Thank you to all three reviewers for your positive and constructive feedback. Before addressing the specific points raised, we would first like to emphasise that the primary objective of this work was not to achieve a new state-of-the-art benchmark by incorporating all the latest network architectures, and carrying out extensive hyperparameter search. Instead, our contribution is primarily conceptual and explorative: we show that, without requiring any additional data, auxiliary task labels can be automatically generated to improve the performance of a primary task. This is a very different way of thinking about machine learning to existing approaches to supervised learning, and we believe that this sets the scene for a new direction from which significant future work can emerge.

– Marginal Performance Improvement (to R1, R2, R3) The major concern among all reviewers is the marginal improvement of MAXL in Table 1, even though reviewers acknowledge that the improvements are robust across datasets and network architectures, and are statistically significant. We accept this. However, improving performance when no additional data is available, is notoriously difficult. For example, improvements due to data augmentation (Mixup [Zhang et al. 2018], CutOut [DeVries et al. 2017]) and gradient manipulation (Shake-Shake regularisation [Gastaldi 2017]), all show $< 2\%$ improvement on CIFAR-10. As such, we are achieving similar improvements to various state-of-the-art regularisation techniques, without even including such regularisation. Furthermore, whilst results in Table 1 are compared to scores presented elsewhere using methods which employ significant regularisation (line 221-222), results in Figure 3 are a purer comparison, where none of the implementations use any regularisation (line 233-234). Here, we see a much more dramatic improvement of MAXL over single-task learning (around 4–6%). This shows that the performance increase due to MAXL alone, when regularisation is not an influence, is actually much greater than Table 1 may at first suggest.

– More Baselines (to R1, R3) Together with the baselines presented in this paper, we did in fact experiment with discrete VAE using gumbel softmax [Jang et al, 2017] and Prototypical Networks [30]. However, both of those methods performed only marginally above baseline *Random*, and below the *K-means* baseline. Considering K-means is a more popular unsupervised clustering approach, we decided to present only K-means in this paper to avoid overcrowding with similar baselines. However, we will add a short discussion of these other baselines experiments in the camera-ready paper.

– Weighting Coefficient / Collapsing Class Problem (to R2)

We agree with R2 that additional experiments on various weight coefficients on entropy loss can lead to a better understanding of the collapsing class problem. In this new table, we show results on CIFAR-100 with (left) and without (right) entropy loss for $\lambda = 0.2$ and 0 respectively, for all hierarchy structures. On the rightmost column, we show the test accuracy. On the second right column, we show the percentage of auxiliary labels which are actually utilised (assigned to by the label-generation network). We see that MAXL with entropy loss utilises the entire auxiliary space, and improves performance compared to using no entropy loss, because in this case, the label space is not fully utilised. We will provide a detailed explanation in the camera-ready paper.

PRI	AUX	Label %	Accuracy
3	10	1.00 | 1.00	90.50 | 90.26
3	20	1.00 | 0.65	90.65 | 90.39
3	100	1.00 | 0.35	90.66 | 90.22
10	20	1.00 | 1.00	78.40 | 77.73
10	100	1.00 | 0.57	78.46 | 78.20
20	100	1.00 | 0.61	74.27 | 73.97

– Auxiliary Hierarchy (to R2) In Table 1, we show an ablative study on the number of auxiliary classes ψ. In all cases, MAXL improved performance over single-task learning, showing that MAXL is robust to choice of this hyperparameter. Our preliminary findings do, however, show that there is no one value for ψ which consistently outperforms all others. As such, automatically determining the optimal hierarchy from training data alone presents intriguing future work in response to the promising results in this first paper.

– Parameter Sharing (to R3) There is no parameter sharing between multi-task network and label-generation network. We wanted to have a fair comparison between MAXL and other baseline methods, and thus we trained the primary task using the same network, for all methods.

– Cosine Similarity (to R3) If auxiliary task gradients have opposite direction to the primary task gradients, MAXL will not be guaranteed to converge to a local minimum according to Proposition 1 in [9]. The success of MAXL lies in the fact that the generated auxiliary labels provide a similar but different gradient compared to the primary task.

– Image Resolution (to R3) Since ImageNet is a very large dataset, we re-scaled the images to a smaller resolution due to the limited hardware available in our lab, in order to generate the large number of experimental results presented. Promising future work would be to investigate using a first-order approximation to the second-order gradient, to speed up MAXL for use with limited hardware on higher-resolution images.