DisCoRL: Continual Reinforcement Learning via Policy Distillation

René Traoré* Hugo Caselles-Dupré* **Timothée Lesort*** Te Sun Guanghang Cai Natalia Díaz-Rodríguez David Filliat

Autonomous Systems and Robotics Laboratory (ENSTA Paris), Theresis Laboratory (Thales), AI Lab (Softbank Robotics Europe), *Equal contribution

Abstract

In **multi-task RL** there are two main challenges: at training time, the ability to learn different policies with a single model; at test time, inferring which of those policies should be applied without an external signal. In the case of **continual RL** a third challenge arises: learning tasks sequentially without forgetting the previous ones. In this paper, we propose DisCoRL, an approach combining state representation learning and policy distillation. We experiment on a sequence of three simulated 2D navigation tasks with a 3 wheel omni-directional robot. Moreover, we tested our approach's robustness by transferring the final policy into a real life setting. The policy can solve all tasks and automatically infer which one to run.

Contribution

1) Applying *State Representation Learning* (SRL) [1] into a continual learning setting of reinforcement learning. The SRL method learns a compact and efficient representation of data that facilitates learning a policy [2].
2) Proposing a CL algorithm based on distillation that does not manually need to be given a task indicator at test time, but learns to infer the task from observations only.
3) Successfully applying the learned policy on a real robot.

Tasks

We present 3 different 2D navigation tasks to a 3 wheel omni-directional robot. We want it to learn to solve them sequentially. The robot has first access to task 1 only, and then to task 2 only, and so on. It should learn a single policy that solves all tasks and be applicable in a real life scenario. The robot can perform 4 high level discrete actions (move left/right, move up/down).
Task 1: Target Reaching (TR): Reaching a red target randomly positioned.
Task 2: Target Circling (TC): Circling around a fixed blue target.
Task 3: Target Escape (TE): Escaping a moving robot.

