We thank the reviewers for their thoughtful insights and comments. Please find below our responses to the main points raised. We have also found a number of typos in addition to the ones listed in the reviews and will correct them in the final version, as well as making improvements based on the comments on readability, clarity, notation, and references.

1. Testing on randomly chosen source and target classes (R1, R2)

As suggested in the reviews, in order to strengthen the validity of our results, we reran the experiments for the 5-shot case on the Tiny ImageNet dataset with source and target classes randomly generated each trial and compared MCW and SVM performance. The result was a much wider range of performance (46.2±15.3 for MCW and 41.7±13.9 for SVM), but the difference in performance between the two methods was consistently in favor of the MCW method ($Accuracy_{MCW} - Accuracy_{SVM} = 4.4 \pm 3.5$). We observe similar phenomena for the other two datasets.

2. Comparison with other algorithms (R2, R3)

We first note that the method from 91-93 was designed for use with a large number of target samples, and used an SVM to combine the networks, similar to what we reported with our SVM experiments. In addition, that method required careful selection of "source tasks" to get outputs that could act as features (such as clustering data points in an unsupervised fashion, with the cluster index as the output), which our method does not require.

Experimentally, we also find that applying the method from 91-93 results in performance that is 3.5±0.8 worse than applying the SVM to the penultimate layers of the networks on the Tiny ImageNet setup.

In the future, we intend to perform some ablative studies and compare to other few-shot learning methods which require source-sample access, such as meta-learning techniques.

3. Revision of Introduction (R2, R3)

We will revise the introduction to provide a short primer on the HGR Maximal Correlation explaining it as a variational formulation of a particular singular value decomposition of joint distributional information with desirable properties for capturing this information, as well as to add intuition for the MCW method in terms of computing weights in a decoupled fashion for each input feature.

4. Focus on the few-shot setting (R3)

Empirically, as shown in Table 1 of the paper, we observed that our method performed best with few samples, with the SVM catching up with our method when we increased the number of samples and eventually overtaking it.

We believe this to be the case because our method operates on each feature independently from the others. As such, it is advantageous in the few-shot setting when there are many features, as our method does not rely on learning ways to combine features. On the other hand, other methods like using an SVM or an additional simple neural net on top of the features requires the learning of a large number of related parameters simultaneously, which is difficult in the few-shot setting as the number of features (and thus parameters) grows large. However, with more samples, the SVM is able to effectively learn how to remove redundant/overlapping features, leading to increased relative performance.

5. Orthogonality constraint in 115-116 (R3)

In the definition of HGR Maximal Correlation given in (1), the functions $f_1, ..., f_k$ are constrained to be uncorrelated, as are $g_1, , , , g_k$, but the expression for each optimal g_i given f_i holds without this constraint due to how the objective separates out (as in (5)). For clarity, we will remove the word "orthogonality" and instead say that the constraint that the variables be uncorrelated is unnecessary. Our revised introduction will also explicitly state this constraint.

6. Few samples resulting in poor estimates for $P^t_{X|Y}$ and P^t_Y from the empirical distributions $\hat{P}^t_{X|Y}$ and $\hat{P}^t_Y$ (R3)

In general, learning weights and features from few samples is very difficult. However, our method learns the weighting for each feature independently, as opposed to other methods which learn all weights over all features simultaneously. We believe this confers an advantage in the number of samples needed to obtain a good estimate for the weights.

We will also revise the paper to note that marginal distributions can be obtained with unlabeled samples, which are often much easier to collect than labeled samples, so obtaining good estimates of them is comparatively easier.

7. Justification/Discussion of method performance (R3)

Based on the feedback from the reviewers, we intend to revise the experimental results/discussion section to include further explanations for why our method performs better or worse in different cases, in addition to the discussion of the choice of the few-shot setting as given in Point 4 of this response.

In particular, for the Stanford Dogs and Tiny ImageNet datasets, we believe that feature redundancy results in less performance gain between the MCW and SVM methods for the Dogs dataset. As a quick experiment, we computed the features for Dogs and Tiny ImageNet again and then, for each feature f_i, we looked at the largest correlation it had with the other features in the experiment. Dogs had a larger average correlation coefficient with other features than ImageNet (0.6 vs. 0.49), which suggests a greater amount of redundancy between features for Dogs. Thus, our method results in less gain since the features are more similar and thus adding additional networks adds less additional information which it is able to leverage to outperform the SVM (which has difficulties in dealing with many uncorrelated features in the few-sample regime) and the single-network methods.