Shotcrete Guidance in Harsh Environments with Uncertainty-Aware Reinforcement Learning

Yurui Du[1], Louis Hanut[2], Herman Bruyninckx[2], Renaud Detry[1, 2]

***Abstract*—This study addresses the challenges of robotic manipulation of amorphous materials in construction, particularly shotcrete application, where high-velocity concrete spraying leads to dust, low visibility, and occlusions. These factors contribute to significant sensory noise and partial observability in 3D visual inputs. We introduce a novel framework combining uncertainty-aware 3D geometric visual input processing through variational inference with deep reinforcement learning (DRL) for ensemble policy learning. This approach enhances robustness in noisy, partially observable environments, showing marked improvements in completion time and material conservation over traditional methods. Our results underscore the efficacy of observational-uncertainty-aware DRL in addressing complex real-world scenarios.**

I. Introduction

Shotcrete, a process widely used in construction and mining industries for stabilizing tunnels and slopes, involves projecting a stream of concrete onto a surface at high velocity to manipulate the deposition of amorphous concrete material [1]. The harmful dust and physical strain from shotcrete has inspired growing interests in automating this task with robots. But the amorphous nature of sprayed concrete makes it more challenging to manipulate than rigid objects due to its complex dynamics [2]. Real-world applications such as shotcrete are even harder because sensory capability is impaired to acquire visual observations in harsh working environments filled with heavy concrete dust [3]. While the agent can learn to shotcrete based on interactions with the environment with the advances in end-to-end imitation learning (IL) and reinforcement learning (RL), industrial shotcrete tasks do not often allow mass data collection because physical interactions between the robot and environment can be too costly and unsafe. As a workaround, end-to-end RL has been widely used in a sim-to-real setting, where a policy is first trained in simulation and later transferred to the real world.

To achieve a successful sim-to-real transfer of the learned policy, a suitable representation of visual input is essential. While it is common to use raw sensory data without preprocessing in end-to-end policy learning [2], [4], such an approach proves inadequate to address several sim-to-real gaps regarding shotcrete applications. Primarily, the presence of heavy concrete dust and occlusions from the plume can significantly impair sensory functions, rendering the raw data excessively noisy for meaningful extraction of 3D geometrical features of concrete deposition. Additionally, real-world industrial shotcrete processes demand online monitoring for field operators to track task progress. This necessitates the denoising and enhancement of raw data to yield semantically understandable visual representations. Furthermore, the limited availability of real shotcrete data poses constraints on learning complex representations directly from raw sensory inputs. In this context, our work opts for heightmaps of concrete deposition as the optimal visual representation to bridge the sim-to-real gap effectively as it is simple enough for simulation-based training yet sufficiently informative to encapsulate critical 3D geometrical features, sensory noise, and partial occlusions encountered in real-world scenarios.

Another major drawback of the sim-to-real transfer of existing works in this domain is that they often overlook observational uncertainty in their decision-making processes, presuming the states to be almost fully observable [5]–[8]. Such assumptions are not viable for shotcrete applications, where observational noise and partial occlusion not only exist but can be predominant, overshadowing actual state changes. Addressing this, our approach distinctively incorporates an explicit estimation of observational uncertainty. This strategy aims to derive a policy that demonstrates robustness against the real-world challenges of sensory noise and partial occlusions, ensuring reliable application in shotcrete tasks.

In summary, our paper presents two main contributions: (a) we propose a novel RL algorithm that utilizes observational uncertainty estimation to address the shotcrete problem; (b) we show the uncertainty estimation method improves the performance regarding completion time and material conservation.

II. Related Work

A. Visual representation of amorphous material

In the realm of end-to-end policy learning for the manipulation of amorphous materials, visual representations predominantly manifest as latent forms derived from raw image data [2], [4]–[8], yet these often overlook the complexities of occlusions. This oversight may precipitate generalization challenges when transitioning from the training domain to the target domain. Efforts to address these challenges have been made, such as tackling self-occlusions within the material via mesh reconstruction [10], or adopting particle-based approaches for dynamic modeling [11]. Despite these advances, a critical gap persists: existing visual representations largely fail to

*Funded by the European Union (robetarme-project.eu). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or HADEA. Neither the European Union nor the granting authority can be held responsible for them.

