To Reviewer 1: Thank you for your positive comments. We have investigated the effect of the disagreement parameter C in Section 4.3. Actually, we also did similar experiments on other teacher-student networks and ensemble sizes (20 and 25). The observations are also similar. We will include these results to improve the comprehensiveness of our ablation studies. Also, we will cover more references to enrich the related work.

To Reviewer 2: Thank you for your detailed comments and suggestions.

Re: motivation. You may misunderstand our motivation. Our AE-KD actually **does not** treat every teacher in the ensemble equally. On the contrary, we introduce a parameter C to allow disagreement among teachers, so that the obtained update direction of parameters is not necessarily a strict descent direction for all teachers. As you suggested, the gradient directions from weak or noisy teachers are not as reliable as those from good teachers (also see line 161-166). With the help of parameter C, the final direction will not accommodate these weak teachers necessarily, and a better weight over teachers can be automatically determined by solving Problem (9) or (11).

Re: comparison with OKDDip. The main differences lie in three ways. First, the learning paradigms are different. OKDDip is essentially an online KD paradigm which adapts two-level distillation. While AE-KD follows traditional teacher-student paradigm and the teachers won't be updated during the training. Next, we use different strategies to learn the weights. In the second-level distillation of OKDDip, group knowledge is transferred to the group leader, where all diverse peers serve as a group of teachers. The weights for diverse peers are computed using self-attention mechanism. In AE-KD, the weights for teachers are computed based on multi-objective optimization in gradient space. We take disagreement among teachers into consideration and prevent the student suffering from adverse guidance. Last, AE-KD has the tunable parameter of disagreement to reconcile all teachers while OKDDip doesn't.

Re: optimization of weights α_m in Eq. (11). Eq. (11) is solved for every minibatch with the calculated gradients over all teacher losses. It is a typical One-class SVM problem with constraint $0 \leq \alpha_m \leq C$, and can be easily solved by LIBSVM or other off-the-shelf solvers.

Re: 350 epochs for resnet20. In experiments, we train all the teachers for standard 240 epochs, following the setting in [1]. And for student resnet20, we train for 350 epochs for better performance.

Re: ensemble networks with various architectures. Our AE-KD performs in the same way, *i.e.*, every teacher provides a gradient and the final direction is computed according to Eq.(9). To validate this case, we take resnet20 as student, and use resnet110, wide_resnet_40_2 and vgg13 as teacher networks with accuracy 74.31%, 75.61% and 74.64%, respectively. Results show that our AE-KD can achieve 70.16% accuracy while our baseline method AVER only has 69.40%, showing our superiority of dealing ensemble KD even with various architectures. We will include this setting in our final version.

To Reviewer 4: Thank you for your detailed comments and constructive suggestions for our experiments.

Re: Section 2. It intended to formally illustrate related work and baseline methods. We will consider your advice.

Re: issues about experiments.

1: multiple runs. Thanks for your suggestion. We experimented with five reset56 teacher networks and the resenet20 student network on CIFAR10, and run our AE-KD for 10 times with different random seeds. The mean and standard variance are 92.49% and 0.02%, respectively. We can see the performance of student network tends to be steady (low variance) due to the distillation from teacher networks. We will cover this in our final version.

2: performance gap between AE-KD* and AE-KD on CIFAR10/100. The performance gap comes from the weight β in Eq.(7) of the feature-based loss in AE-KD*. In real implementation, for CIFAR10 we determine the optimal β by cross-validation in $\{10^{-1}$,1,10,100, 1000$\}$, and for simplicity, we solely adopt the same β for CIFAR100, thus the performance of AE-KD* on CIFAR100 can drop a bit than AE-KD since we do not tune β specially for CIFAR100. However, what we emphasize here is that given a KD method (logits or feature based), when it comes to the ensemble setting, our method can develop a better way to distill from their ensemble by adaptively assigning weights. Of course, suitable parameters bring in better results. For example, if we tune β specially for CIFAR100, our AE-KD* can have 71.95% accuracy on resnet20 student network with five resnet56 teacher networks, surpassing AE-KD (71.37%).

3: experiments with more teachers. With more teachers involved, the proportion of weak teachers will tend to be steady. Thus for AVER, their influence on the performance will gradually reduce to a certain level, and the accuracy will increase dramatically with more teachers at first, but stabilize eventually as Figure 2. However, our method remains steady and robust for most time. We experiment with 30, 40, 50 teachers on CIFAR10. The performances of AVER and AE-KD are 92.59%, 92.67%, 92.69% and 92.77%, 92.82%, 92.86%, respectively. AE-KD still outperforms AVER with a comparable and stable gap for large ensemble size.

4: weak teacher models. Our teacher models follow those in [1] with the same or similar accuracy. Admittedly, stronger teacher models do produce better students. We choose network from [2] on CIFAR10 as teacher models with 97.37% accuracy and the performance of AE-KD on resnet20 student network reaches 95.14% compared to AVER (93.66%). The results are even better and AE-KD retains its superiority.

5: line 244. Yes. It's a typo. Thank you for pointing it out and we will fix it in our final version.

[1] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. ICLR 2020.
[2] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. ICLR 2019.