Abstract—Reinforcement learning (RL) methods for social robot navigation show great success navigating robots through large crowds of people, but the performance of these learning-based methods tends to degrade in particularly challenging or unfamiliar situations due to the models’ dependency on representative training data. To ensure human safety and comfort, it is critical that these algorithms handle uncommon cases appropriately, but the low frequency and wide diversity of such situations present a significant challenge for these data-driven methods. To overcome this challenge, we propose modifications to the learning process that encourage these RL policies to maintain additional caution in unfamiliar situations. Specifically, we improve the Socially Attentive Reinforcement Learning (SARL) policy by (1) modifying the training process to systematically introduce deviations into a pedestrian model, (2) updating the value network to estimate and utilize pedestrian-unpredictability features, and (3) implementing a reward function to learn an effective response to pedestrian unpredictability. Compared to the original SARL policy, our modified policy maintains similar navigation times and path lengths, while reducing the number of collisions by 82% and reducing the proportion of time spent in the pedestrians’ personal space by up to 19 percentage points for the most difficult cases. We also describe how to apply these modifications to other RL policies and demonstrate that some key high-level behaviors of our approach transfer to a physical robot.

I. INTRODUCTION

While robot navigation has been explored extensively, smooth integration of mobile robots into human-populated spaces is yet to be achieved. Robots that interact with people are expected to navigate in a way that is predictable and unobtrusive, maintaining both the safety and comfort of surrounding people [1]. The social robot navigation field is seeing a growing number of RL-based approaches that implicitly predict human motion and plan robot paths without explicit models of human behavior [2]. These RL-based approaches have achieved great success in enabling effective navigation around large crowds of people, outperforming traditional approaches [3]–[6]. However, the performance of RL policies is contingent on having representative training data, so these policies are sensitive to differences in pedestrian behavior seen during deployment versus training (a problem generally referred to as domain shift). Domain shift is always a concern with learning-based methods but is of particular importance in social robot navigation because of the wide range of human behavior and the potential physical hazards and psychological risks associated with mobile robots operating in close proximity to humans [7].

Thus, to widely deploy RL-based approaches for social robot navigation, these methods should recognize their own level of uncertainty in situations in which pedestrians behave unpredictably from the perspective of the RL policy. We say that pedestrians behave unpredictably if their behavior deviates significantly from their normal or expected behavior as defined by the RL policy’s implicit model of human behavior. Once an RL policy recognizes that it is in an unfamiliar situation and cannot accurately predict the behavior of nearby pedestrians, it should respond with appropriate caution (Figure 1). Thus, it must distinguish between predictable (green) and unpredictable (pink) pedestrians and maintain appropriate caution while still navigating efficiently.

![Fig. 1: (a) RL-based robot navigation policies are trained with humans that behave according to some pedestrian model. (b) During deployment, these policies will encounter pedestrians that behave differently. Existing RL policies generally do not consider this and continue to treat all pedestrians the same, presenting concerns for human comfort and safety. (c) RL policies should distinguish between predictable (green) and unpredictable (pink) pedestrians and maintain appropriate caution while still navigating efficiently.](image-url)

Thus, to widely deploy RL-based approaches for social robot navigation, these methods should recognize their own level of uncertainty in situations in which pedestrians behave unpredictably from the perspective of the RL policy. We say that pedestrians behave unpredictably if their behavior deviates significantly from their normal or expected behavior as defined by the RL policy’s implicit model of human behavior. Once an RL policy recognizes that it is in an unfamiliar situation and cannot accurately predict the behavior of nearby pedestrians, it should respond with appropriate caution (Figure 1). Thus, it must distinguish between predictable (green) and unpredictable (pink) pedestrians and maintain appropriate caution while still navigating efficiently.