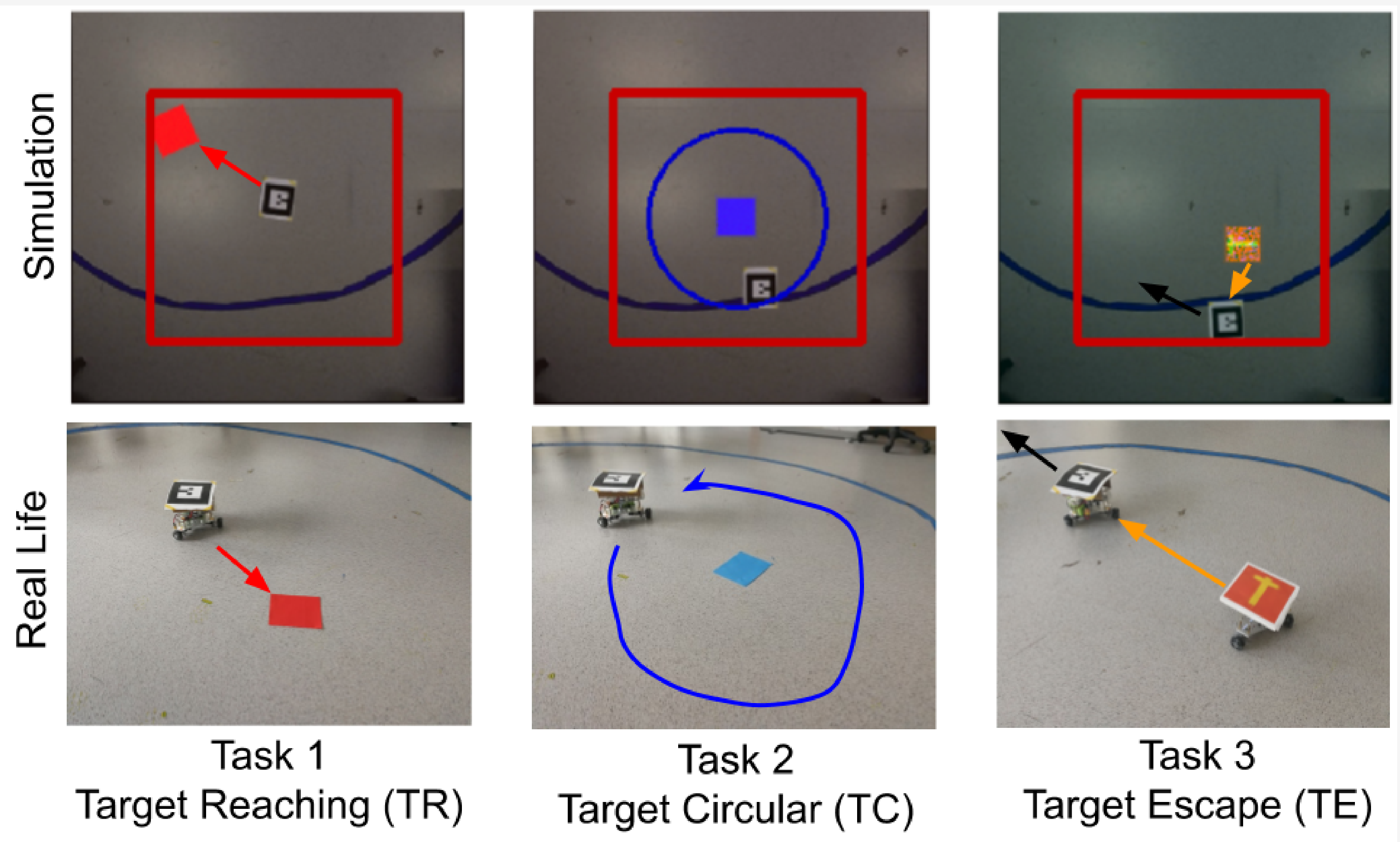

Figure 1: The three tasks, in simulation (top) and in real life (bottom), sequentially experienced. Learning is performed in simulation, the real life setting is only used at test time.

Approach

Method: The SRL model is trained with an auto-encoder and an inverse model. Once the SRL model is trained on task i, we keep its encoder E_i and use it to learn our policy π_t on top with PPO2 [3]. We sample then a dataset D_{π_t} on policy with π_t. When all tasks have been learned, we merge all D_{π_t} and distill the knowledge into a new model $\pi_{d:1,..,n}$. This method makes simple multi-task continual learning and does not need task label at test time.
Distillation: The distillation process [4] consists in transferring knowledge from one or several neural network(s) (the teacher(s)) to another (the student). The teacher annotates a database with soft labels (actions probabilities). In our setting, the student is trained to fit these soft labels in order to learn a policy. This process is used here to distill our two policies into one model.

