Response to Reviewer 6

Backbone clarification: We completely agree it is critical to compare models fairly, and hope to convince you we took substantial measures to ensure there were reliable comparisons.

In Table 2, the top blocks are the published results of each particular method. Each paper chooses its own backbone, and the authors optimized the choice of model for their approach. It is fair to compare methods on the quoted results on the *same* benchmarks (which is, after all, the point of benchmarks).

On top of this we provide 3 results on 2 different backbones. The first two results are produced by the the the *same* original MAML/MAML++ backbone (Low-End), trained with either the original MAML formulation or, more reliably, the MAML++ formulation. We will clarify these are the same backbone. In addition we also provide an improved backbone for MAML++ (High-End MAML++) which utilizes a shallow but wide DenseNet with squeeze excite attention. We compare how the two MAML++ backbones perform with and without SCA to demonstrate that performance gains were due to the SCA technique. The final model can then be compared with all previous methods. The proposed combination of MAML++ and SCA and has produced the top results ever published across tasks on Mini-ImageNet and CUB to date.

We would have liked to apply SCA directly on LEO as the technique is general and can be applied to any meta-learning system. However, LEO is closed-source and we have found it very hard to reproduce even after much effort. Hence the comparisons with the MAML backbone.

Figure 2 demonstrates there are changes to the probabilities resulting from the learnt unsupervised loss. Some top classes do change, e.g. in the rightmost chart. The effect of these changes is not in the class-labels alone but also allows better model initialization: training a critic on a pretrained initialization does not provide as much benefit. Hence we did not intend any strong conclusions beyond illustrating that the unsupervised loss has a non-trivial effect.

Discussion on related work. Both papers are highly relevant and we thank the reviewer for alerting us to them. We will reference them in the paper. In Yu et al. (2018), the authors learn a loss function to enable sample-efficient adaptation of an RL agent. Instead we tackle the few-shot classification task by learning a transductive model. Our model adapts a meta-learned initialization by both labelled and unlabelled examples; we do use a learnt unsupervised model, but the combination is critical. Their work is purely unsupervised in the inner loop, using RL as the outer loop, whereas ours consists of both supervised and unsupervised inner loops and a supervised outer loop.

In Evolved Policy Gradients, the author also learns a loss function to train a policy network towards an RL task. The loss function receives both states and rewards, it has an RL inner phase and an evolutionary algorithm outer loop which both also has access to rewards. We target a different problem, that is few-shot learning, using a supervised outer loop, and a transductive inner loop (composed by supervised and unsupervised inner loop phases).

Broader formulation of transductive meta-learning: We understand there is no particular dependence on MAML. However feedback from colleagues on earlier drafts suggested the paper clarity was improved when using a concrete case rather than the previous more general description. However we should indeed alert the reader to this generality in the conclusion.

Response to Reviewer 3

Utilization of the target-set information: We use gradients of a whole batch of target set images to update the base model since this produces better performance. The results in the paper represent experiments where we used target sets of size 75 to learn the loss function, however one could instead randomly choose the batch size of the target set to train a loss function that generalizes on a wide range of batch sizes.

Additional diagrams: Space constraints prevented us including further diagrams. We will prepare a longer version which will incorporate your suggestions. Code will also be available upon acceptance of this paper.

Response to Reviewer 7

Indeed, we are currently using the Meta-Dataset (https://arxiv.org/abs/1903.03096) in other work; however, in the context of this paper, effective runs on this were beyond our computational capacity in the time frame.