![Fig. 1: (a) RL-based robot navigation policies are trained with humans that behave according to some pedestrian model. (b) During deployment, these policies will encounter pedestrians that behave differently. Existing RL policies generally do not consider this and continue to treat all pedestrians the same, presenting concerns for human comfort and safety. (c) RL policies should distinguish between predictable (green) and unpredictable (pink) pedestrians and maintain appropriate caution while still navigating efficiently.](image-url)

To explore this idea, we incorporate uncertainty-awareness into an existing RL policy called SARL [6] by (1) modifying the training process to systematically inject significant deviations into a model of pedestrian behavior, (2) augmenting the observation space of the value network algorithm to recognize and quantify deviations of pedestrian behavior from the assumed model, and (3) adding a term to the reward function to encourage caution toward progressively more unpredictable pedestrians while navigating normally around predictable ones. We conduct ablation studies in simulation to understand the cumulative impact of these modifications and find that they substantially improve the performance of SARL around particularly difficult and previously unseen pedestrian behavior. Compared to the original policy, our modified policy maintains similar navigation times and path lengths while notably reducing the number of collisions and the proportion of time spent in the simulated pedestrians’ personal space. We then describe how the same modifications...
can be made to other socially-aware RL policies and demonstrate on hardware that our policy successfully identifies and maintains caution around real-world pedestrians who exhibit behaviors that are not part of the training distribution.1

II. RELATED WORK

We first note high-level procedural novelties of our work and then distinguish our approach from closely related work.

Traditional approaches to social robot navigation explicitly predict human trajectories and then plan paths around them. Because these approaches use explicit human models, they can be adapted to detect when a pedestrian consistently deviates from these models, and then respond by calculating a conservative path that still maintains pedestrian comfort and safety [8]–[15]. Some RL-based policies also incorporate explicit pedestrian trajectory predictions [16], [17], but the vast majority do not. Our work extends the anomaly detection and response process to RL-based policies that implicitly model agent interactions and are thus not directly amenable to techniques designed for explicit human models.

The evaluation procedures in most prior work in RL-based social robot navigation are ill-suited for determining policy performance under significant deviations from the assumed pedestrian model (i.e., under domain shift) because (i) the evaluations are conducted using the same pedestrian model as was used during training, albeit in randomly generated scenarios and (ii) they usually only report average performance values, which reveal very little about policy performance in particularly difficult and unfamiliar situations [18]–[24]. To better evaluate and quantify policy performance under domain shift, we evaluate our policies on pedestrian models that are outside of the training distribution, and we report Conditional Value at Risk (CVaR) values, which describe expected performance on the hardest cases [25].

One approach to addressing the domain shift problem in RL-based social robot navigation could be to train on more realistic data (e.g., higher-fidelity pedestrian models or real-world pedestrian data) [26], [27]. While this approach would expand what is included in the training distribution, there are undoubtedly myriad situations and behaviors that still lie outside the training distribution, so the need to identify and account for these unfamiliar situations still persists. In our paper, we attempt to express this gap in realism by training on a relatively simple and homogeneous pedestrian model and testing on scenarios that include a mix of three different pedestrian models with randomized parameters.

A collection of papers manage pedestrian unpredictability by training the robot to avoid regions around pedestrians called “Danger Zones” or “Warning Zones” that comprise all their physically plausible next states [28]–[30]. The size and shape of these Zones depend on pedestrian velocity and observed demographic (e.g., child vs. adult). In our approach, we directly adjust each pedestrian’s discomfort distance instead of defining additional Zones, and these adjustments are based on inferred unpredictability values that describe the RL policy’s training limits, rather than being directly related to the pedestrians’ observed physical features.

One prior work quantifies pedestrian deviation from a given model, and then adaptively switches from a fast to slow RL policy when any pedestrian within a neighborhood of the robot is deemed unpredictable [31]. This approach forces the robot to respond either efficiently or cautiously towards all surrounding pedestrians, even if the majority of nearby pedestrians are acting predictably. In contrast, our approach identifies specific individuals around which the robot should be more cautious and integrates this information directly into the RL policy, allowing the robot to exercise individualized caution as appropriate while still navigating efficiently around all other predictable pedestrians.

Our work builds on SARL [6], which is an RL-based method for crowd-aware robot navigation that predicts the optimal robot action given the current state of the robot and the configuration of the crowd. While SARL performs well around the types of pedestrians on which it was trained, it is highly dependent on its training data [19], generalizes poorly to novel situations (as we demonstrate in Section IV), and has no way of recognizing when it is in an unfamiliar situation.

