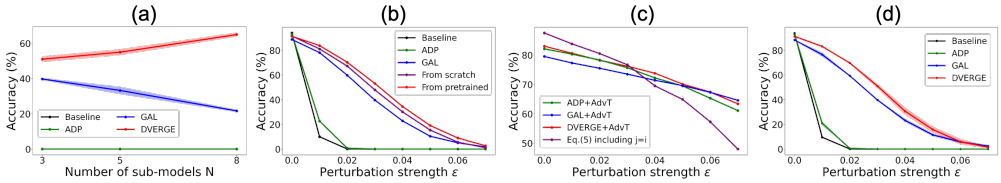

(e)

Method	Training time (h)
Baseline	1.0
ADP	2.0
GAL	7.5
DVERGE	10.5
AdvT	11.5

(a) Black-box robustness at ϵ of 0.03 vs. # sub-models; **(b)** Training DVERGE from scratch; **(c)** Baselines with AdvT and Eqn (5) including $j = i$; **(d)** Black-box robustness with error bars; **(e)** Training time comparison on a single TITAN Xp GPU for 200 epochs. Except (a), all results are obtained on ensembles with three ResNet-20s with the same setup as in the paper.

Q1: Computational overhead (R1, R2 & R4) DVERGE induces similar computational cost as AdvT. They both need extra back propagations to either distill non-robust features or find adv. examples. However, DVERGE uses only intermediate features for distillation so it is marginally faster than AdvT, as shown in Tab (e). Though ADP requires the least time budget, it does not improve the robustness much as shown in Fig 4. And the significantly improved robustness would be worth the extra training cost of DVERGE over GAL.

Q2: Scalability under distributed setup (R1) As indicated in Eqn (5) and Alg 1, the feature distillation and loss computation of each sub-model are separated, so model parallelism is well supported. The only communication required between sub-models would be transferring distilled images, which can be done in an asynchronous way. Exploring asynchronous distributed training will be an interesting future work for DVERGE.

Q3: Effect on clean accuracy (R1 & R2) Traditional ensemble methods diversify sub-models' prediction for high accuracy. DVERGE diversifies the *adv. vulnerability* of sub-models, which enhances the robustness to transfer attacks, but not necessarily improves the clean accuracy. Fig 3 shows DVERGE diversifies the vulnerabilities of sub-models the best. Combining DVERGE with prediction-based diversity metrics will be studied in the future.

Q4: Robustness vs. the number of models N (R1 & R2) Black-box robustness under a fixed attack strength ϵ vs. # sub-models N is given in Fig (a). Compared with others, DVERGE has a clear and steady increase in robustness with larger N. Finding orthogonal features for an arbitrarily large N is difficult intuitively, so robustness will saturate at some point, which will also be the case for other methods. When facing ImageNet with more complex models, it could take larger N to saturate as the feature space will have a larger capacity that makes the diversification of features easier.

Q5: Ablation study on training from scratch and fixing the layer (R2) We train DVERGE from a pre-trained ensemble based on the intuition that well-learnt features of the pre-trained models are more informative for distillation and diversification. Fig (b) shows that training from scratch still offers improved robustness over others. Thanks for the alternative design choice. As to the choice of the layer, training with only layer 7, 13, or 20 yield 2.5%-20.4% black-box robustness drop (ϵ=0.03) compared with random layer sampling. We will conduct a rigorous study in the future.

Q6: Additional references on ensemble (R3) We thank the reviewer for new references and will cite them in the revision. The given ref [4] has a similar motivation to DVERGE, yet the two works diversify the sub-models from distinct perspectives. Our distilled images are in the input space, which serve as transfer adv. attacks targeting the vulnerability within the whole sub-model, while [4] directly diversifies the latent feature of a specific layer. We believe DVERGE can better diversify each sub-model's vulnerability and lead to significant robustness improvement.

Q7: ADP and GAL + AdvT (R3) Fig (c) shows the results of ADP and GAL with AdvT. DVERGE+AdvT remains the best across the majority of the ϵ spectrum. It is not surprising to see DVERGE does not yield significant improvement over others when incorporating AdvT, as AdvT will force each model to learn a similar set of robust features, which will leave less capacity to capture diverse non-robust features, as discussed in line 285-292.

Q8: Error bars in the plots (R3) Standard Deviation over 3 runs are included in Fig (a) and (d). DVERGE has a little higher variation, potentially due to the random distillation layer selected in the last training epoch. Yet superior black-box robustness can still be observed. We will also include error bars in all other plots in the revision.

Q9: Additional questions from R3 We thank the reviewer for the suggestion on Alg 1 and Fig 4 and will update them in the revision. We use ResNet-20 as it is a std. arch. for CIFAR-10 (Sec 4.2 of He et al. *CVPR'16*) and both ADP and GAL adopt it. Bayesian NN is out of the scope of this paper, yet it is likely that DVERGE imposes different feature priors on sub-models. The # of data modes available in the non-robust feature space might provide a bound on the ensemble size, which could be a promising direction to derive a better theoretical understanding of DVERGE.

Q10: Difference with AdvT (R4) DVERGE is fundamentally different from AdvT: AdvT exclusively promotes the learning of robust features; DVERGE allows the learning of non-robust features (detailed explanation see line 166-171). When $j = i$ is allowed in Eqn (5), the model will also be trained with distilled images (adv. examples) generated against itself, which will have a similar effect to Eqn (6). We empirically study this case by training an ensemble with the objective as Eqn (5) without the $j \neq i$ constraint. As shown in Fig (c), it gives higher clean accuracy than DVERGE+AdvT, but the black-box robust accuracy degenerates faster. This may result from the difference between training with distilled images (*targeted* attacks that minimize the distillation objective) and using PGD (*untargeted* attacks that maximize the classification loss). We will discuss more on this in the revision.

Q11: Additional questions from R4 Note the "sensitivity" appears only for white-box robustness, which is not our main focus (line 463-468). So we leave its exploration to future works. Meanwhile, training with larger ϵ consistently leads to a higher black-box transfer robustness as desired. We thank the reviewer for the extra baselines. The focus of DVERGE is to promote a diversity metric to enhance the robustness of the ensemble, so we focus on comparing with ensemble diversity training methods. We believe the individual model-based defenses in the given ref [a] are orthogonal to DVERGE, and incorporating these techniques may bring further robustness improvement in the future. The color in Fig 1 indicates the prediction label, and we will make it clear in the revision.