IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 April 2024, accepted 13 June 2024, date of publication 17 June 2024, date of current version 25 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3415708

RESEARCH ARTICLE

Generating Social-Aware Locations for a Robot in a Human Group by Image Generation Using Adversarial Learning

SEVENDI ELDRIGE RIFKI POLUAN AND YAN-ANN CHEN, (Member, IEEE)
Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320315, Taiwan

Corresponding author: Yan-Ann Chen (chenya@saturn.yzu.edu.tw)

The work of Yan-Ann Chen was supported by Grant MOST 107-2218-E-155-007-MY3 and Grant MOST 110-2221-E-155-022-MY3, Taiwan.

ABSTRACT With the advance of deep learning techniques, social robots can have more powerful perception and interaction capabilities. However, the problem of finding a socially aware standing location for the robot to join a conversation group is not well addressed. Thus, we propose a generative-based and image-based approach to generate a social-aware group formation to obtain the possible locations for the robot. Furthermore, to overcome the problem of formulating human comforts, we try to leverage human behaviors with the concerns of human comforts when joining the conversation group. We utilize a self-supervised technique to generate this kind of human experience from the real-world dataset. Through extensive experiments, we show that the proposed method outperforms the social force method by 62% with respect to data from human experiences. In addition, our approach also provides controllable parameters to generate the location with the required features using the GAN noise vector.

INDEX TERMS Adversarial learning, conversation group, edge artificial intelligence, generative AI, Internet of Things, robot standing position, social robot.

I. INTRODUCTION

Nowadays, integrating social robots into human interaction undergoes a transformative evolution, propelled by advancements in AI, sensors, and control technologies. In executing various interaction tasks, robots must possess sophisticated capabilities. These may include speech recognition [1], natural language processing [2], object detection [3], face recognition [4], facial emotion recognition [5], place localization [6], object interaction [7] and lastly, social interaction [8], [9]. From an application-oriented viewpoint, social robots have a significant impact on areas such as healthcare [10], education [11], customer service [12], and social assistance [13]. However, deploying these promising technologies in real-world settings raises important concerns. A key concern is the potential discomfort that may arise from the behavior of the robots. For example, when a social robot joins a conversation, it can cause unease and changes in the communication behaviors of the group members [14], [15]. This phenomenon may be caused by unfamiliarity with technology, artificial interaction, and perceived threat to privacy and control, which can negatively impact the dynamics and effectiveness of the group, potentially resulting in trust issues for humans. Therefore, it is essential to take into account human feelings when designing these technologies.

Several studies [16], [17], [18], [19], [20] have explored the problem of social-aware navigation and group joining for robots using group formations (F-formations). For example, reference [18] estimates optimal robot placement in an O-space of F-formations for a social group interaction. However, the assumption of structured formation may not always hold in conversational groups. This presents significant challenges when dealing with unstructured groups that have unpredictable dynamics. In contrast, reference [21] disregards formations, determining the social placement of an online avatar within a group based on social force fields.

The associate editor coordinating the review of this manuscript and approving it for publication was Yang Tang.

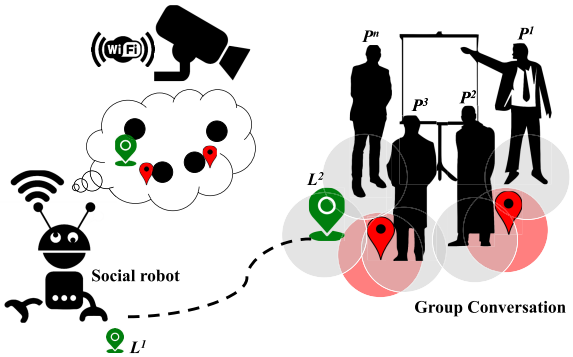


FIGURE 1. A scenario of a social robot joining a group conversation.

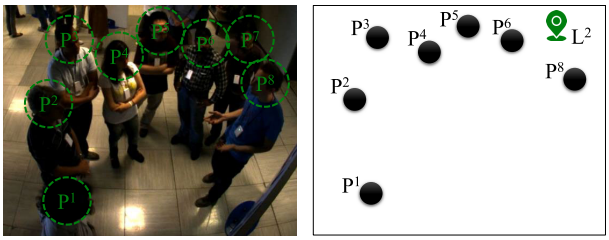


FIGURE 2. (a) The photo of a conversation group in a real scene and (b) Plot locations of all participants in this group and P^7 is removed for the reference of the desired location L^2.

This approach uses repulsion forces to maintain a minimum distance between individuals but relies on predefined rules and heuristics, which may not accurately represent real-world conditions. Moreover, the study [22] focuses on learning-based path planning with social awareness for a robot joining a group. It employs LSTM (Long Short-Term Memory) and GAN (Generative Adversarial Network) to generate the robot's movement trajectories, considering potential collisions and the field of view of individuals within the group. However, the robot halts once it is reasonably close to the group, leaving a precise and fine-grained stop location for the robot's final standing position unexplored.

This work addresses the *standing location problem* for a social robot that intends to join a conversational group. To ensure the robot's social integration, we have to consider the comfort of other group participants and identify a socially acceptable standing location. Fig. 1 shows the application scenario where a social robot is trying to join a free-standing conversational group with four participants. The problem is how we place this social robot inside the interaction range of the group with a human-centric manner. Assume that the robot can collect visual information from the surveillance camera in this environment or directly from an embedded camera. According to the position or orientations of each participant, the robot will predict a suitable standing location and navigate the robot from its current position to the computed one, i.e., from L^1 to L^2 in Fig. 1. The designated location L^2 is considered socially acceptable for the robot, taking into account each individual's personal comfort zone (indicated by the gray circles). In contrast, locations marked by red pins may be deemed unsuitable due to their proximity to group members. Note that comfort zones vary among individuals and are difficult to formulate. To address the complexity inherent in unstructured groups, we explore the possibility of learning acceptable standing locations based on observed human behaviors and experiences.

