Broader impacts: CO problems in industry and deep RL methods. With POMO, our long-term primary aim is to solve hard, practical CO problems that arise in industry, manufacturing and logistics in particular. We demonstrated that POMO can solve three benchmark CO problems of different natures using the same neural net and the same training method. Human guidance during training was minimal, as we only needed to provide problem-specific scoring functions to the machine and the rest was automatic. Such adaptability and autonomy are extremely important features for industrial applications, where optimization process must be performed under various constrains and should adapt to changing environments rapidly.

From finding more efficient routing plans for goods to optimal assignments among groups of tasks and machines, majority of the problems encountered in operation research (OR) are CO problems. These problems are still dealt with hand-crafted heuristics as a common practice. In the field of computer vision and natural language processing, classical methods based on manual feature engineering by experts have now been superseded by automated deep learning algorithms. Our work signifies that such change is also possible for OR using deep-RL methods.

Traditional heuristic techniques used in OR, however, have remarkable performance. This creates a high barrier for deep RL approaches to gain meaningful attention from industry, yet. Machine learning-based heuristic solvers for CO problems that have a comparable performance to traditional rule-based solvers are extremely rare. We strongly believe that our POMO approach needs a special recognition to properly fuel the AI research efforts towards the OR practice.

Contribution. Contribution of our paper is three fold. First, we identify symmetries in RL methods for solving CO problems that can be leveraged to create a powerful heuristic solver. Second, we propose POMO method that rigorously takes advantage of the found symmetry: (1) Multiple trajectories are generated, each having a different but equivalent optimal solution as its goal for exploration. (2) A newly devised, low-variance baseline for policy gradient is used to update the policy. It is based on the total rewards from heterogeneous trajectories, and thus it is less vulnerable to local minima. (3) Instance augmentation technique is used to further exploit the symmetry at the inference stage. Third, we empirically demonstrate the effectiveness of the POMO method on three classic NP-hard benchmark problems with new state-of-the-art results.

Reviewer 1. [Weakness, Line 252, Contribution] We appreciate your valuable comments. Our answers regarding your comments are given above, and we will include them in our paper. [Correctness] In line 6, we will change the text to "... with an extended exploration techniques towards all optimal solutions." | We will add training times to our paper. [Clarity] We will change "With enough training" to "When training converges" in line 104. | With different starting points, no two trajectories can be identical. By the shared baseline, such heterogeneous trajectories are compared to each other to find the common optimal reward. Without the shared baseline, trajectories are assessed independently and the best performing trajectory is no longer guaranteed to be reinforced most. We will add more explanations to line 121.

Reviewer 2. [Summary, Weakness] POMO is more than a sampling technique. Behind the significant improvement on the experimental result is the new policy gradient calculation with lowered variance. Also, achievement of better trained neural net compared to the previous work based on the same model proves that our training method is highly resistant to falling into a local minimum as explained in Section 4.2. [Feedback] We will fix the typos. We will include values of N in Section 5.1 (N=50 for TSP50, N=100 for TSP100, etc.).

Reviewer 3. [Weakness] POMO assigns a single rollout for each optimal solution, and multiple rollouts are natural byproduct from the existence of multiple optima that indeed maximizes entropy on the first action. Relation between POMO and a max-entropy RL method is interesting, and we will add a discussion related to this topic. | The core architecture from the past work has not been changed. As for hyperparameters, we will rerun experiments and update our results to match them to avoid confusion from readers. This has marginal effect in our experience. [Correctness] We will change the word "Traditionally" to "In previous deep-RL research on CO problems." Abstract will have the change "... with an extended exploration techniques towards all optimal solutions." [Others] We will fix lines 32 and 29 as suggested, and we will include references to augmentation techniques used in supervised learning.

Reviewer 4. [Weakness, Feedback] Ablation studies for POMO using baselines other than attention model (AM) is an interesting topic that probably needs another paper. All the other baselines are based on "improvement type" methods (see paragraph "Construction vs. improvement") that offer no obvious way to define multiple starting points, each leading to different optimum. A separate ablation study only on instance-augmentation component is possible, however, using AM without multiple starting points, the results of which we will add to our paper. Some of the result is displayed in the table on the right; it is interesting that $\times 8$ augmentation (i.e., choosing the best out of 8 greedy trajectories) improves the AM result to the same level achieved by 1280 sampling trajectories. | We will also add graphs showing POMO's solutions for random problem instances to help visualize its empirical performance.

Method	TSP 50		TSP 100	
	Len.	Gap	Len.	Gap
AM, greedy	5.80	1.76%	8.12	4.53%
AM, sampling	5.73	0.52%	7.94	2.26%
AM, $\times 8$ aug.	5.73	0.53%	7.95	2.37%