The authors sincerely thank all the reviewers for their very constructive and helpful comments.

**Response to Reviewer #1:**

Table 1: Use pruning (Z Liu 2017) on top of JointRD on CIFAR100.

| Model | ResNet-34 | | | plainCNN-34 (JointRD) | | | ResNet-50 | | | plainCNN-50 (JointRD) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pruning Rate (%) | 0 | 30 | 60 | 0 | 30 | 60 | 0 | 30 | 60 | 0 | 30 | 60 |
| Accuracy (%) | 78.42 | 78.52 | 75.71 | 78.47 | 78.55 | 75.38 | 78.39 | 78.56 | 77.20 | 78.16 | 78.17 | 77.03 |

Table 2: Use KD(logits)(Hinton 2015) on top of JointRD

| DataSet | Teacher: | Student | Baseline (%) | KD (%) |
|---|---|---|---|---|
| CIFAR100 | ResNet-50:<br>78.39% | ResNet-18 | 77.92 | 78.67 |
| | | plain-CNN 18 | 77.91 | 79.05 |
| | | ResNet-34 | 78.58 | 78.98 |
| | | plain-CNN 34 | 78.47 | 79.23 |

**Q:** *Why JointRD instead of pruning and KD.* **A:** Thanks for your valuable comments. We would like to highlight that our method can be used on top of either pruning or KD instead of being a substitution of them. The results in Table 1 and 2 show that after removing the shortcut, the network can also benefit from pruning or KD as much as the original ResNet.

Table 3: Compare with smaller network obtained by pruning and KD: results on ImageNet. Prune: 40% of the channels are pruned.

| model | memory(kb) | latency(ms) | Baseline (%) | KD (%) |
|---|---|---|---|---|
| ResNet50 | 242194 | 166.71 | 76.11 | Teacher |
| Prune 40% | 222642 | 152.86 | 74.68 | 75.13 |
| plainCNN50 | **222182** | **137.45** | **76.08** | **76.32** |

As per your suggestion, we have conducted the filter pruning and KD experiments on ImageNet dataset in Table 3. The performance of plain CNN 50 model trained by the proposed JointRD, is better than the pruned ResNet 50 concerning memory, latency and accuracy, for both cases with or without using KD.

**Q:** *L107-111 gradient.* **A:** In line 171, we reported that a cosine annealing policy is used to decay the penalty factor of the losses from the teacher network. Empirically, this decay policy works better than a constant one: 78.16% verses 74.23% for CIFAR100 on plain-CNN 50. **Q:** *Ablation studies.* **A:** Thanks. For CIFAR100 on plain-CNN 50, the accuracy of original setting, path 1 removed, path 2 removed, and path 3 removed are 78.16%, 75.54%, 76.37%, and 73.74% respectively. **Q:** *Minor comments.* **A:** Thanks. We will correct these typos and proofread the manuscript to make it more readable.

**Response to Reviewer #3:**

**Q:** *Usage of Equation 2.* **A:** In our case where the student has the same number of channels as the teacher, the transformation is used to loosen the constraint of channel-wise Mean-Squared Loss, where only a transformation of the student channel-wise features are required to align with the teacher.

**Q:** *Ablation study.* **A:** Thanks for this nice concern. For the results provided in Table 7, the Dirac initialization are not used, we would refine this table and sentences around it to have an explicit illustration. The performance of using KD together with Dirac is provided in Table 3 and Table 4 instead (the KD(MSE)+Dirac column). As we can see, the proposed method also brings significant benefit over KD(MSE)+Dirac.

**Response to Reviewer #4:**

**Q:** *Comparison to related work.* **A:** Thanks for this constructive comments. The most significant difference of the proposed teacher-student framework from the existing knowledge distillation is the use of the gradients from the teacher models during the training process. Classic knowledge distillation (KD) work either impose loss terms to force the student to learn similar classification soft-label or feature maps like the teacher. Different from the existing methods, our framework allows the student network to use the gradients calculated from the teacher network during optimization. In other words, we not only guide the student by informing them the target points but also provide them with step-by-step direction guidance for them to find the way to the target points. As there are multiple paths in the framework, we also proposed an effective forward and backward process for these paths. These gradients from the teacher model turn out to be very important to achieve good accuracy, compared with only using KD (logits)(Hinton 2015) or KD(MSE)(B Heo 2019): 78.16% verses 70.93% and 63.82% for CIFAR100 on plain-CNN 50 (Table 10). In addition, as shown in Table 1 and 2, our method can be used on top of both pruning and KD(logits). We will include more discussion in the final version.

**Q:** *Small models and small tasks.* **A:** Thanks for this nice concern. The shortcut is explored for avoiding the gradient vanishing for training very deep neural networks on large datasets. For small tasks such as CIFAR100, the accuracy of ResNet-18 and plain CNN18 (naive training) is 77.92% and 77.44%, respectively. Thus, the accuracy improvement using our method is subtle on these shallow models. In contrast, for the ImageNet benchmark, the plain CNN-50 learned using our approach achieves a 76.08% top1-acc with an about 18% latency reduction. We will emphasize this issue and include more discussions in the final version.