Generative AI has shown its effectiveness in understanding and executing human instructions in fields such as text generation [23] and image generation [24]. In this study, we propose a generative-based approach, PosGAN, which utilizes GAN [25] to devise a socially aware standing location. The GAN architecture is adopted because it generates high-quality synthetic outputs that preserve the key attributes of socially aware human behaviors. In addition, our method, which embraces an end-to-end principle, inputs the captured images directly into our model. This eliminates the need for intermediary data transformations such as conversion into coordinates or facing orientations. To model socially aware human behaviors, we rely on open group conversation datasets [26], [27] derived from real-world scenarios. We employ a self-supervised technique to construct the training set, which involves sequential removal of individuals from a group.

The contributions of this work are as follows. First, we address the challenge of finding a socially acceptable position for a robot to join a group conversation. Second, we propose an end-to-end approach for determining the robot's standing location in a group using the image generation technique. Third, we exploit a self-supervised method to generate our dataset for training. Finally, it provides a simplified solution to reduce engineering efforts and the requirement of prior knowledge, rather than relying on complex prediction steps.

The remainder of this paper is organized as follows. Related work is covered in Section II. Section III formulates the problem and presents our work to generate a socially acceptable location. Section IV presents evaluation results. We have a discussion of our method in Section V. Finally, Section VI concludes this paper.

II. RELATED WORK

In this section, we examine various studies on social robots, group formation, robot trajectory planning, and generative AI. A brief summary of the reviewed work is presented in Table 1.

A. SOCIAL ROBOTS

Deploying social robots offers several advantages, such as their use in stores to increase sales, with a success rate of two out of three stores reported in case studies by [28]. In addition, social robots can guide museum tours [29], which guide visitors through the museum along a predefined path, and track the visitor group [30]. To ensure effective robot navigation during such tasks, Simultaneous Localization and Mapping (SLAM) serves as a key technique. SLAM allows

TABLE 1. Summary of related work.

Ref.	Category	Outcomes	Key Inference & Challenges
[28]	Social robots	Social robots in retail	Social acceptance of social robots; Employment of social robots in stores; Cost-effective labor; Case studies on social robot implementation; Manager satisfaction and desire for reuse;
[29]		Autonomous robotic museum guide	Integration of perception; Cognitive robot architecture; Intelligent agent; Guided tours;
[30]		Robot system for guided museum tours	Location and orientation measurement; Spatial awareness with sensor poles; Group detection method; Recognition of standing groups; Utilization of laser range finder; Enhancing visitor experience;
[31]		Embedding robots in welfare services	Shift towards using robots in care; Emergence of care robotics; Socio-technical transition; Re-configuration of social and technological elements; Qualitative study on robot use in elder care; Management and policy measures;
[32]		Ethical concerns in robot use	Use of robots in elder care; Monitoring health and safety; Improving lives of the elderly; Providing companionship;
[33]		Social robot navigation in dynamic indoor settings	SLAM-based localization; Utilization of the Pepper robot platform; Validation using ROS;
[34]		Robust semantic visual SLAM for dynamic environments	SLAM Systems Development; Semantic segmentation; Local mapping for Spatial representation; Tracking in dynamic environments; Parallel threads in DS-SLAM; Dense semantic octo-tree map;
[16]	F-formations	Novel architecture for real-time social group detection	F-formation detection; Prediction of the robot's approach angle; Detection of outliers; Learning model based on CRF and SVM;
[35]		Sociological studies on "F-formation"	Robot body orientation; Spatial arrangement; Spatial relationship; Information-presenting robot; Robot's body rotation impact on F-formation arrangement;
[36]		Museum guide robot	F-formation establishment; "Pause and restart" implementation; Spatial formation ("F-formation"); Elicitation of visitor's attention towards the robot;
[37]		Socializing model for HRI	F-formation concept; Impact of mobility on social acceptance; Optimal pose for social comfort; Omnidirectional mobility;
[38]		Navigation system development	Gesture recognition; Mobile robot navigation; Robot moving pattern; Perception of robot intelligence;
[39]	Trajectory planning	Trajectory planning in dynamic environments	Rapidly-exploring Random Tree (RRT) algorithm; Potential field-based motion planning; Kinodynamic constraints in robot motion; Model Predictive Control (MPC) in robotics; Robot motion models;
[40]		Motion model for predicting agent trajectories	Safe navigation in robotics; Multi-agent encounter scenarios; Timed elastic bands in trajectory planning; Global proxemic considerations;
[41]		Socially acceptable path planning	Urban city path planning; Navigation management policies; Shared spaces for pedestrians; Optimal route selection strategies; PMD navigation; Imitating PMD behavior; Social force models;
[42]		Safe navigation for robots	Traditional approach treating pedestrians as obstacles; Algorithms for collision-free motion; Social force models; Human-like collision avoidance;
[43]		Socially acceptable trajectories for autonomous platforms	Multimodal human motion behavior; Recurrent sequence-to-sequence model; GAN integration; Novel pooling mechanism for information aggregation; Collision avoidance capabilities; Computational complexity considerations;
[22]		Social-aware navigation for crowds	Modeling social-aware space; Virtual character behaviors in crowds; Static and dynamic group formations; Fast marching method for pathfinding; Production of speed map for navigation;
[44]	Generative AI	Translation of Aerial IR Recordings to RGB Images	Deep Learning architecture; Inferring temperature information; Drone technology; Improved Conditional-Generative Adversarial Network (IC-GAN); U-Net-based generator; Color space transformation;
[45]		Reverse GAN system for complex image captioning	Reverse Generative Adversarial Network (ReverseGAN); Image-to-image conversion; Graph Convolutional Neural Network (GCN); Attention mechanism;
[46]		Self-supervised collaborative synthesis in medical images	Medical image analysis; Multi-source-modality images; Auto-encoder network; Missing data imputation;
[47]		Self-supervised learning in medical image analysis	Framework development (DiRA); Computer vision; Deep semantic representation learning; Lession localization; Adversarial and restorative learning;
[48]		Self-supervised visual relationship detection	Masked Bounding Box Reconstruction (MBBR); Context-Aware Representations; Representation learning; Few-shot learning;

robots to create and update maps of their environment in real-time. For instance, [33] employs an efficient SLAM-based localization and navigation system for service robots in dynamic indoor environments. Furthermore, [34] utilizes a semantic visual SLAM approach for intelligent mobile robots in dynamic environments. However, the problem of evaluating the comfort of group participants when a social robot joins the conversation is not well addressed.

