We sincerely thank the anonymous reviewers for their supports and constructive comments.

Response to R1:

Q1: Compare with vanilla ANN with dropout. A1: The classification accuracy is improved not only because of the better generalization ability. The gradient used in vanilla ANN is L2-gradient rather than L1-gradient which is not accurate. The knowledge from CNN will help ANN to find the correct optimization direction. Only adding dropout to vanilla ANN will not solve this problem. The experimental results on CIFAR-10 are shown in Table 1 column 2 and 4.

Q2: Does PKKD change the inference process? A2: As shown in Line-150 in the main paper, the inference process is exactly the same as the vanilla ANN, which means that the proposed method will not introduce extra multiplication operations during inference. We will further emphasize this in the final version.

Q3: Apply the method to the traditional CNN distillation methods. A3: It is really a good question. We believe that the proposed PKKD method is also useful for the distillation between traditional CNNs. When the architectures of the two networks are different, their 'ground-truth' weight distributions and feature maps of intermediate layers are also different. The experimental results of distilling ResNet-18 with ResNet-152 are shown in Table 4.

Response to R2:

Q1: Using CNN rather than bigger ANN to distill. A1: Theoretically we can use bigger ANN to distill smaller ANN. However, it is not always possible to find a bigger teacher. Instead, we can always find a homogeneous CNN as teacher model. Besides, ANN uses L2-gradient to replace L1-gradient when doing back-propagation, which may lead to a wrong optimization direction. Using CNN rather than bigger ANN to distill can alleviate the problem.

Q2: Compare with other self-distill methods. A2: The experimental results using the proposed method and method of snapshot KD on CIFAR-10 is shown in Table 1 column 2 and 3, which shows the superiority of the proposed method.

Table 1: Compare with self-distill methods and dropout.

	PKKD	Snapshot KD	ANN + dropout
VGG-small	**95.03**	93.95	93.88
ResNet-20	**92.96**	92.33	92.20
ResNet-32	**93.62**	93.17	93.09

Table 2: Operations in different networks.

Model	Method	#Mul.	#Add.	XNOR
ResNet-18	CNN	1.8G	1.8G	0
	ANN	0.1G	3.5G	0
	BNN	0.1G	1.8G	1.7G

Q3: Moving experiment results to the main paper. A3: The results will be placed in the main paper in final version.

Response to R3:

Q1: Number of #Mul. A1: Thank you for pointing out the problem. We follow the vanilla ANN setting [3] and '#Mul' was zero in that paper. The actual numbers are listed in Table 2. And we will fix it in the final version.

Q2: Compare with other SOTA BNNs. A2: The top-1 accuracies of PKKD-ANN and PCNN of ResNet-18 on ImageNet are **68.8** and 57.3. We will include these results in the final version.

Q3: Extra training time. A3: The training time of using traditional KD and PKKD for ResNet-18 on ImageNet is 40m 33s and 59m 45s per epoch.

Table 3: Compare with [R1], [25] and [26] on CIFAR-10.

	PKKD	[R1]	[25]	[26]
ResNet-20	**92.96**	92.38	92.22	92.27

Table 4: PKKD and KD in CNN distillation.

ResNet-18	69.8/89.1
PKKD	**73.1/91.3**
Traditional KD	72.5/90.9

Q4: Typos. A4: We will fix all the typos in the final version.

Response to R4: All the following experiments and references will be included in the final version.

Q1: Effectiveness of kernel method. Suggest comparing on ImageNet. A1: To alleviate the influence of stochastic initialization, we have reported the mean accuracies of **5** different runs in Tab.1 in the main paper, which means the results are convincing. We will emphasize this in the final version. Results on ImageNet with ResNet-50 are shown in Table 5. We run 40 epochs due to the limited rebuttal period, but still shows the priority of PKKD.

Table 5: Ablation study on ImageNet with ResNet-50. 'K / NK' stands for using kernel or not. 'P / NP' stands for using progressive or fixed teacher.

CNN	ANN	K + P (PKKD)	NK + P	K + NP	NK + NP
73.2/91.6	70.4/89.9	**73.4/91.7**	72.5/90.9	72.0/90.6	71.3/90.4

Q2:Compare with existing methods. A2:We compare PKKD with [R1], [25] and [26] in Table 3 on CIFAR10. PKKD is better since it focus on alleviating the difference of distributions between CNN and ANN, which is the key problem.

Q3: Reference FitNets. Difference from FitNets. A3: We will reference FitNets in final version. The difference from FitNets is that FitNets use 1x1 Conv to make sure the size of teacher feature maps equal to student, so that the KD method can be applied correctly. In PKKD, the original size of feature maps from ANN and CNN are already the same. Thus, the purpose of using kernel (non-linear transformation, which is also different from FitNet) is to alleviate the difference of two distributions by mapping them to infinite dimensional space, and 1x1 Conv is further used to align the features in the new space derived from kernel. The difference has been briefly mentioned at the beginning of Section 3.1. More details will be included in the final version.

[R1] F. Tung G. Mori. Similarity-Preserving Knowledge Distillation. ICCV, 2019.