is competitive even in small capacity difference setting. As shown by Table 4, DiSK does not loose its competitive edge over the KD even when the student is relatively similar sized as teachers, and shows gains of up to 2.5% relative to KD. We conjecture that the observed gains arise from the fact that DiSK provides scaffolding for hard points to the student in initial training stages, which promotes the student to learn easy examples first. As training progresses, DiSK removes the discounted help from hard inputs. As a result, the student evolves from simpler hypothesis to the ones

[1]We cover more ablative experiments and refer to Appendix A due to page limit.

Table 3: DiSK performance under large capacity mismatch on CIFAR-100 & Tiny-Imagenet: We draw mismatched teachers and students from the Resnet family, and report accuracy of CE trained teachers and students, performance of students distilled using KD and DiSK, and gains of the latter relative to KD.

Architecture		CIFAR-100					Tiny-Imagenet				
		Accuracy (%)					Accuracy (%)				
Teacher	**Student**	**Teacher**	**CE**	**KD**	**DiSK**	**Gain**	**Teacher**	**CE**	**KD**	**DiSK**	**Gain**
Resnet10-ℓ	Resnet10-xxs	71.99	32.05	32.64	**37.56**	**4.92**	52.14	17.44	17.59	**18.62**	**1.03**
	Resnet10-s		52.16	54.92	**58.14**	**3.22**		34.65	35.77	**37.43**	**1.66**
	Resnet10-m		65.24	66.96	**70.03**	**3.07**		44.74	46.01	**48.03**	**2.02**
Resnet10	Resnet10-xxs	75.25	32.05	34.25	**37.84**	**3.59**	56.04	17.44	17.96	**18.55**	**0.59**
	Resnet10-s		52.16	54.95	**58.36**	**3.41**		34.65	36.11	**37.37**	**1.26**
	Resnet10-m		65.24	67.27	**70.15**	**2.88**		44.74	46.08	**48.19**	**2.11**
Resnet18	Resnet10-xxs	76.56	32.05	34.16	**37.8**	**3.64**	62.48	17.44	17.47	**18.53**	**1.06**
	Resnet10-s		52.16	55.76	**58.11**	**2.35**		34.65	35.59	**37.5**	**1.91**
	Resnet10-m		65.24	68.09	**69.86**	**1.77**		44.74	45.91	**47.7**	**1.79**
Resnet34	Resnet10-xxs	80.46	32.05	33.93	**37.78**	**3.85**	63.06	17.44	17.67	**18.91**	**1.24**
	Resnet10-s		52.16	54.19	**58.02**	**3.83**		34.65	35.43	**37.68**	**2.25**
	Resnet10-m		65.24	66.78	**69.89**	**3.11**		44.74	45.89	**47.6**	**1.71**

Table 4: DiSK performance with small capacity mismatch on CIFAR-100: We pick standard student and teacher configurations used in the KD literature, and report accuracies and gains similarly to Table 3. Feature matching KD baselines are due to Chen et al. (2022).

Architecture		CIFAR-100							
		Response Matching KD					Feature Matching KD		
		Accuracy (%)					Accuracy (%)		
Teacher	**Student**	**Teacher**	**CE**	**KD**	**DiSK**	**Gain**	**FitNet**	**SemCKD**	**SimKD***
Resnet32x4	Resnet8x4	81.45	73.89	76.25	**76.92**	**0.67**	74.32	76.23	78.08
	ShuffleNetV2		73.74	79.13	**80.23**	**1.1**	75.82	77.62	78.39
	Wide-Resnet-16-2		74.26	76.28	**77.67**	**1.39**	74.70	75.65	77.17
	MobileNetV2x2		69.24	76.05	**77.24**	**1.19**	73.09	73.98	75.43
Wide-Resnet-40-2	Resnet8x4	78.41	73.89	75.15	**76.05**	**0.9**	75.02	75.85	76.75
	ShuffleNetV2		73.74	75.81	**78.33**	**2.52**	-	-	-
	Wide-Resnet-40-1		72.81	74.44	**75.92**	**1.48**	74.17	74.4	75.56
	MobileNetV2x2		69.24	73.92	**76.32**	**2.40**	-	-	-

*SimKD accuracy is not emphasized as it employs additional layers beyond the given student architecture and thus not directly comparable to other methods.

consistent with both easy and hard inputs. This justifies our dual parameter (λ) update in Algorithm 1, wherein we periodically increase and decrease λ to enforce and relax the budget constraint.

