We thank the reviewers for the time and expertise they have invested in these reviews. We appreciate positive comments about the paper like connecting the existing meta-learning frameworks with unsupervised/self-supervised feature learning frameworks, clear presentation of the ideas, good control parameter experiments, and novel approach to the unsupervised meta-learning problem. Furthermore, the insightful comments we received point us towards several directions in which this work can be extended in the future. We focus on answering to the posed questions by the reviewers and provide feedback related to improvements suggestions.

Answers to reviewer #1.

How much does hand-crafted knowledge play a role in the performance of the proposed method? Domain knowledge, such as knowing what transformations of the image retain the classification, does play a role in the performance of the augmentation function. See left side of Table 1 where various hand-crafted augmentations performed at accuracy of 30.16%, 32.80%, 35.09% against the 24.17% baseline. It turned out that the learned auto-augment outperformed all the hand-crafted methods with 39.93%.

In particular, a simple baseline would be to perform few-shot training on models trained with the augmentation methods proposed. That seems like a more informative/reasonable baseline than training from scratch. We agree, we implemented this, and here are the results. For 5-way, 1-shot learning on Omniglot the accuracy was: training from scratch 52.5%, training from scratch with augmentation 55.8%, UMTRA 83.8%. For MiniImagenet the numbers were: from scratch without augmentation 27.6%, from scratch with augmentation 28.8%, UMTRA 39.93%.

The impact of this work could be more substantial if the proposed method were generalized beyond data-augmentation to unsupervised learning in the few-shot setting using other proposed self-supervised techniques. While the main body of the paper focuses on augmentation techniques, its application to the video domain in the supplemental material is using self-supervision: we randomly select 16 frames from one video for training and another 16 frames from the same video for validation. For results, see table at supplemental material page 4.

The approach of using Auto-Augment within UMTRA would suggest significant computational overhead for a new dataset: This overhead indeed exists during the meta-training time. Note however, that this is an offline process which needs to be done once per domain. No augmentation is done during the target learning phase.

How would augmentation work for non-image data? The choice of augmentation is domain specific. For example, in our supplemental material, for videos we do not do augmentation and just pick two different parts of the video. Look at the Figure 5 at supplemental material.

Answers to reviewer #3

that all samples are in a different class (...) how a person might test this assumption on any particular problem where classes are not previously articulated (as they are in imagenet). (...) How would we estimate if "N $\ll$C"?. (...) Does one need to fully articulate the task space – how do you project all possible arrangements of the data into separate classes down into some effective number of classes, c? Indeed, for the unlabeled datasets that we use for meta-learning, we do not know the exact value of C. This is ok, because UMTRA does not take C as a parameter. We only need a rough estimate of it to ensure that "N $\ll$C" holds, we do not need to fully articulate the task space. For instance, if we download 10,000 random videos from YouTube, we can estimate that there will be at least hundreds of different activities in them.

Answers to reviewer #6

Algorithm (N, K)	**(5, 1)**	**(5, 5)**	**(5, 10)**
Training from scratch	26.86	39.65	50.61
UMTRA	**33.43**	**50.19**	**58.84**

I would run experiments on one other domain to assure readers of the generality of the system across domains with widely varying appearances within the dataset. As requested, we run UMTRA on the CelebA dataset (which is unbalanced in terms of the number of examples in each class) for identity recognition and obtained the results shown in table above. Given 5 new identities and one image of each, UMTRA is able to learn the task better than training from scratch. The supplemental material also contains results on another domain, video classification.

Answers to reviewer #7

The most important further contribution would be demonstrating that the augmentation approach can either be effectively automated in all cases, or that a rough hand-designed augmentation approach can be found that works in other domains. We believe that some degree of domain knowledge is a necessary input into the design of the augmentation algorithm. For instance, we know that photos are crops of 2D projections of 3D scenes, which imply certain invariances that can be exploited by the augmentation. This augmentation can be then used across all scenarios involving photos. For instance, the CelebA results above were obtained with the same augmentation used for MiniImagenet.