III. OUR UNCERTAINTY-AWARE RL POLICY

To reduce SARL’s dependence on its training data, allow the policy to recognize when it is in an unfamiliar situation, and improve the ability of the policy to generalize to novel scenarios, we modify the training process (§III-A), the model architecture (§III-B), and the reward function (§III-C).

A. Training Process

Our uncertainty-aware RL policy is trained in a modified CrowdSim environment [6], where we generate arbitrarily many pedestrians with randomized initial positions and goals.1

---

1Code for reproducing our methods and analysis is available on GitHub: https://github.com/sarapohland/stranger-danger.
By default, pedestrians choose their action at each time step based on the ORCA policy [32]—a navigation strategy commonly used to model human navigation behavior [2]. For a pedestrian with a preferred velocity of $v_{\text{pref}}$, an ORCA action, $\vec{a}_{\text{ORCA}} \in \{\vec{v} \in \mathbb{R}^2 : ||\vec{v}||_2 \leq v_{\text{pref}}\}$, comprises $x$ and $y$ velocities and makes progress towards a goal while avoiding collisions with other agents. To generate quantifiable deviations from this policy and systematically produce highly-heterogenous pedestrians for training, we augment the policy with Gaussian noise. Each pedestrian is instantiated with a deviation value $\rho \sim U(0, \rho_{\text{max}})$ for $\rho_{\text{max}} \in [0, 1]$, which represents how much the pedestrian deviates from the default ORCA policy. At each time step, the pedestrian takes an action $\vec{a} = (1-\rho)\vec{a}_{\text{ORCA}} + \rho \vec{a}_{\text{rand}}$, where $\vec{a}_{\text{rand}} \sim N(0_2, v_{\text{pref}}I_2)$ is a 2D Gaussian-random action. We call this noisy policy Noisy ORCA to differentiate it from the standard ORCA policy. The left plot in Figure 2 provides one example of Noisy ORCA pedestrians. We intentionally do not ensure that this action is collision-free and rational, as real people may take actions that appear irrational and result in collision.

We found that successfully training an RL policy on Noisy ORCA pedestrians is not trivial. These pedestrians generate spurious signals from their random motion, making it difficult for the robot to simultaneously learn how to exploit behavioral patterns in ORCA while also avoiding the unpredictable deviations from this policy and systematically produce highly-heterogenous pedestrians for training, we augment the policy with Gaussian noise. Each pedestrian is instantiated with a deviation value $\rho \sim U(0, \rho_{\text{max}})$ for $\rho_{\text{max}} \in [0, 1]$.

C. Reward Function

The reward function used to train our RL policies encourages the robot to reach its goal while maintaining social norms and avoiding collisions with people. In an environment with $n$ pedestrians, where $d_i$ is the distance from the robot to the $i$th person and $d_k$ is the distance from the robot to its goal, the default (i.e., $\rho$-independent) reward is:

$$r = k_{\text{succ}}H(-d_k) + k_{\text{coll}} \sum_{i=1}^n H(-d_i) + k_{\text{disc}} \sum_{i=1}^n \min\{0, d_i - d_{\text{disc}}\},$$

where $H$ is the step function and $d_{\text{disc}} = 0.1$ is a constant referred to as discomfort distance. In this function, the first term rewards the robot for reaching its goal, the second penalizes it for colliding with a person, and the third encourages it to maintain a comfortable distance from each person.

To incorporate the intuition of avoiding close interactions with unpredictable pedestrians while freely navigating around predictable ones, we modify the discomfort distance in the reward function to be $\rho$-dependent. Given a deviation value $\rho_i$ for the $i$th pedestrian, their discomfort distance in the modified function is $d_{\text{disc}}(\rho_i) = a\rho_i + b$, where $a = 1.0$ and $b = 0.2$ for our experiments. Everything else from the initial reward function remains unchanged. We call this $\rho$-dependent reward function the modified reward function.

IV. EXPERIMENTAL EVALUATION

We conduct two simulated ablation studies of our RL policy to analyze how the policy behaves in various situations. We also implement our RL policy on a physical robot and discuss some takeaways from our hardware experiment.

