Table 1: Ablation study of proposed ideas for knowledge distillation.

Ablation	Accuracy
Baseline	11.07%
No ideas	$27.41 \pm 4.76\%$
Idea 1	$55.10 \pm 2.49\%$
Idea 2	$44.21 \pm 14.0\%$
Idea 1 & 2	$\mathbf{63.40 \pm 1.80\%}$

Table 2: Classification accuracy by the training epochs of generators.

Epochs	Accuracy
0	$23.21 \pm 1.25\%$
10	$42.52 \pm 2.19\%$
50	$52.03 \pm 2.51\%$
100	$61.70 \pm 3.94\%$
200 (ours)	$\mathbf{63.40 \pm 1.80\%}$

Table 3: Comparison between different lengths of noise variables z.

Length	Accuracy
8	$49.23 \pm 3.02\%$
10 (ours)	$\mathbf{63.40 \pm 1.80\%}$
12	$60.65 \pm 1.29\%$
16	$61.39 \pm 2.56\%$
20	$59.29 \pm 0.84\%$

Thank you for the detailed reviews. We address your comments and attach additional experimental results. We group the issues based on the topics and show the related reviewers for clarification: for instance, R1 denotes Reviewer 1. All experiments in this letter have been done for Student 3 on the SVHN dataset.

Tucker decomposition for initialization (R1, R3). We initialize the student networks by Tucker-2 decomposition. Specifically, we take three steps to train each student network: we 1) initialize the weights by running SVD on the teacher networks, 2) update them by Tucker-2 to minimize the reconstruction errors, and 3) fine-tune them by artificial data points from the generators. Since applying step 1 alone produces low accuracy, we take it as a baseline to apply steps 1 and 2 together and report it in the paper as *Tucker (T)*. If we apply step 3 alone without steps 1 and 2, as Reviewer 1 suggested as *KegNet + random init*, the accuracy is not as good as shown in the paper since the student networks have no prior knowledge about the learned weights of the teachers. It is a challenging problem to apply KegNet without initializing the student networks by Tucker-2, as it is more difficult to train them properly.

Soft labels by unnormalized distributions (R1, R3). We propose two ideas in line 189 to improve the performance of knowledge distillation. The first is to use multiple generators when generating artificial data points. The second is to sample label vectors from the elementwise uniform distribution; instead of using a typical one-hot vector or a categorical distribution as a label vector $\hat{y}$, we sample each element independently from $\mathrm{uniform}(0, 1)$ and create an unnormalized probability vector as an input label. As a result, we do not impose any correlation between the different classes but generate diverse data points that cover a larger manifold. Table 1 shows the result of ablation study of these ideas. Currently, even a simple idea is enough to improve the performance of our model by a large margin, but we may apply a more principled approach to achieve a similar objective following the suggestion of Reviewer 1.

Performance improvements during training (R3). Reviewer 3 commented that the superiority of our approach may have come from the structural prior imposed by CNN-based generators. To address the concern, we report the accuracy of student networks, coupled with various generators trained for different numbers of epochs. Table 2 clearly shows that it is essential to train enough the generators to get a superior performance; this is because the artificial data generated from random generators are not close enough to the true data manifold which we aim to estimate.

The length of noise variables (R2). The noise variable z has been designed to follow the class-independent manifold of the data distribution p_x; compare it with y which embeds class-dependent manifold of the distribution. Thus, it is reasonable to assume that z lies in a low-dimensional embedding space as done in previous works [25]. At the same time, it is important to choose a proper length of z as it determines the capacity and learning complexity of our model. Table 3 compares the accuracy of student networks trained with noise variables of different lengths. Accuracy is the best when the length is 10 as in the paper, but the differences between different lengths are negligible compared with the other experiments in Tables 1 and 2; this implies that our approach is not very sensitive to the length of z.

More complex datasets (R1, R2). We ran additional experiments for other datasets such as CIFAR10 and CIFAR100 which are larger and more complex than the datasets that we used in the paper. As a result, we have checked that it is challenging to achieve a good performance on these datasets, because they have more complex data manifolds which are difficult to be estimated by our simple KegNet structure. It seems that a more complex architecture is needed to capture such a complex manifold, and thus we leave it as an open problem for future works.

Minor points (R1, R2, R3). (R1) We have typing errors in lines 245 and 246; we used ResNet14 instead of ResNet20. (R1) We compressed only the convolutional layers in the teacher networks as described in line 252 and did not touch the dense layers, based on previous works on compressing CNNs by Tucker-2 [13]. (R2) Our objective is to distill the knowledge of a neural network in the absence of training data. This is done by generating artificial data that follow a similar manifold to the unseen data, and the result can be used for various applications such as model compression, interpretation, or knowledge transfer especially when the data are not accessible. (R3) We had visualized the generated images for Fashion MNIST, but it was difficult to recognize the images because they described ambiguous clothes, hats or shoes rather than clear digits. It is a future work to make the model generate more recognizable images.