Previous studies [31], [49] have reported that neither caregivers nor care recipients have explicitly expressed interest in the care robot, and ethical and legal concerns about the use of care robots remain unresolved [32]. Additionally, spatial dynamics within discussion groups are critical. Reference [50] highlights the importance of peripersonal space, a protective buffer around the human body, where intrusion can cause discomfort. Trust is another key concern. As noted by reference [51], robot assistants need to ensure safety, comfort, and trustworthiness to avoid frustrating Human-Robot Interaction (HRI) experiences or severe consequences [52]. Thus, social acceptance is an important issue when creating the functionality of the social robot. In this work, we investigate the problem where the social robot can seamlessly participate in group conversations while also considering the comfort of human participants.

B. F-FORMATIONS

Several studies have employed F-formation techniques in the integration of social robots into human environments. Reference [16] utilizes F-formations using spatial coordinates extracted from human bodies to determine the ideal angle for robots to approach a group. Reference [17] introduces a framework that combines pose prediction with socially aware robot navigation, building upon the F-formation algorithm's graph-cuts and human motion data. Reference [19] further applies deep learning to model robot approach behavior, carefully maintaining the F-formation through strategic repositioning and reorienting. Reference [20] uses spatial formations to propose the best placement for a mobile robotic telepresence system or a virtual agent in a simulated environment to a social group interaction. Further, reference [35] emphasized the establishment of correct spatial relationships with individuals, making keen use of F-formations. While, reference [36] introduced a model for a mobile museum guide robot that can adhere to F-formations and execute ''pause and restart'' strategies efficiently. Lastly, reference [37] proposed a socializing model that allows a robot to maneuver to its most socially optimum position within F-formations during group interactions with humans.

Although F-formation techniques have shown promise in enabling social robots to join structured group conversations, they have limited effectiveness in less organized conversation groups. To address this concern, reference [38] developed a topology map-based approach to robot navigation during human-robot interaction tasks while considering the comfort zone of people. However, this method still relies on a rule-based approach to position the robot when approaching a group and may produce less-than-ideal outcomes owing to limitations such as unstructured conversational groups or dynamic environmental conditions. Therefore, future research should focus on developing more advanced techniques that allow social robots to join group conversations while considering the comfort of human participants, without limiting only to the F-formations. This work aims to contribute to addressing this research gap.

C. TRAJECTORY PLANNING

In the field of social robotics, various studies have concentrated on trajectory planning and robot navigation in dynamic human environments. Reference [39] developed a Rapidly-exploring Random Tree (RRT) based trajectory planning algorithm. For navigation, reference [40] proposed an approach to ensure safe and legible navigation in multi-agent encounters through implicit cooperation between humans and robots. Comparing different navigation techniques in terms of mobile robot trajectories, Reference [53] carried out an analysis using various ROS-based SLAM systems. Reference [54] presented a real-time SLAM technique that estimates robots' trajectories and creates 3D maps using a 3D shape matching method. With a focus on the social aspect of navigation, reference [41] worked on a socially acceptable global route planner and evaluated its legibility. In contrast, reference [42] introduced a system in mobile robots that mimics human-like collision avoidance using a pedestrian model from human science. Highlighting the prediction of future actions, reference [43] used a sequence-to-sequence model to anticipate potential actions of agents and stimulate diverse predictions. A similar study, reference [22], developed a trajectory prediction model for social robots that navigate in groups engaged in open discussions. However, these studies do not thoroughly address where a robot should stop and stand when reaching its target group. This is a crucial issue that our work attempts to solve by suggesting a socially aware stopping location that can aid existing trajectory planning methods.

D. GENERATIVE AI

Generative AI is a form of artificial intelligence that utilizes vast datasets and deep learning models to produce new content such as images, text, or audio through automatic creativity and imagination. It is frequently combined with self-supervised techniques to provide abundant learning input. For example, in aerial technology, reference [44] uses an improved Conditional Generative Adversarial Network (IC-GAN) to translate aerial infrared recordings into RGB images. In image captioning, reference [45] introduces the Reverse Generative Adversarial Network (ReverseGAN) with a graph convolutional neural network. Furthermore, reference [55] enhances pathological image super-resolution with the help of MASR-GAN. In healthcare, reference [56] forecasts the demand for healthcare service

using an attention-free model and self-supervised Generative Summary Pretraining (GSP), exceeding existing baselines. Addressing missing data in medical images, reference [46] presents a self-supervised collaborative learning framework that demonstrates superiority in generalization and specialty. For Visual Relationship Detection (VRD) representation learning, reference [48] develops a self-supervised approach, Masked Bounding Box Reconstruction (MBBR), for learning relationship-aware object representations and obtains improved results in predicate detection within sentences. In the field of graph anomaly detection, reference [57] presents a self-supervised approach named SL-GAD that significantly improves performance over existing methods. In medical imaging, reference [47] introduces a unified self-supervised learning framework, DiRA, to collaboratively generate fine-grained semantic representations. Inspired by the success of these generative AI strategies in various fields, our study uses a generative adversarial network with self-supervised techniques to tackle challenges in providing a socially aware location within a group conversation.

III. PROBLEM FORMULATION AND SYSTEM DESIGN

In this work, we address the challenge of recommending an appropriate standing position for a robot to join a free-standing conversation group, taking into account the comfort levels of the participants with the newcomer's movements. Estimating comfort levels as a function is significant complexity due to the challenges associated with quantifying the human mind. Alleviating this issue, we operate by inferring social awareness from empirical human behavior data using learning techniques. We first utilize the technique of self-supervised learning to obtain data on human behavior. We then introduce *PosGAN* an image generation method, which is based on conditional generative adversarial network [58], to learn how to draw the full social-aware formation when inputting the current group structure. After that, we find one unoccupied position in the plotted formation to be the robot's position. Fig. 3 shows the architecture and procedure of our system.

A. PROBLEM DEFINITION

This paper addresses the standing location problem considering acceptances of other human members when a robot is a member of a free-standing conversation group. Thus, we have to care for the comforts of other human members for determining the location. These comforts may come from the distancing and facing direction. Assume that there are n human members, $H = \{h_1, \ldots, h_n\}$, of a free-standing conversation group. We denote the status of a member i by S_i. Status information may contain the location, facing direction, or posture of the member where only location is mandatory. A social robot r is going to participate into the group. We then define a function $C(S_{h_i}, S_r)$ to represent the comfort level of a human participant i with respect to the robot. The result of $C(\cdot)$ may be related to social distancing or physical contact.