[1]Department of Electrical Enginnering, KU Leuven, Leuven, Belgium, {yurui.du, renaud.detry}@kuleuven.be

[2]Department of Mechanical Engineering, KU Leuven, Leuven, Belgium, {louis.hanut, herman.bruyninckx}@kuleuven.be

account for sensory noise and external occlusions independent of the material. Our approach, inspired by [12], utilizes a 2D heightmap representation of concrete, chosen for its inherent resilience to variations in concrete appearance. We go beyond existing methodologies by creating a shotcrete simulator, based on the model proposed in [13]. In this simulator, the concrete deposition state is represented by the heightmap, and its corresponding observations are rendered by adding sensory noise and external occlusions derived from empirical data collected during real-world shotcrete operations. We adopt the same state representation in actual shotcrete applications by converting stereo images into such heightmaps, facilitating a direct, zero-shot application of policies trained in the simulator to real-world environments.

B. Bridge sim-to-real gap

The concept of sim-to-real transfer of learned manipulation policy is not novel in the domain of DRL. There have been extensive works over an extended period [14]–[25]. Central to these investigations is the interpretation of their foundational principles from the standpoint of contextual Markov Decision Processes (MDPs). This perspective allows for the assimilation of simulated and real-world scenarios as singular RL problems, distinguished only by varying contextual parameters. Pioneering efforts in this domain have leveraged recurrent neural agents, employing domain randomization techniques that modify transition dynamics, observations, or rewards. Such approaches enable RL agents to construct and utilize an internal latent memory, summarizing historical data to facilitate the learning of adaptive, context-sensitive policies.

Transitioning these principles to construction robotics, especially shotcrete, reveals inherent limitations. While RL agents are adept at inferring context from past observations and reacting appropriately, the 'naive' application of domain randomization can falter in the face of excessively noisy observations and the profound uncertain dynamics characteristic of partially observable environments. In such scenarios, the agents may struggle to differentiate between contexts during testing, leading to subpar generalization capabilities. Our work addresses this issue by approximating the observational uncertainty using variational inference to provide a reconstructed true state given the observations. Futhermore, we use the confidence of reconstructed states in bootstrap ensemble DRL [26] to train an approximate Bayesian optimal manipulation policy that shows sim-to-real robustness.

III. Background

Contextual Reinforcement Learning (CRL) extends the standard RL framework to account for the effects of context in the learning process. Here, sim-to-real transfer can be interpreted as simulated and real-world scenarios with different contexts, thus having different dynamics, observation probabilities and reward functions. The CRL framework is represented by the tuple $(\mathcal{C}, \mathcal{S}, \mathcal{O}, \mathcal{A}, P, O, R, \gamma)$, where an agent interacts with an environment across various contexts, aiming to maximize cumulative rewards. Here, $\mathcal{C}$ is a finite set of contexts, $\mathcal{S}$ denotes the state space, $\mathcal{O}$ represents the observation space, $\mathcal{A}$ is the action space, $P : \mathcal{C} \times \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the state transition probability function, $O : \mathcal{C} \times \mathcal{S} \to \mathcal{P}(\mathcal{O})$ is the observation function, $R : \mathcal{C} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor.

In this context-aware and partially observable setting, the policy $\pi : \mathcal{C} \times \mathcal{O} \to \mathcal{P}(\mathcal{A})$ maps contexts and observations to a probability distribution over actions. The objective is to find an optimal policy π^* that maximizes the expected cumulative reward, defined as $\pi^* = \arg\max_\pi \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T} \gamma^t R(c_t, s_t, a_t)\right]$, where $\tau = (c_0, s_0, o_0, a_0, \ldots, c_T, s_T, o_T, a_T)$ is a trajectory generated under policy π, with each element $o_t \in \mathcal{O}$ representing an observation.

The inclusion of the observation space $\mathcal{O}$ is crucial for dealing with environments where agents have access to only partial observations of the state, a common scenario in real-world applications. This formulation allows the agent to make decisions based on limited information, learning to estimate belief states and adapting its strategy accordingly.

In the next section, we discuss how to effectively estimate actual states from noisy observations and how to achieve sim-to-real transfer for the policy learned in simulation.