A. Simulation Experimental Setups

We evaluate our RL policies in a modified CrowdSim environment by conducting randomized episodes across six distinct categories of robot-pedestrian interactions (circle and perpendicular crossing, oncoming and outgoing flow, single and perpetual random goals) comprising a superset.
of scenarios presented in prior works [24], [33]. We perform two sets of experiments in this environment using (i) Noisy ORCA pedestrians and (ii) pedestrians operating according to more realistic but varied policies. Noisy ORCA pedestrians are quantifiably diverse, allowing us to evaluate our policies under a formalized concept of domain shift. However, because real people do not move with Gaussian random noise, Noisy ORCA pedestrians cannot be expected to reflect true deviations in human behavior. For more realistic deviations, we design experiments where pedestrians operate under standard ORCA [32], CADRL [34], and Linear policies with a wide range of different parameters. In these experiments, each pedestrian is uniformly randomly assigned one of these three policies, and the parameters for their assigned policy are uniformly randomized as well. These pedestrians present realistic but previously unseen and highly heterogenous pedestrian behavior in our suite of scenarios.

B. Performance Metrics

To compare navigation policies, we evaluate the robot’s ability to efficiently navigate to its goal while preserving the safety and comfort of surrounding pedestrians using the following metrics: (1) Success rate: percentage of trials where the robot successfully reaches its goal within 30 seconds. (2) Timeout rate: percentage of trials where the robot fails to reach the goal in the allotted time. (3) Collision rate: percentage of trials where the robot collides with at least one pedestrian. (4) Relative navigation time: time required to navigate to the goal (relative to the fastest time). (5) Relative path length: distance traveled by the robot to its goal (relative to the shortest path). (6) Number of collisions: total number of collisions between the robot and any pedestrian across all trials. (7) Personal space cost: overall personal space cost incurred by the robot (as defined by [35] with parameters from [36]). (8) Personal space violation: percentage of time spent within the personal space of a pedestrian (as defined by [37]). (9) Intimate space violation: percentage of time spent within the intimate space of a pedestrian (as defined by [37]). Metrics 1 – 5 describe robot path efficiency, while metrics 6 – 9 quantify the comfort of nearby pedestrians.

C. Simulation Results & Analysis

Our ablation studies analyze the cumulative impact of modifying the training process (§III-A), the model architecture (§III-B), and the reward function (§III-C) of the original SARL policy. Since these modifications have sequential dependencies, we define our ablation study as follows: the original policy with no modification is referred to as SARL, the policy with only the modified training process is referred to as Training, the policy with both the modified training process and model architecture is referred to as Model, and the policy with all three modifications (training process, model architecture, and reward function) is referred to as Reward or “our full uncertainty-aware policy.” We show that the combination of these three policy modifications improves policy performance on particularly complex scenarios.

1) Ablation Study on Noisy ORCA Pedestrians: We run 500 trials with increasingly noisy pedestrians to quantify the policies’ performance under domain shift. In the left plot of Figure 4 we see that ORCA and all variations of the socially-aware RL policy perform comparably when the pedestrians navigate with only small deviations from ORCA (i.e., $\rho_{\text{max}} < 0.3$). However, as the maximum randomness of the pedestrians increases, the performance of all of the policies drops significantly except for that of Reward. We see the most significant performance drops for ORCA and standard SARL, which is expected because they implicitly expect pedestrians to behave according to the original ORCA policy. We also see a reasonably large drop for Training and Model, which indicates that the modified reward function in Reward is crucial for learning robot responses that generalize well to significant changes in pedestrian behavior.

2) Comparing Different Discomfort Distances: Since the greatest performance improvement comes from using a $\rho$-dependent discomfort distance in the reward function, it is natural to suspect that we can improve performance by simply tuning the constant $\rho$-independent distance in the original reward function. We explore this in the right plot of Figure 4 The policy trained with a discomfort distance of 0m frequently collides with pedestrians, while policies trained with discomfort distances greater than 0.1m often time out
from being too cautious, resulting in low success rates. The policy trained with a discomfort distance of 0.1m does well when \( \rho_{\text{max}} \leq 0.3 \) but does not generalize to higher variability. Our adaptive discomfort distance results in the best performance across all levels of pedestrian unpredictability.

