We thank all reviewers for the comments. The code will be released if our paper can be accepted.

R1: **Unrealistic Clinical Scenario and More Baseline:** We agree with the reviewer that in practice we may only have one domain for training. To this end, we consider HAM10000 for training and the others for testing for skin lesion classification task. Besides directly training with classification loss (DeepAll), we also compare with CCSA and MixUp. Noted that other DG baselines are not applicable in this case, as they require multiple domains available to simulate domain shift. The results are shown in Table 1. In most of the cases, our proposed method can outperform the DeepAll, CCSA as well as Mixup. For CCSA, as the contrastive loss is applied on only one source domain while target domain is unseen, it is likely to suffer from overfitting problem. For Mixup, the combination is conducted only in a CONVEX manner, which may not be able to generalize well to the out-of-distribution target domain.

Last but not the least, it is likely that we can have data from multiple domains for training. For example, different hospitals can collaborate to build an AI system for medical image classification.

Others: We thank the reviewer for pointing out our typos. We will revise them in our revised manuscript. We agree with the reviewer that standard error is more suitable. Previously, we follow MASF [7] to report std value.

R2: **Discussion of Linear Combination for Domain Adaptation:** We thank the reviewer for the reference suggestion. The major difference between our work and the works suggested by reviewer are two folds, 1) we focus on domain generalization where target domain is not available while the reference works focused on domain adaptation where target domain is available; 2) we do not conduct combination in an explicit manner but only assume linear dependency among samples with same category. Moreover, these two works only consider combination in a CONVEX manner where target domain is available, which may not be able to handle the out-of-distribution testing domain which is unseen during training (unlike domain adaptation). We will add the discussion in our revised manuscript.

KL: We agree with the reviewer adopting KL is quite common and the reasons why we adopt KL are two folds: 1) it can be optimized by using reparametric trick, which is simple and effective, 2) we find other metrics (e.g. JS divergence) may not be able to outperform KL, such observation is consistent with the finding in MASF [7].

Simple Combination of [37]: We do not agree with the reviewer that our method is a simple combination of [37] and KL. We discuss [37] in both Section 1 and 3 that [37] simply conducted image augmentation in the pixel level, while we introduce a novel rank regularization, which is quite different. We also discuss the results of [37] in our supplementary material and give sufficient analysis.

"Heterogeneous" in Abstract and Domain Gap: The word "heterogeneity" is to describe the difference between source and target domain, which was adopted in [37]. Regarding the domain gap, it is very common in the field of domain generalization to use the same backbone for all the domains for the SAME task (e.g. the literatures in related works). On the other hand, although the variation of domains is relatively small for medical data, we may not get sufficient data during training, which can also have negative impact to the generalization capability.

Why Outperform [7,21]: We have shown the rationale in both Section 1 and 4 that simply conducting meta-learning/divergence minimizing on source domain may not be desired as it can lead to overfitting problem to the source domain, while target domain is not available during training stage. We introduce linear dependency, which is inspired by the augmentation technique proposed in [37], such that better generalization capability can be observed.

R3 & R4: **Setting Details:** We have shown the experimental details in our supplementary material. For low rank constraint, We replace the rank regularization with a nuclear norm, which is a popular way for implementation.

R4: **Rank:** We agree with the reviewer that imposing rank regularization may suppress more information. However, in our manuscript, we expect that the latent features (the output second-last FC layer which is deep) to contain more category information instead of others. Based on Figure 2, we find that the best performance can be observed when setting $rank = C$. Moreover, the performance drops when increasing the number of rank, which means the information, which may not benefit domain generalization, can be contained. We also compare with the traditional low rank method (LR) where the LR constraint is imposed on the same feature space. Last but not the least, we have shown in Figure 3 the relationship between epoch and rank value and the results further justify our formulation. How to select the rank for the feature in other layers will be investigated in our future work.

Table 1: Domain generalization results with HAM10000 as source domain.

	DMF	D7P	MSK	PH2	SON	UDA	Average
DeepAll	0.3003	0.4972	0.1667	0.4945	0.5025	0.4945	0.4093
CCSA	0.2762	0.5082	0.4652	0.4667	0.5275	0.5055	0.4582
MixUp	**0.3514**	0.4029	0.3000	0.4333	0.6296	0.4615	0.4298
Ours	0.2943	**0.5191**	**0.5087**	**0.5500**	**0.6949**	**0.5714**	**0.5231**