Here, we ignore the comfort of a human member toward others. We formulate this standing location problem as an optimization problem. Given the statues of all the human members who already form a conversation group, we try to compute

$$S_r^* = \arg\max_{S_r} \sum_{i}^{n} C(S_{h_i}, S_r). \tag{1}$$

The location of S_r^* is the final standing location for the robot to join the group.

For measuring comfort levels, it is complicated to define the function $C(\cdot)$ by a mathematical formulation. Therefore, we try to utilize the location determined by humans as the reference and exploit the concept of self-supervised learning [59] for learning. If our predictions align closely with the locations or group formations as determined by humans, we assume that we can achieve higher comfort levels. However, deviations from these human-determined positions or structures will result in reduced comfort levels, corresponding to the degree of displacement. Fig. 2(a) illustrates an example of a conversation group where each participant's position can be determined based on their head location. These positions are then represented by solid circles in Fig. 2(b). The self-supervised technique leverages the dataset itself to acquire insights. For example, by removing the circle P^7 and designating L^2 as the ground truth for learning, we can deduce that L^2 is one of the socially acceptable locations for joining this group.

B. DATA PREPARATION

The input to our model is an image frame j of a conversation group, which may be cropped from the image captured by the surveillance camera. In the initial stage, our Head Detection Module (HDM) plays a crucial role. HDM identifies the position of each participant based on their head position in the frame, utilizing a deep neural network that incorporates R-CNN under the ResNet-50 architecture. In particular, this R-CNN model is trained using the SCUT-HEAD dataset. HDM marks these positions with symbols (e.g., solid circles) on an image y_j. To generate training data, we remove the symbol of participant k from y_j to be another image x_j^k where k is the participant index. Thus, (x_j^k, y_j) is the pair for training. To augment the training pairs, we do the above procedure for each participant.

C. STANDING LOCATION GENERATION (POSGAN)

Our generator and discriminator architectures are based on [58]. For one pair (x_j^k, y_j), we then input a random noise vector z and x_j^k into the generator where z will be regenerated each time and x_j^k is used as conditional information to control the generated image. During the training period, the generator produces an image m_g^{jk} that may add some new symbols to the image and form a socially acceptable structure. Then the discriminator uses the images m_g^{jk} and y_j to learn if they are generated one and real one, respectively. The loss functions

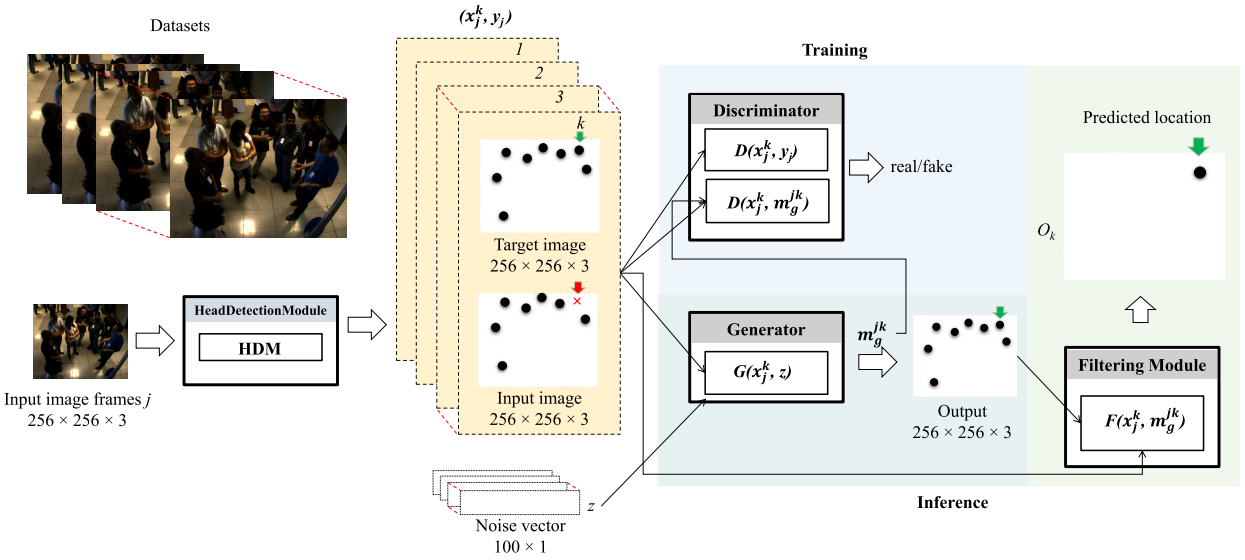


FIGURE 3. PosGAN architecture: socially aware standing location generation in a human group.

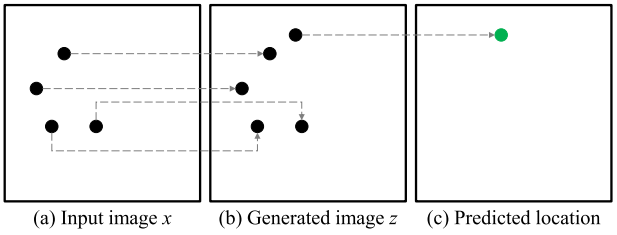


FIGURE 4. Example of determining the predicted position by filtering.

$\mathcal{L}$ of our GAN model are as follows.

$$\begin{aligned}
&\text{objective} \quad \min_G \max_D \quad \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \\
&\mathcal{L}_{cGAN}(G, D) = \frac{1}{w} \sum_j^w \log D(x_j, y_j) \\
&\quad + \frac{1}{w} \sum_j^w \log\left(1 - D(x_j, G(x_j, z))\right) \\
&\mathcal{L}_{L1}(G) = \frac{1}{w} \sum_j^w \|y_j - G(x_j, z)\|_1. \qquad (2)
\end{aligned}$$

G and D represent the generator and the discriminator, respectively. w represents the number of training samples, and λ is a hyperparameter to control the clearness of the generated image.