3) Ablation Study on Diverse, Realistic Pedestrians: We further evaluate the impact of uncertainty integration on navigation by conducting trials with more realistic pedestrian behavior and summarize our results in Table I. These trials contain a mix of up to twenty pedestrians in 100 randomized scenarios. Each pedestrian is uniformly randomly assigned behavior from the unpredictability feature \( \rho \) modified reward function is critical in learning generalizable behavior from the unpredictability feature \( \rho \). Overall, Reward reduces the number of collisions by 82% relative to standard SARL and reduces the time spent in pedestrians’ personal space by 6.2 percentage points on average and by 16.3 to 19.1 percentage points for hard cases. It also reduces time spent in their intimate space by 2.0 percentage points on average and by 9.2 to 11.2 percentage points for hard cases.

Thus, our proposed modifications are able to generate RL policies that notably improve pedestrian safety and comfort in particularly challenging scenarios while still maintaining good expected performance overall.

D. Robotic Experiment

To evaluate the ability of our uncertainty-aware policy to identify and respond to novel situations in the real world, we implement our full uncertainty-aware policy on a TurtleBot with a single, front-facing RGBD camera. We use a pre-trained YOLOv3 neural network [38] to detect pedestrians and design an algorithm to obtain their state information and track them over time. We compare the behavior of the standard SARL policy to our full uncertainty-aware policy in three familiar scenarios containing predictable pedestrians (Crossing, Passing, and Overtaking) and two unfamiliar

<table>
<thead>
<tr>
<th>Navigation Policy</th>
<th>Success Rate</th>
<th>Timeout Rate</th>
<th>Collision Rate</th>
<th>Relative Navigation Time</th>
<th>Relative Path Length</th>
<th>Number of Collisions</th>
<th>Personal Space Cost</th>
<th>Personal Space Violation</th>
<th>Intimate Space Violation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORCA</td>
<td>69%</td>
<td>0%</td>
<td>31%</td>
<td>1.09 (1.32 / 1.56)</td>
<td>1.04 (1.24 / 1.36)</td>
<td>156</td>
<td>24.5</td>
<td>91.9 / 124.7</td>
<td>31.4%</td>
</tr>
<tr>
<td>SARL Training</td>
<td>74%</td>
<td>0%</td>
<td>26%</td>
<td>1.40 (1.92 / 2.06)</td>
<td>1.27 (1.64 / 1.72)</td>
<td>153</td>
<td>10.4</td>
<td>47.2 / 58.2</td>
<td>17.0%</td>
</tr>
<tr>
<td>SARL Model</td>
<td>85%</td>
<td>2%</td>
<td>13%</td>
<td>1.44 (2.11 / 2.22)</td>
<td>1.30 (1.79 / 1.92)</td>
<td>39</td>
<td>13.5</td>
<td>49.5 / 57.6</td>
<td>13.5%</td>
</tr>
<tr>
<td>SARL Reward</td>
<td>87%</td>
<td>2%</td>
<td>11%</td>
<td>1.39 (1.77 / 1.86)</td>
<td>1.31 (1.61 / 1.66)</td>
<td>28</td>
<td>7.7</td>
<td>61.0 / 76.8</td>
<td>13.1%</td>
</tr>
</tbody>
</table>

TABLE I: Ablation study of our uncertainty-aware social navigation policy and two baseline policies (ORCA and SARL) on 100 randomized scenarios with up to 20 pedestrians running ORCA, CADRL, and Linear policies. In addition to average values, we report (10% CVaR / 5% CVaR) values to describe expected performance on the hardest 10% and 5% of all trials. We see that integrating uncertainty-awareness allows the policy to generalize better to novel and challenging situations.
scenarios containing unpredictable pedestrians (Standing and Stopping). Since ORCA pedestrians seen during training are constantly moving, Standing and Stopping behaviors are not part of the training distribution, so policies trained on ORCA pedestrians are limited in their ability to accurately predict future actions of pedestrians exhibiting these behaviors. Hardware evaluations on more unusual and haphazard pedestrian motions (e.g., zig-zagging) are not performed because the robot’s responsiveness is limited by long image-processing times and a narrow camera field-of-view. A more capable hardware platform would support more extensive experimentation in future work. Regardless, we find that our full uncertainty-aware policy behaves similarly to SARL in familiar situations, while our policy successfully identifies unpredictable pedestrians and keeps a larger distance from them in novel situations, as compared to the original SARL policy. Therefore, the key high-level behaviors of our approach transfer from simulation to hardware in these scenarios. Video clips for the Crossing and Stopping episodes of our robotic experiment are included in our video.2