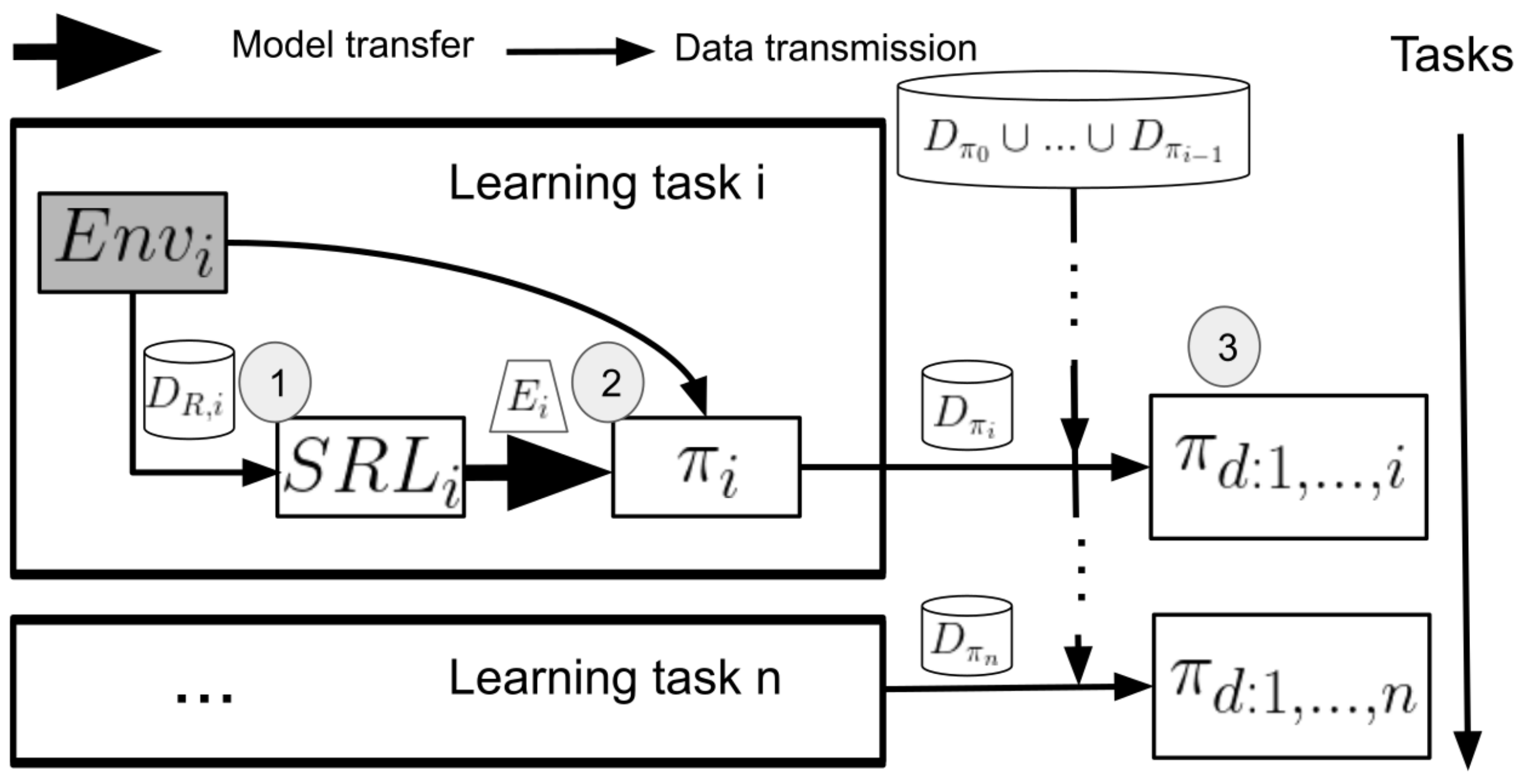

Figure 2: Overview of our full pipeline for Continual Reinforcement Learning. White cylinders are for datasets, gray squares for environments, and white squares for learning algorithms, whose name correspond to the model trained. Each task i is learned sequentially and independently by first generating a dataset $D_{R,i}$ with a random policy to learn a state representation with an encoder E_i with an SRL method (1), then we use E_i and the environment to learn a policy π_i in the state space (2). Once trained, π_i is used to create a distillation dataset D_{π_i} that acts as a memory of the learned behaviour. All policies are finally compressed into a single policy $\pi_{d:1,..,i}$ by merging the current dataset D_{π_i} with datasets from previous tasks $D_{\pi_1} \cup ... \cup D_{\pi_{i-1}}$ and using distillation (3).