For inference, when we input an image with marked symbols to the trained generator, it will add some symbols to the image according to the social norm behaved by humans. To decide the stop-standing location of the robot for joining the conversation group, we remove the symbols that are in the same locations between the input image and the generated image in the filtering module. Fig. 4 shows the concept that the remaining symbol is the position that we recommend to the social robot. But, sometimes, the remaining symbols are more than one. To select one location, we perform heuristic filtering on the generated image. One possible heuristic filtering is to randomly choose one symbol as the output location. Finally, we have to extract the location in the coordinates (x,y) by examining the area with black pixels and finding the center of the symbol as the location. Then, we may map the location from the image coordinate to the coordinate in the environment to navigate the robot.

D. IMAGE SYMBOLS

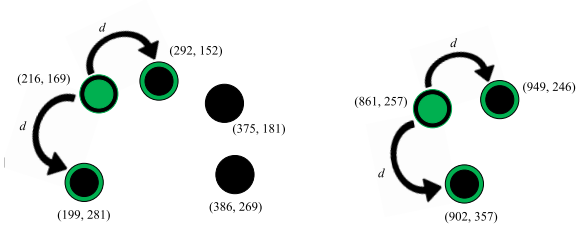
Previously, we address that the image symbol for representing the positions of participants is a solid circle. But, our approach may support replacing the symbol of solid circles with the symbol of different shapes. The advantage of using different symbols is that the new symbol may be used to represent more information. For instance, we may utilize a solid triangle to represent the locations and also the facing direction of the person. If we change to use other symbols, we simply use that symbol in the data preparation step, in Section III-B, and we do not have to make any modifications to the following steps.

If we exploit the solid triangle as the symbol, we use the isosceles triangle with an obtuse vertex angle, where the vertex angle is used to point toward the facing direction. For extracting the location from the generated image with the solid triangle, we first find the symbol by examining the region of black pixels and use the center of the longest edge (the edge opposite to the vertex angle) of the triangle as the location for the robot.

IV. EXPERIMENT RESULTS

A. DATASETS

We use the SALSA dataset [26] which provides detailed annotations of group formation and participants' movements



(a) 5-member group (b) 3-member group

FIGURE 5. Computation of ED2C where the solid green circles represent a generated location. The ED2C is calculated by finding the Euclidean distances from the generated location to its nearest two neighbors.

in two social events, a cocktail party and a poster session. These events involve various dynamic social interactions, such as group formations, conversations, and movements, which offer diverse scenarios for our study. We mainly use the position information of each individual in a conversation group. For preparing the dataset, we use the poster session of SALSA as the training set and the cocktail party of SALSA as the testing set. In our data cleaning process, we excluded one-person groups and eliminated overlapping points on plots to disregard extreme cases. Using self-supervised techniques, we eliminated one participant at a time from each frame of a group. For further data augmentation, we also relocated the group to various positions in the space, maintaining the structure of the group. This approach resulted in the generation of 21, 419 training samples and 18, 925 testing samples for our experiments. The sample numbers of group size $\{2, 3, 4, 5, 6, 7, 8\}$ for training are $\{4297, 4839, 6098, 2143, 2933, 1025, 84\}$ and those of group size $\{2, 3, 4, 5, 6, 7\}$ for testing are $\{6407, 6334, 3389, 2124, 621, 50\}$.

B. COMPARISON METHODS

1) BASELINE

We adopt a random approach as the baseline method to compare with ours. This method simply randomly generates a location in (x,y) coordinates within the image dimension. In our expectation, this method may work the worst since it does not consider any information about group formation or robot information.

2) INCEPTION

We use Inception v3, a deep convolutional neural network [60], for the comparison. Since this model cannot generate the position on the image, we use the locations in (x, y) coordinates of each group member as input. Let the final output also be a location in (x, y) coordinates. The model was trained for 200 iterations, and we selected the best training model with the smallest MSE.

3) SOCIAL FORCE

The social force model [21] is used to determine the position to visualize an avatar in an online game conversation considering social distancing. They use a social force field to adapt to the formation of a conversation group and use a repulsion force to maintain the minimum distance between individuals. To implement this model for our scenario, we compute the average distance between each current member and the group center as the social force and compute the minimum distance between each current member as the repulsion force. Then we repeatedly generate a random location to be the possible location until the distance between the location and the center is around the social force and the distance between the other members is more than the repulsion force.

TABLE 2. Evaluation on the mean ED2C of generated locations.

Methods	mean ED2C (cm)	Displacement Error (cm) ↓
Ground truth	128.96	0.0
Baseline	921.28	792.31
Social force	80.54	48.42
Inception	632.98	504.01
PosGAN	**110.58**	**18.38**

C. PERFORMANCE COMPARISON

1) PREDICTION ERRORS

In our method, we remove one participant from a conversation group to serve as the ground truth for a socially-aware location, enabling a direct measurement of the prediction errors relative to this ground truth. The average prediction errors for the baseline, social force, Inception, and PosGAN (our proposed method) on the test set are 769.99, 271.12, 540.42, and 175.62 centimeters, respectively. These results suggest that our method provides the most accurate prediction in relation to the ground truth. However, these numerical comparisons of prediction errors may not comprehensively represent the degree of social acceptance toward the predicted results because there could be multiple locations suitable for joining a conversation group that deviates from the ground truth. Therefore, we advocate for an evaluation strategy that considers the social context.

2) DISTRIBUTIONS OF ED2C

To further compare performance, we employ a metric, *ED2C (Euclidean Distance to two Closest neighbors)*, to assess the social acceptance of the determined location within a conversational group. ED2C measures the combined Euclidean distances to the two nearest members of the conversation group, as visualized in Figure 5. The reason of selecting two nearest neighbors is that placing the robot at the generated location could directly affect the neighboring participants on its left and right. As such, ED2C can effectively represent proximity patterns within the social interactions.

To compare the performance across the entire dataset, we calculate the ED2C for all test cases and normalize the distribution of ED2C values. The normalized ED2C distribution is expected to closely align with the distribution

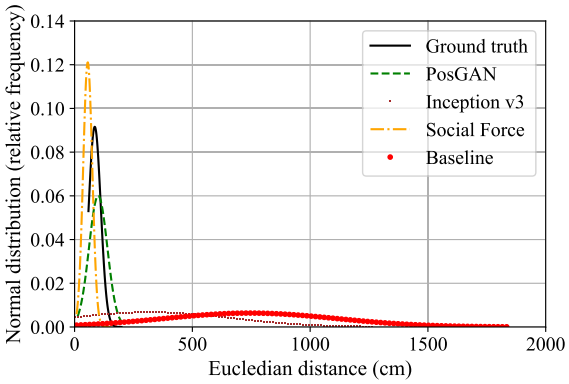


FIGURE 6. ED2C distribution of our and compared models.