V. Extensions to Other RL Policies

In our work, we demonstrate the need to train and evaluate RL-based social navigation policies with the consideration of domain shift. While we focus on modifying SARL [6], our proposed modifications are not restricted to this particular policy. Even given significant advances in model architectures for RL-based social navigation, the overall framework of many policies remains conducive to these modifications.

Training process: We develop a curriculum training process, where the RL policy is initially trained as normal. As training progresses, Gaussian noise is increasingly added to the pedestrians’ actions. While there are many approaches for modeling pedestrian behavior during RL policy training, the training process of any policy can be modified in this way as long as the action values for pedestrians in the environment are accessible. Regardless of how pedestrian actions are determined, each pedestrian can be initialized with a corresponding deviation value ρ and their actions adjusted with ρ-dependent Gaussian random noise.

Model architecture: We propose a modification to the RL policy model architecture by (1) training an uncertainty estimation network and (2) incorporating the uncertainty estimations as agent-level features in the observation space. The uncertainty estimation network can be developed entirely independently from the RL policy, so this component is completely policy-agnostic. For the uncertainty estimations to be seamlessly incorporated into the original RL policy, the observation space of the original policy must contain agent-level features (e.g., position and velocity values for each nearby pedestrian). This is true for many existing policies. Additional modifications would have to be made for end-to-end RL policies that operate directly on raw sensor measurements (e.g., images or 2D lidar).

Reward function: We propose a modification to the reward function that encourages the robot to maintain additional space around pedestrians that deviate from the assumed model of pedestrian behavior. To make this same modification in other RL policies, their reward function simply must contain some notion of “safety space,” “discomfort distance,” or “clearance” that captures a sense of maintaining proper distance from pedestrians. If this is the case, the modification of increasing each agent’s discomfort distance based on their deviation value ρ is subsequently straightforward, though specific constants may need to be tuned for the particular model of interest.

VI. Conclusions

In this work, we articulate the domain shift problem for RL policies in social robot navigation and present an approach that improves generalizability of RL policies to novel scenarios while maintaining their efficiency in familiar ones. We find that SARL [6] generalizes poorly to significant deviations in pedestrian behavior, thereby presenting serious concerns for pedestrian safety and comfort in a real-world deployment. We posit that for socially-aware RL policies to be viable in real-world mobile robots, these policies must recognize when people deviate from the (implicitly) assumed pedestrian model and take appropriate caution. We present effective methods for modifying the training process, the model architecture, and the reward function of SARL that substantially improve the generalizability of the policy. Comparing our modified policy to the original SARL policy on randomized scenarios containing realistic ORCA and non-ORCA human policies, our modifications reduce the number of collisions by 82% and reduce the proportion of time spent in the pedestrians’ personal space by 16 percentage points for the hardest 10% of all trials and by 19 percentage points for the hardest 5% of trials. This increase in pedestrian comfort is achieved while maintaining similar navigation times and path lengths. We also discuss how these same modifications can be applied to other socially-aware RL policies.

While we believe our work takes an important step toward enabling the deployment of socially-aware RL policies on mobile robots, there are some limitations that should be addressed. First, we modify the reward function to encourage the robot to maintain greater space between itself and unpredictable pedestrians. While this heuristic for caution is reasonable in many situations, it is less effective in tight spaces, where the robot is unable to maintain such a distance. It would be interesting to explore other heuristics, such as slowing down, speeding up, or some combination of adjusting distance and speed when approaching unpredictable pedestrians. Another limitation is that we use the ORCA policy as our primary pedestrian model for this study because this is the model commonly used in other RL-based social navigation work. However, this model is relatively simplistic. It would be interesting to train RL policies using other models of pedestrian behavior and evaluate generalizability to even more diverse and realistic pedestrian scenarios.

2A video summarizing our methods and results is available on YouTube: https://youtu.be/9IDhXvCC58w.