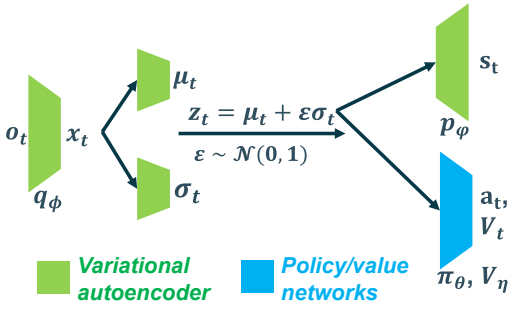

Fig. 1. Uncertainty-Aware DRL using VAE

IV. Methods

In this study, we propose an innovative ensemble deep reinforcement learning framework, integrating a custom Variational Autoencoder (VAE) [27] with a Proximal Policy Optimization (PPO) [28] based actor-critic model. Our method encompasses two core components: a tailored VAE for uncertainty estimation via true state reconstruction from noisy partial observations, and an ensemble approach that leverages multiple learning agents to enhance decision-making robustness.

A. VAE for state reconstruction and uncertainty estimation from noisy, partial observation

The VAE architecture plays a pivotal role in our framework, efficiently serving 3 purposes: feature extraction, state reconstruction, and uncertainty estimation. An intuitive description of the network architecture is shown in Fig. 1. The encoder $q_\phi(z_t|o_t)$ is designed to process noisy, partially occluded

2D heightmap o_t. It utilizes a series of convolutional layers for image processing, followed by fully connected layers to generate a latent representation z_t, which is shared with the policy and value networks of PPO.

The VAE decoder $p_\varphi(\hat{s}_t|z_t)$, structurally mirroring the encoder, focuses on reconstructing the observation's corresponding true state estimate $\hat{s}_t$ from z_t. It comprises a sequence of transposed convolutional layers that incrementally upscale z_t, culminating in the reconstruction of the shape of o_t. The observational uncertainty is approximated as the variance of $\hat{s}_t$, computed by multiple samples of z_t. The VAE is trained by minimizing the reconstruction loss and KL-divergence between the approximate posterior $q_\phi(z_t|o_t)$ and prior distribution $p(z_t)$.

$$\begin{aligned}\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{z\sim q_\phi(z_t|o_t)}[\log p_\theta(\hat{s}_t|z_t)] \\ +\beta \cdot \text{KL}[q_\phi(z_t|o_t)\,||\,p(z_t)]\end{aligned} \tag{1}$$

B. Uncertainty-Aware Ensemble DRL Mechanism

We propose a novel architecture that orchestrates the training and interaction of multiple agents within an ensemble. Each agent in the ensemble is instantiated with the custom VAE and an actor-critic model based on PPO. The ensemble framework manages individual agent-environment interactions. Agent-specific buffers store experience tuples $(s_t, o_t, a_t, r_t)_D$, facilitating independent learning while allowing for cross-agent knowledge sharing. This design not only encourages diversification in learned policies but also fosters a cooperative learning environment, enhancing overall decision-making robustness.

A distinctive feature of our framework is the ensemble learning mechanism, which incorporates a novel approach to policy calibration. Agents in the ensemble compute the variance in their reconstructed states as a measure of certainty. This variance informs the weighting of each agent's policy contribution to a collective decision-making process. Higher certainty (lower variance) leads to a greater influence of an agent's policy on the ensemble's combined policy.

The combined policy is formulated as a weighted sum of individual policies, with weights inversely proportional to the variance in reconstructed states. Subsequently, we compute the KL-divergence between each agent's policy and the combined policy. This divergence serves as an additional loss term during training, ensuring that individual policies do not diverge significantly from the ensemble consensus, thus maintaining a coherent and collaborative decision-making strategy across the ensemble. Follow the theoretical framework of contextual Markov decision processes (CMDPs), where simulation and real-world environment are MDPs conditioned by different contexts $\{c_i, c \in \mathcal{C}\}$. Our uncertainty-aware DRL framework is shown in Algorithm 1.

Algorithm 1 Uncertainty-Aware Ensemble PPO

Sample contexts $\{c_i, i \in N\}$ with replacement from $\mathcal{C}$, initialize policy θ_i, value η_i, and VAE network ϕ_i for each c_i
for each iteration **do**
  Collect data $\{s, o, a, s', o'\}_D$ using π_i in c_i
  for each mini-batch B in D **do**
    Compute $\mathcal{L}_{\text{VAE}}, \mathcal{L}_{\text{PPO}}, \mathcal{L}_V$
    Compute confidence score $\sigma^2_{t,i}$ for each $o_{t,i}$ in c_i
    Compute imitated policy $\pi = \sum_{i=1}^{N} \frac{\pi_i}{\sigma^2_{t,i}}$
    Compute uncertainty-aware loss $\mathcal{L}_{\text{UA}} = D_{\text{KL}}[\pi||\pi_i]$
    Update θ_i, η_i, ϕ_i using $\mathcal{L} = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{PPO}} + \mathcal{L}_V + \mathcal{L}_{\text{UA}}$
  end for
end for

V. Experiments

This section delineates the experimental evaluation of our proposed Uncertainty-Aware Proximal Policy Optimization (UAPPO) against two baselines: the vanilla PPO and Model Predictive Control (MPC).