DiSK students achieve near teacher accuracy while saving up to $8\times$ *MACs &* $5\times$ *Params.* As reported in Table 3, student ('Resnet10-m') trained with the teacher ('Resnet10-ℓ') achieves close to the teacher accuracy of 71.99%. In this process, it saves $4\times$ compute and requires $4\times$ less parameters. Similarly, student ('ShuffleNetV2') trained with the teacher ('Resnet32x4') achieves close to the teacher accuracy of 81.45%. In this process, it saves $24\times$ compute and requires $5\times$ less parameters.

DiSK cleverly selects a subset of datapoints and smoothens the loss landscape. As illustrated in Figure 1 and 2, DiSK judiciously selects a subset of hard-to-learn data points for the students and provide discounted help to the student focus on easily learnable inputs. As a result, it eliminates some bad local minima in the student loss-landscape, and smoothens tihs surface.

DiSK enables the student to reach saturation capacity. In Table 3, the performance of KD often suffers as the teacher size is increased, e.g., student ('Resnet10-s') accuracy decreases substantially with the teacher 'Resnet34' versus the teacher 'Resnet18'. In contrast, DiSK saturates the student performance across different teachers. For instance, student ('Resnet10-m') accuracy is $\approx 70\%$ for all the teachers. Thus, we point out that DiSK enables the student to reach saturation. This may be due to the fact that guide g identifies the same set of 'easy' points across different teachers.

5 Related Work

We refer the reader to Gou et al. (2021) for a comprehensive survey on knowledge distillation.

Response Matching. Zeng & Martinez (2000); Bucila et al. (2006) distill the response of an ensemble of classifiers into a single neural network by creating a pseudo-labeled dataset using the ensemble of

classifiers. Ba & Caruana (2014) extend this to the setting with the neural network as the teacher. Hinton et al. (2015) propose vanilla KD that distilled knowledge from an ensemble of neural networks into a single network by matching their output logits. This work provided a simple recipe for aligning the teacher and student predictive distributions using the Kullback-Leibler (KL) divergence. Recently, Beyer et al. (2021) modify the KD procedure to include patient and consistent teacher resulting in substantial gains. Knowledge consistency is enforced by using the same aggressive data augmentation and image views in the student as in the teacher. Patience is promoted using a very long training schedule. This results in a computationally very expensive training process.

Stanton et al. (2021) analyze response matching KD and suggests that difficulty in optimization leads to poor knowledge distillation. Thus, the teacher and student predictions do not always match, even on the training data. Cho & Hariharan (2019) study Vanilla KD through the lens of mismatched student and teacher capacities. It shows that small students are unable to mimic complex teachers. They proposed early stopping teacher training as to remedy to achieve a student-learnable teacher. The above works fail miserably when the gap between student and teacher complexities is large. Specifically, the student cannot learn the complex teacher decision boundaries primarily due to the small student capacity. It becomes imperative to selectively choose only easy-to-learn data points and transfer the teacher knowledge from these points and ignore the hard-to-learn data points during distillation. Thus, our proposal targets the problem of severe capacity gaps between student and teacher models. Additionally, our experiments with standard student and teacher configurations show that DiSK is still competitive when the capacity difference is small.