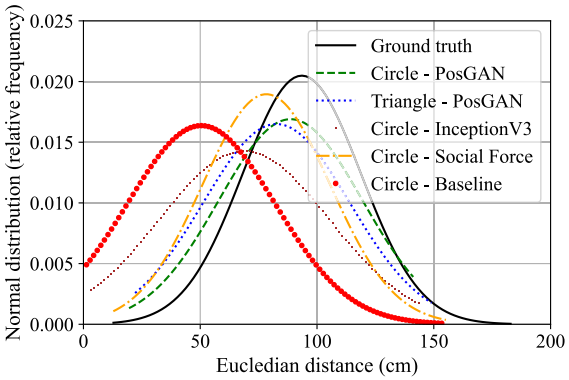


FIGURE 7. ED2C distribution of our and compared models when varying symbol shapes.

associated with the ground truth position. Fig. 6 and Table 2 show the results of the comparison. In Fig. 6, we display the histogram representing the distances to the two closest neighbors from the excluded participant throughout the test set for each method evaluated. The peak of a curve indicates the distance that occurs most frequently within the test set. We think the distribution should resemble that of the ground truth to suggest that the predicted locations mimic actual human behavior, thereby ensuring higher comfort levels for other participants. The figure shows that the distribution curve of our method is close and similar to that of the ground truth. It validates that our approach follows the social norms performed by humans. For social force, the curve is close to the ground truth one, but it will give closer distancing, which violates the human norms. For baseline and inception, their histogram curves are far from the ground truth one.

In Table 2, we present the average ED2C for each method, defined as the mean distances to the two nearest neighbors in the test set, and the displacement error, which is the absolute difference relative to the ground truth. We expect that ED2C will approximate the ground truth, indicating that the locations are on average determined in a manner similar to human decisions. As we can see, our method has the lowest displacement error and exceeds the performance of social force (the method ranked second) by approximately $(48.42 - 18.38)/48.42 = 62\%$. Although the results of social force can be close to the ground truth, it provides the locations which are closer to participants in average, and this may lower the comfort levels of others. For the baseline and inception methods, they fail to predict locations that mimic human-like decision-making.

3) VARYING SYMBOL SHAPE

We conducted an experiment to compare the performance of using a solid circle and a solid triangle as symbols. In the previous experiment, we found that the symbol size and position of the group in the image may affect the generation performance. Thus, we do a translation for the position of the group such that the center of the group is aligned to the center of the image, and we also do a scaling to enlarge the relative distances among group members. Since these transformations are linear, we can transform the result back to the original setting. Fig. 7 shows the results of varying the symbol shape. As we can see, PosGAN with circle symbols is slightly better than PosGAN with triangle symbols. The reason may be that the circle is a symmetric shape and that it is easier for the model to pinpoint the location. But, our proposed methods are better than others by showing our ED2C distributions, which are closer to that of the ground truth. Although the triangle shape may not outperform the circle one, it may be used to provide more information about the robot.

4) FACING DIRECTION

Exploiting the triangular symbol introduces the ability to estimate the facing direction of the robot while considering social acceptance. To evaluate the accuracy of the estimation of the facing directions, we utilize the root mean square error (RMSE) to measure the difference between the predicted and actual facing directions. In this experiment, we only compare with InceptionV3 since other approaches cannot be modified to predict orientation. The observed RMSEs for our triangle-symbol-based method and InceptionV3 are 0.111 and 0.421 respectively, indicating that our approach yields better accuracy. Through this, we validate the utility of different symbols in furnishing additional information. It is critical to note that the computations for position and orientation occur simultaneously and positioning inaccuracies could lead to orientation prediction errors. This simultaneous computation may be a contributing factor to why our approach outperforms InceptionV3.

D. EXAMPLES OF PREDICTION RESULTS

1) GROUP STRUCTURE IMAGES

In addition to metric-based method comparison, we show the visualization the predicted results for better understanding in Fig. 8. In the figure, each row displays the comparison with respect to one sample. The first column provides the inputs to these methods by depicting the original conversation group from which one participant has been subtracted. The second column displays the original group structure of the

