Table 1: Comparison with other KD methods on CIFAR100.

Model	Baseline	VID	AT	FSP	Jacobian	SVD	Heo	Ours
ResNet18	77.09	78.93	78.45	78.75	78.45	78.53	79.21	**82.92**
ResNet50	77.42	79.21	78.73	79.02	79.03	78.82	79.72	**84.74**
MobileNetV2	69.04	70.62	70.34	70.48	70.26	69.35	71.02	**73.57**

Table 2: Comparison experiments on ImageNet.

Model	Baseline	KD	FitNet	Self-Distill	DML	Ours
ResNet18	69.76	70.45	70.26	70.51	70.39	**70.92**
MobileNetV2	71.52	72.23	71.95	72.37	72.29	**72.81**
ShuffleNetV2	69.36	70.14	69.86	70.26	70.21	**70.74**

We sincerely appreciate all reviewers for their most detailed and thoughtful reviews.

Overall Response: Comparison with more KD methods (R1'Question-3 and R4'Question-2). Tab.1 shows the comparison with additional 6 kinds of KD methods, including the VID (Ahn *et al.*, CVPR2019) requested by R1, and 5 kinds of KD methods (from Tab.1 of the paper) requested by R4. It is observed that our method outperforms the second-best KD method by 3.71%, 5.02%, 2.55% on ResNet18, ResNet50, MobileNetV2 @ CIFAR100, respectively.

Response to R1: Question-1 and Question-2. Tab.2 shows the comparison experiments on ImageNet with ResNet18, MobileNetV2 and ShuffleNetV2. It is observed that our method leads to 1.16%, 1.29%, 1.38% accuracy improvements on ImageNet with ResNet18, MobileNetV2, ShuffleNetv2 respectively, outperforming the second-best KD method by 0.41%, 0.44%, 0.48%, respectively. **Question-3.** We have re-conducted experiments on ModelNet10/40 with ResGCN8 and each experiment is repeated 5 times. On ModelNet10, the accuracy of our method and DML are 94.35 $\pm 0.21\%$ and $93.63 \pm 0.17\%$, respectively. On ModelNet40, the the accuracy of our method and DML are $91.78 \pm 0.12\%$ and $91.67 \pm 0.09\%$, respectively. Note that DML is a logit-based KD method while the orthogonal loss can only be utilized in feature-based KD methods so DML can not be improved by the orthogonal loss. **Question-4.** Please refer to the overall response (the VID Column). **Question-5.** We do not introduce new hyper-parameters for balancing these two losses because we find that the fixed 1:1 loss ratio is good enough. Additional experiments show that the performance of our method can be improved by 0.15% by introducing and adjusting a new hyper-parameter here (ResNet18, CIFAR100). **Question-6.** The proposed method has given strong experimental clue to the intuition that *task-oriented features are more essential in the knowledge transfer related domains.* These improvements might open up new research opportunities for exploring the transfer-based ML including mathematical theory and algorithm. **Question-7.** Thanks for your advice, we will remove this sentence in the final version.

Response to R4: Reply to additional feedback. The training loss for KD baselines includes both distillation loss and the supervised loss. Thanks for the valuable question and we will modify our writing to clarify this confusion. **Question-1.** We apply both supervised loss and distillation loss to the auxiliary classifiers instead of only supervised loss because the distillation loss can facilitate the training of auxiliary classifiers. A more accurate auxiliary classifier is able to extract better task-oriented features, which improves the performance of knowledge distillation in turn. *How to Avoid the Case?* The case that auxiliary classifiers have learned to minimize their loss but the backbone model fails to learn useful features will not happen because (i) Compared with the whole backbone, the auxiliary classifiers have much fewer parameters so they have weaker learning ability than the backbone model. (ii) The auxiliary classifiers are trained to minimize both distillation loss and supervised loss, which is harder than only the supervised loss. (iii) In the overall loss function, we multiply the loss of auxiliary classifiers by hyper-parameters whose value is less than 1 ($\alpha = 0.03$, $\beta = 0.5$), which forces the whole model to pay more attention to the learning of the backbone and the final classifier, instead of the auxiliary classifiers. **Question-2.** With the same or fewer parameters in the training period, our method still outperforms the other KD methods. For instance, ResNet18 with our method has 12.10M and 11.22M parameters in the training and testing period respectively, achieving 82.92% accuracy in the testing period. In contrast, ResNet50 with Hinto's KD has 23.70M parameters in both the training and testing period, achieving 78.58% accuracy in the testing period. Note that the parameters of auxiliary classifiers are very tiny (only 7% of the whole model, on ResNet18). **Question-3.** We have added comparison experiments with them. Please refer to the overall response.

Response to R5: The proposed method has given strong experimental clue to the intuition that *task-oriented features are more essential in the knowledge transfer related domains.* These improvements might open up new research opportunities for exploring the transfer-based ML including mathematical theory and algorithm.

Response to R6: Question-1. This question can be considered in two aspects. First, when the feature resizing layer is used to increases the dimension of student features, due to the energy-preserving property of orthogonality, the energy of student features will not be amplified by the feature resizing layer. Second, when the feature resizing layer is used to decrease the dimension of teacher features, the weight of feature resizing layer works as an orthogonal projection to the teacher features, which leads to less information loss than non-orthogonal projection. **Question-2.** As shown in the ablation study (Tab.7), compared with naive feature distillation, the task loss brings 3.50% accuracy improvement, which accounts for 80% of the total accuracy gain, indicating that the task loss is the key of our method.