Feature Matching. In response matching, teacher supervision is limited to its logits. We can enforce intermediate layer feature matching for refined teacher supervision. FitNets (Romero et al., 2015) extend the KD by including the feature matching in the middle layers. Zagoruyko & Komodakis (2017) used feature map attention as the teacher supervision. Tung & Mori (2019) preserve pairwise similarity in feature maps amongst data points during the distillation. Chen et al. (2022) modify the student by projecting the student features onto the teacher feature space and by reusing the teacher classifier. While our work focuses primarily on selective distillation in Vanilla KD for simplicity. We can easily extend the proposed framework to incorporate it into feature-matching distillation.

Privileged Information. Vapnik & Izmailov (2015) propose the 'learning under privileged information' (LUPI) framework wherein a support vector machine is trained using privileged information unavailable during the inference stage. Later, Lopez-Paz et al. (2016) unified LUPI and Vanilla KD into generalized distillation, wherein the teacher is learned using the privileged information. Next, the student is trained using the ground truth and the teacher labels. These works rely on privileged information in the application domain and are shown to work on toy setups. Since our guide function, g is available only during training, it can be thought of as privileged information from a teacher.

Curriculum Learning & Hard Instance Mining. Curriculum Learning (CL) (Bengio et al., 2009; Hacohen & Weinshall, 2019; Graves et al., 2017) sorts the data based on their hardness as measured by some scoring function (ex., predictive entropy, softmax margin score, etc.). It presents the data points during training in the order of increasing hardness. Similarly, Hard Instance Mining (HIM) (Zhou et al., 2020) reduces the weight of the easy example and increases the weight on hard inputs to promote hard-example learning. We point out that our method is only conceptually related to these works via input hardness. Our method helps the student with hard examples by providing explicitly discounted help g. We learn the helper function g (through teacher representation) that decides whether the student needs help on a given input. Thus, we do not prioritize learning hard examples keeping in mind the fact that student capacity is much smaller than the teacher.

6 Conclusion

We develop a new knowledge distillation method that utilizes teacher predictions in novel ways combining predictive targets with scaffolding the student on hard-to-learn points by means of a guide function. Our method is particularly relevant when there is large gap between student and teacher capacities. We show that our method allows for convergence to better minima based on two key properties. Our guide function allows for censoring hard-to-learn examples, and the predictive targets set by the teacher on remaining points allow for eliminating bad local minima and smoothening the resulting student loss landscape. Against vanilla KD we achieve improved performance and are comparable to more intrusive techniques that leverage feature matching on benchmark datasets.

REFERENCES

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/ea8fcd92d59581717e06eb187f10666d-Paper.pdf.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent, 2021. URL https://arxiv.org/abs/2106.05237.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933395. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7028–7036, 2021.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11933–11942, 2022.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801, 2019. doi: 10.1109/ICCV.2019.00489.

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, mar 2021. doi: 10.1007/s11263-021-01453-z. URL https://doi.org/10.1007%2Fs11263-021-01453-z.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1311–1320. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/graves17a.html.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2535–2544. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hacohen19a.html.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge, 2015.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *ICLR (Poster)*, 2016. URL http://arxiv.org/abs/1511.03643.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Adriana Romero, Samira Ebrahimi Kahou, Polytechnique Montréal, Y. Bengio, Université De Montréal, Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *in International Conference on Learning Representations (ICLR*, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

Samuel Don Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=7J-fKoXiReA.

Kaizhao Sun and Andy Sun. Dual descent alm and admm, 2021. URL https://arxiv.org/abs/2109.13214.

F. Tung and G. Mori. Similarity-preserving knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00145. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00145.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. URL http://jmlr.org/papers/v16/vapnik15b.html.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. URL https://arxiv.org/abs/1612.03928.

Xinchuan Zeng and Tony R. Martinez. Using a neural networks to approximate an ensemble of classifiers. In *Neural Processing Letters*, pp. 2000, 2000.

Tianyi Zhou, Shengjie Wang, and Jeffrey Bilmes. Curriculum learning by dynamic instance hardness. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8602–8613. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/62000dee5a05a6a71de3a6127a68778a-Paper.pdf.