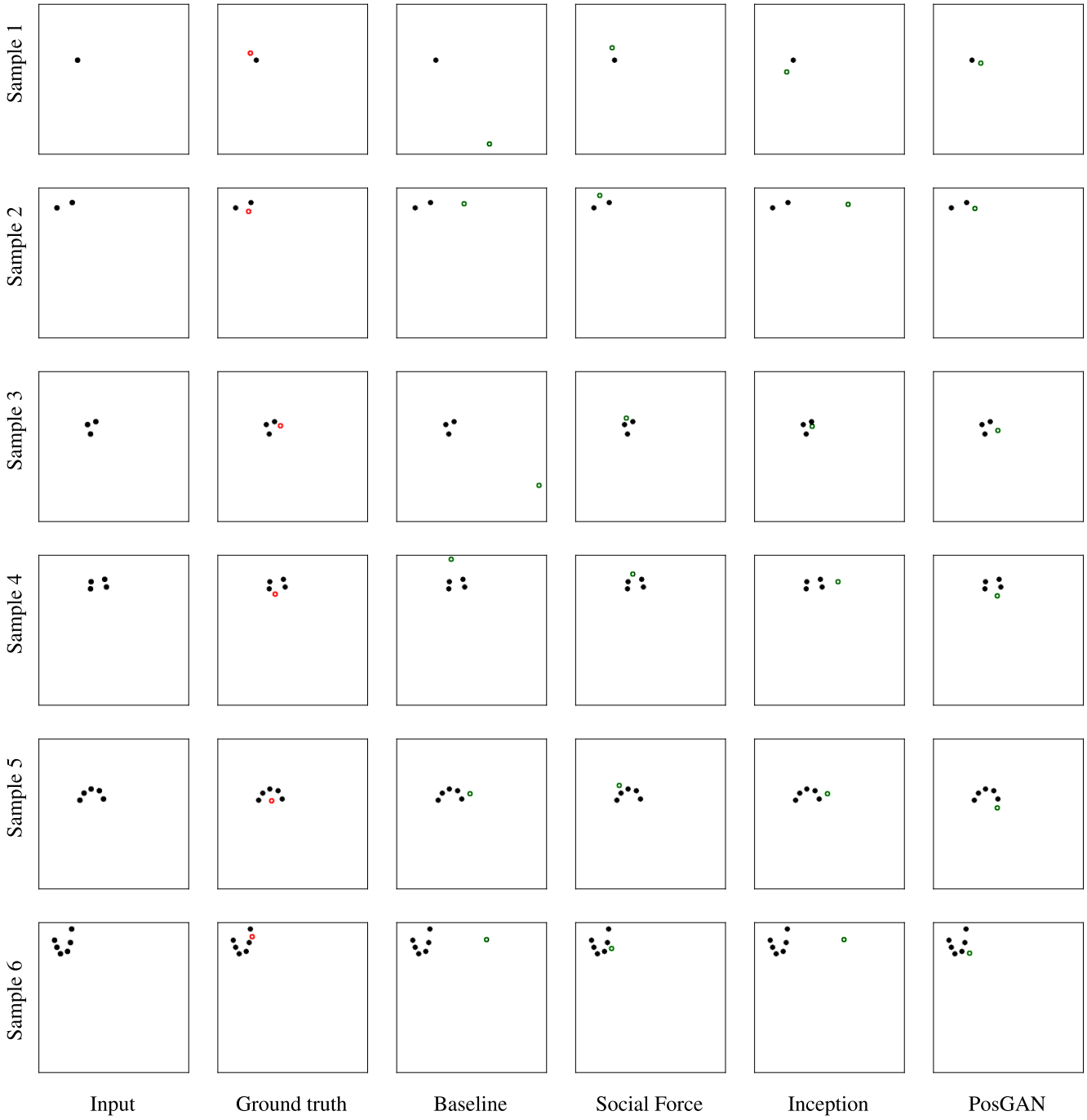


FIGURE 8. Visualized comparisons of random group samples ranging from 2 to 7 in size. (Green hollow circles represent predicted locations, while red hollow circles indicate actual positions).

sampled data, and a red hollow circle indicates the excluded participant in the input. Columns three to six depict the predicted locations derived from the baseline, social force, inception, and PosGAN methods. The green hollow circles denote the predicted positions. Note that our method produces group formations (column six) that closely resemble the actual scenarios depicted in column two. While our method's results do not exactly match the ground truth locations, we still follow social distancing norms and take into account the overall group configuration. In contrast, the social force method provides a reasonable structure, but occasionally results in tighter spacing and fails to adequately address distancing. Leveraging insights from the group structures of real-world social interactions, our method is capable of suggesting positions that conform to accepted social norms.

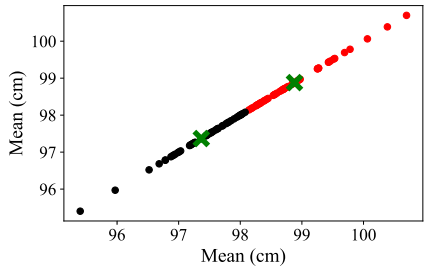


FIGURE 9. Clustering of mean distances based on varying the values in static noise vectors (Green cross indicates the center of a cluster group).

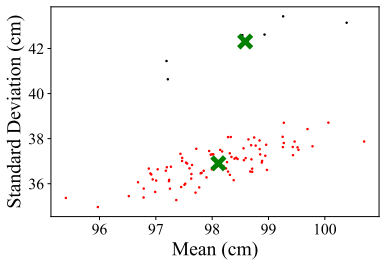


FIGURE 10. Clustering of mean distances and standard deviations based on varying the values in static noise vectors (Green cross indicates the center of a cluster group).

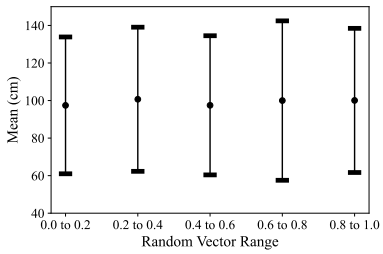


FIGURE 11. Comparison of mean distances based on random noise vectors from the five ranges.

2) REAL-WORLD SCENES

Fig. 12 shows that we make the standing location determination for each group of a captured image, utilizing a three-row structure where each row represents one sample of a standing group discussion. Our HDM can identify the position of each group member based on their head position, using a deep neural network incorporating R-CNN under the ResNet-50 architecture. We use the SCUT-HEAD dataset, consisting of 4,405 images and 111,251 head labels, to train our model for obtaining a more generalized prediction. We use all the methods to infer the standing locations. We plot the robot at the proposed standing locations in each group, and we may validate the results in the real-world scenes.

V. DISCUSSION

A. EFFECTS OF NOISE VECTOR

Several research [61], [62], [63] control the generated images of GAN by exploiting the values in the random noise vector. Here, we explore the effects of random noise vectors z on the generated group formations. After we have a trained model, we input a group image with different random vectors to investigate the influences. In our model, the tensor shape of the random vector is (100, 1). So, the possible settings for the vector are huge. Therefore, to simplify the experiment, we assume that each dimension of the random vector has no correlation.

- Static Noise Vector: In the first experiment, we assign the same value to each dimension within the noise vector. For instance, if we set the value to 0.1, the tensor z will be $[0.1, 0.1, 0.1, \ldots, 0.1]^T$ and $z \in \mathbb{R}^{100\text{x}1}$. We generate 100 different noise vector settings by varying the value from 0.01 to 1.0 with increment 0.01. For each noise vector, we evaluate the model using the test set with the vector. We compute the distance between the generated location and the location of the closest member for each test sample. Then, we determine the mean and standard deviation of all these distances.
  To understand the influences on the generated locations, we first group all the mean distances of the test set relative to the noise vectors into two categories using k-means clustering with $k = 2$. The clustering result is presented by Fig. 9. Although the centroids of these two clusters are close, we may still have some findings where a group of noise vectors may generate locations closer to the member, and the other group of noise vectors may generate farther locations. Second, we cluster two groups by means and standard deviations. Fig. 10 shows that two groups have similar average distances, but one group has a higher deviation. Thus, we may use the noise vector in the group with a higher deviation to obtain more diverse locations.
- Random Noise Vector: In this experiment, we maintain the same settings as above. However, in this case, the noise vector for each dimension is assigned a random value from one of five distinct ranges: $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, and $[0.8, 1]$. Fig. 11 shows the results. From the results, we note that there are only slight differences in the distances between the generated locations and the location of the closest group member. We may achieve a closer distance by using values within the first and third ranges.