Results

- Testing different sampling methods

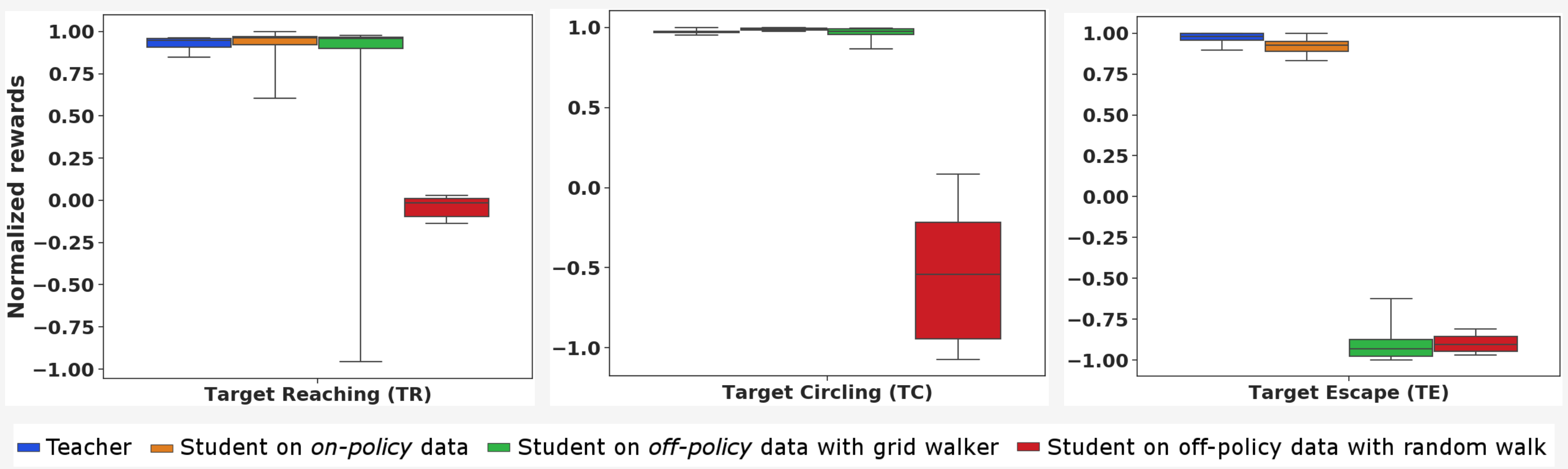

Figure 3: Distilled policies' efficiency on 8 seeds using various data generation strategies for each task separately. Each policy is distilled on 15k tuples of sampled observations and action probabilities, for 4 epochs.

- Final performance

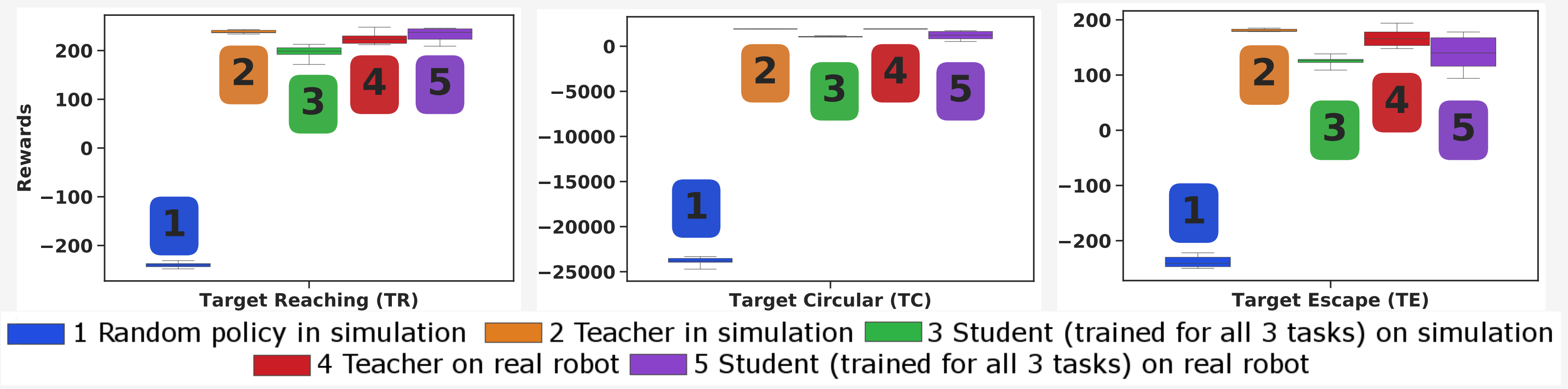

Figure 4: Main result: distillation in a continual learning setting of three teacher policies into a single student policy. The resulting policy is able to perform all three tasks both in simulation and in the real world, while minimizing forgetting.

Conclusion

In this paper we presented **DisCoRL**, an approach for continual reinforcement learning. The method consists of sequentially summarizing the learned policies into a dataset to distill them into a student model. It allows to learn sequential tasks in a stable pipeline without forgetting. Some loss in performance may occur while transferring knowledge from teacher to student, or while transferring a policy from simulation to real life. Nevertheless, our experiments show promising results in simulated environments and real life settings.

- **Repository**: `https://github.com/kalifou/robotics-rl-srl`
- **Acknowledgement**: This work is supported by the EU H2020 DREAM project (Grant agreement No 640891).

References

[1] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat, "State representation learning for control: An overview," *Neural Networks*, 2018.

[2] A. Raffin, A. Hill, K. R. Traoré, T. Lesort, N. Díaz-Rodríguez, and D. Filliat, "Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics," *Workshop on "Structure and Priors in Reinforcement Learning" (SPiRL) at ICLR*, 2019.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.