A. Simulated shotcrete experiment

1) **Setup** In this study, UAPPO, PPO, and MPC algorithms were tested in an OpenAI Gymnasium simulation, where the objective was to apply shotcrete on a 2m x 1m flat surface to achieve a 5cm target thickness. Any concrete applied outside this area or beyond the target thickness was considered waste.
   Both UAPPO and PPO were trained using 2.4 million transition samples $\{s, o, a, s', o'\}$. The training involved constant environmental transition dynamics, whereas the evaluation phase presented the agents with variable dynamics, ranging from 10% to 1000% of those in training. This setup aimed to test the algorithms' adaptability to diverse and unforeseen scenarios.
   The evaluation phase also featured a heightened challenge by making the agents' observations highly partially observable. 80% of each observation's information was masked, simulating the limited information availability in real-world conditions and testing the robustness of the policies under such constraints.
2) **Metrics** Three metrics are used to evaluate performance: completion time, computation time and wasted volume. To mitigate the effects of stochasticity in the evaluation process, the performance metrics of each agent are computed from four repetitive validation trials.
3) **Simulation Results** The results presented in Table I indicate that UAPPO outperforms other methods in terms of completion time and material conservation, while MPC exhibits the least favorable performance across all evaluated metrics. This observation aligns well with our initial expectations. In contrast to PPO, the superior performance of UAPPO can be attributed to its use of VAE for estimating observational uncertainty, which not only aids in guiding exploration during the training phase but also enhances robustness during testing. Conversely, MPC is encumbered by significant computational demands due to its reliance on online optimization. To conform to the constraints of

TABLE I
PERFORMANCE COMPARISON IN SIMULATION

Models	UAPPO	PPO	MPC
Completion Time (s)	1985 ± 5	2035 ± 28	2057 ± 5
Computation Time (ms)	45 ± 12	48 ± 15	98 ± 15
Wasted Volume (m^3)	0.0867 ± 0.004	0.0911 ± 0.015	0.0927 ± 0.003

real-time planning, MPC necessitates a reduction in the planning horizon, subsequently diminishing its planning efficacy.

VI. CONCLUSIONS

This paper introduced an innovative ensemble deep reinforcement learning framework, merging a custom VAE with PPO. This dual approach effectively harnesses the VAE's capacity to distill high-dimensional sensory data into meaningful latent representations, thereby enhancing state representation in complex robotic tasks.

Our method elevates the decision-making capabilities of individual agents in scenarios characterized by noisy and partial observations, and it cultivates a collaborative dynamic through an ensemble learning mechanism. This system minimizes individual policy deviations and promotes collective behavior, aligning with a wisdom that aggregates all policies in the ensemble, weighted by their respective confidences.

The empirical results underscore our approach's superiority in learning efficiency and policy robustness compared to traditional methods. The agents adeptly adapt to new situations and make decisions based on highly uncertain observations by utilizing the variance of reconstructed states as a certainty measure. Despite its computational demands and current limitations in sample efficiency and scalability, the framework offers promising avenues for future research, including exploring its application to larger robotic swarms and the transferability of learned policies across diverse tasks.

REFERENCES

[1] Galan, I., Baldermann, A., Kusterle, W., Dietzel, M., and Mittermayr, F. (2019). Durability of shotcrete for underground support–Review and update. Construction and Building Materials, 202, 465–493.

[2] Wu, Y., Yan, W., Kurutach, T., Pinto, L. and Abbeel, P., 2019. Learning to manipulate deformable objects without demonstrations. arXiv preprint arXiv:1910.13439.

[3] Chen, L., Li, P., Liu, G., Cheng, W., and Liu, Z. (2018). Development of cement dust suppression technology during shotcrete in mine of China-A review. Journal of Loss Prevention in the Process Industries, 55, 232–242.

