We thank the reviewers for their detailed comments and questions. Typos are fixed.

R1 **Line 137-138**: In these experiments we change the size of the shuffled blocks all the way to 1 and even try shuffling the channels of the input. Therefore, the only object that is preserved here is the set of all pixel values which can be looked at as a histogram/distribution and be fully characterized by its moments. We have now clarified this in the paper. **Line 179-188**: It has been observed in the literature that *any* two neural network minimizers can be connected via a non-linear low-loss path. In contrast, as mentioned by the reviewer, due to non-linear nature of neural nets one would not expect the model obtained by the linear interpolation of two neural nets in the parameter space to perform well, and that is precisely why we find this phenomenon interesting and investigate it. Note that insights from this observation can be used for improving ensemble methods. **Line 108-109**: One can come up with quantitative measure by comparing domains higher level representations by passing them through a network but here what we meant was looking at data itself, and we have included some samples in the paper. Real includes real images, while Clipart is a cartoon version, both have colors and texture. On the other hand ChexPert is black and white and has texture while Quickdraw has no color or texture. In light of addressing your main concerns, we respectfully ask you to consider accepting the paper.

R2 **Back n forth to Appendix**: Thank you for your suggestions. We have now moved the distance in feature space results to the main part of the paper and moved the 'train' part of figure 5 to the Appendix. **L151-152**: We mean ambiguous data points in the target domain. We will revise the text to clarify this. **L191-192**: Two P-T models in the same basin are not trivially (approximately) equal, as higher layers compute quite different features (Table 1). We agree that using interpolation coefficient as x-axis does not tell the whole story. But with ReLU you can artificially scale up weight magnitudes and distances. It's the connectivity rather than the actual distance that is important here. Regarding RI-T having different random seeds, we added experiments of different RI-T models from *the same* random initialization values. They still converge to *different disconnected* basins. We also added extrapolation (with coefficient in [-1, 2]) to our connectivity analysis. **Spectrum section**: Thank you for your question, what we meant is among two models that can classify with certain margin, which translates to having low cross-entropy loss, then we can look at concentration of spectrum and concentration towards low values shows less confidence. More mathematically speaking, the confident model requires a lower-rank to get ϵ-approximation of the function and therefore, has lower capacity. However, the model at init (either random or pre-train) has higher loss and is not fitted to the data.

R3 **Examples of negative transfer?** Pre-training with random labels *can* hurt transfer learning [Maennel et al 2020]. **Practical usage**: Our findings on basin in the loss landscape can be used to improve ensemble methods. Our observation of higher order statistics improving training speed could lead to better network initialization methods.

R4 **Novelty**: The block-shuffling experiments and the linear interpolation of two neural networks experiments are not trivial and it was unclear what to expect beforehand. They were not done before either. Moreover, experiment outcomes matching the intuitions does not necessarily renders the experiments uninteresting or unimportant. **Claims**: We clarify that we mention in the paper "lower layers are in charge of more general features" (comparing to more class-specific features in higher layers) and we do not claim they are in charge of feature reuse. (line 240-242) **Explain methods used**: CKA[Kornblith et al] is the latest work on estimating feature similarity with superior performance over earlier works. The algorithm in [Sedghi et al] is the one that provides correct singular values for convolutions and has been used for various applications. We cited the works and referred the reader to original papers for details for space constraints. We have now added details of both methods to the Appendix. **Related work in NLP**: We have now added [Brown+2020, Devlin+2018, Tamkin+2020, Raffel+2019, Liu+2015, Radford+2019, Roberts+2020] to the related works on transfer learning in NLP. **Difference between RI-T & P-T**: The reviewer has brought up this point multiple times that better performance of P-T, P-T's being in the same basin, and the results of shuffling experiments is *expected* due to the fact that it starts from an optimized point. We need to emphasize that the pre-training model has been optimized to fit a *different* dataset (e.g. ImageNet). It would not necessarily converge faster or perform better in downstream tasks (esp. with shuffled blocks) simply because it is 'already optimized'. In fact, if the outcome is the opposite, we could perfectly justify it by saying that it is trapped in a local optima because it is 'already optimized' (to fit a different task). Moreover, optimization speed is affected differently from accuracy. **Two P-T's in the same basin being a result of large number of epochs**: We have performed many experiments and also results from Section 4 show that once the models pass the initial finetuning stage, they belong to the same basin and a large number of epochs is not needed. We also added new experiments of two RI-T models from *the same* initializations, and unlike P-T, they do *not* converge to the same basin even after many epochs. **Higher layers being less similar compared to lower layers**: The two P-T models are in the same basin but are not identical functions. Stochasticity of SGD leads to non-identical models. Lower layers are the foundation of the function and higher layers depend on lower layers. If the lower layer feature is more robust there are many solutions at higher layers and each of them can perform well. Moreover, in a ResNet model you may not be using the weights in the higher layers and may be using skip connection. Nevertheless, P-T's are more similar compared to other models (Table 1). In light of addressing your main concerns, we respectfully ask you to consider accepting the paper.