[4] Wang, A., Kurutach, T., Liu, K., Abbeel, P. and Tamar, A., 2019. Learning robotic manipulation through visual planning and acting. arXiv preprint arXiv:1905.04411.

[5] D. Seita et al., "Deep Imitation Learning of Sequential Fabric Smoothing From an Algorithmic Supervisor," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2020, pp. 9651-9658.

[6] Weng T, Bajracharya SM, Wang Y, Agrawal K, Held D. Fabricflownet: Bimanual cloth manipulation with a flow-based policy. In Conference on Robot Learning 2022 Jan 11 (pp. 192-202). PMLR.

[7] Yan, W., Vangipuram, A., Abbeel, P. and Pinto, L., 2021, October. Learning predictive representations for deformable objects using contrastive estimation. In Conference on Robot Learning (pp. 564-574). PMLR.

[8] Matas, J., James, S. and Davison, A.J., 2018, October. Sim-to-real reinforcement learning for deformable object manipulation. In Conference on Robot Learning (pp. 734-743). PMLR.

[9] Nguyen, H., Daley, B., Song, X., Amato, C. and Platt, R., 2020. Belief-grounded networks for accelerated robot learning under partial observability. arXiv preprint arXiv:2010.09170.

[10] Huang, Z., Lin, X. and Held, D., 2022. Mesh-based dynamics with occlusion reasoning for cloth manipulation. arXiv preprint arXiv:2206.02881.

[11] Lin, X., Wang, Y., Huang, Z. and Held, D., 2022, January. Learning visible connectivity dynamics for cloth smoothing. In Conference on Robot Learning (pp. 256-266). PMLR.

[12] Zhang, Y., Yu, W., Liu, C.K., Kemp, C. and Turk, G., 2020. Learning to manipulate amorphous materials. ACM Transactions on Graphics (TOG), 39(6), pp.1-11.

[13] Lu, B., Li, M., Wong, T. N., and Qian, S. (2021). Effect of printing parameters on material distribution in spray-based 3D concrete printing (S-3DCP). Automation in Construction, 124, 103570.

[14] Peng, X. bin, Andrychowicz, M., Zaremba, W., and Abbeel, P. (2018). Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. 2018 IEEE International Conference on Robotics and Automation (ICRA), 3803–3810.

[15] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 23–30.

[16] Tan, J., Zhang, T., Coumans, E., Iscen, A., Bai, Y., Hafner, D., Bohez, S., and Vanhoucke, V. (2018, June 26). Sim-to-Real: Learning Agile Locomotion For Quadruped Robots. Robotics: Science and Systems XIV.

[17] Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W., and Abbeel, P. (2018, June 26). Asymmetric Actor Critic for Image-Based Robot Learning. Robotics: Science and Systems XIV.

[18] Andrychowicz, O. M., Baker, B., Chociej, M., Józefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schneider, J., Sidor, S., Tobin, J., Welinder, P., Weng, L., and Zaremba, W. (2020). Learning dexterous in-hand manipulation. The International Journal of Robotics Research, 39(1), 3–20.

[19] Sadeghi, F., and Levine, S. (2017, July 12). CAD2RL: Real Single-Image Flight Without a Single Real Image. Robotics: Science and Systems XIII.

[20] James, Stephen, Andrew J. Davison, and Edward Johns. "Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task." Conference on Robot Learning. PMLR, 2017.

[21] Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., and Hutter, M. (2019). Learning agile and dynamic motor skills for legged robots. Science Robotics, 4(26).

[22] Chen T, Murali A, Gupta A. Hardware conditioned policies for multi-robot transfer learning. Advances in Neural Information Processing Systems. 2018;31.

[23] Kumar, A., Fu, Z., Pathak, D., and Malik, J. (2021, July 12). RMA: Rapid Motor Adaptation for Legged Robots. Robotics: Science and Systems XVII.

[24] Muratore, F., Gienger, M., and Peters, J. (2021). Assessing Transferability From Simulation to Reality for Reinforcement Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(4), 1172–1183.

[25] Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., and Hutter, M. (2020). Learning quadrupedal locomotion over challenging terrain. Science Robotics, 5(47).

[26] Ghosh, D., Rahme, J., Kumar, A., Zhang, A., Adams, R. P., and Levine, S. (2021). Why generalization in rl is difficult: Epistemic pomdps and implicit partial observability. Advances in Neural Information Processing Systems, 34, 25502-25515.

[27] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[28] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).