Here, we explore two potential methods for managing distancing, providing users with the flexibility to adjust the distancing according to their application's specific requirements. For instance, in congested environments, system administrators can generate closer distancing using a random noise vector with a range of either $[0, 0.2)$ or $[0.4, 0.6)$. Conversely, for less congested settings, an opposite setup may be employed. Moreover, to maintain consistency in results across generations, system managers may opt for a static noise vector with a value in a group exhibiting minimal variation.

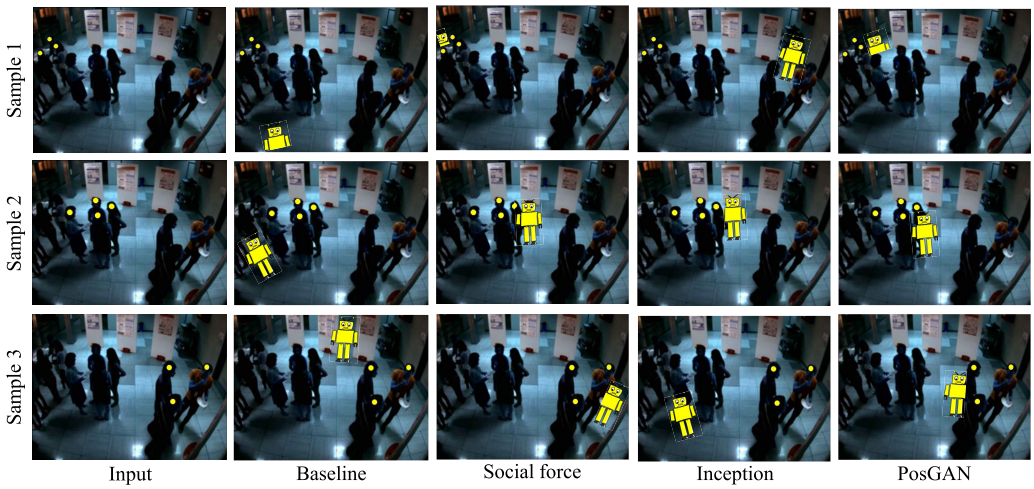


FIGURE 12. Examples of prediction results on images of the real scene.

B. SYSTEM DEPLOYMENT

Our system offers applications for various social settings such as retail, public spaces, and collaborative workspaces, where social robots need to fit into human interactions while maintaining social distance. For example, social robots could provide non-intrusive information and assistance in places such as airports or museums, enhancing user experience. Similarly, in education, our approach could improve interactions between robots and students in group learning, creating a comfortable, effective learning atmosphere.

When considering deploying this system, there are certain prerequisites to ensure the effective functioning of social robots in dynamic environments. We propose three main requirements. Firstly, a camera must be installed, either fixed within the environment or embedded in the robot, to recognize the positions of each user. Secondly, the system requires adequate computational power to handle recognition and prediction tasks. Our current inference model possesses 256, 638, 979 parameters and demands 12.5 GFLOPs (giga floating point operations). In theory, this model is compatible with Nvidia Jetson Nano, which offers 472 GFLOPS (giga floating point operations per second) [64], Nvidia Jetson TX2 NX capable of 1.33 TFLOPS (tera floating point operations per second) [64], or more powerful edge devices. Lastly, the robot must incorporate an actuating system, a navigation method, and a communication interface to facilitate interaction with users and navigation within the real environment.

Addressing privacy concerns, primarily related to collecting and processing personal data, is crucial. In our current design, the robot can process and make decisions on the edge device locally to avoid disclosing user information to the cloud. This approach prioritizes user privacy and security. While we ensure that sensitive user information, such as facial data, is not utilized, we focus instead on the user's head position for predictions. However, the use of cameras poses potential privacy risks if unauthorized parties access them for unrelated purposes. Adherence to existing privacy laws is essential during the deployment of the system. Measures such as data anonymization, explicit user consent, and transparency in data usage policies should be integral parts of the system's deployment. These are significant considerations for deployment, despite falling outside of our current system design.

To explore the societal impacts of deploying this system, we refer to several related systems. The initial concern is that an unsatisfactory location predicted by our model could lead to discomfort among individuals. Furthermore, the introduction of a social robot might also result in discomfort. These issues may include user stress [65], increased dissatisfaction [66], [67], reduced authenticity of the experience [68], and diminished trust [69]. To minimize societal impacts and improve acceptance, the robot could incorporate a mechanism to identify undesirable human actions to prevent further proximity. For example, the robot might stop its movement if a group member executes a stop gesture or indicates a desire to leave.

Furthermore, as an ethical point, our system, which is based on learning technology, could unintentionally reflect biases in human behavior. For example, social distancing norms might differ between male-male and female-male pairs. Currently, our focus is on the system's design, leaving exploring and mitigating potential biases as an area for future research.

VI. CONCLUSION

We have proposed PosGAN, which uses generative AI to suggest a socially acceptable standing location for a robot to join a conversation group. We investigated the possibility of using the image generation technique to leverage the recognition power of deep learning in visual information. By considering human behaviors in the group conversation,

we trained and tested our model using the dataset collected in real-world settings. The experiment results showed that our method outperforms other approaches by the metric of the distancing to the members in the conversation. Our approach may generate a location with moderate distancing that is not too close or too far from the members, as a human did. Therefore, our model will allow the social robot to determine a socially acceptable location, which is not well addressed in current research. In future work, the research will focus on incorporating SLAM algorithms with PosGAN to enhance the robot's spatial awareness in human groups.

REFERENCES

[1] A. Guerrieri, E. Braccili, F. Sgrò, and G. N. Meldolesi, "Gender identification in a two-level hierarchical speech emotion recognition system for an Italian social robot," *Sensors*, vol. 22, no. 5, p. 1714, Feb. 2022.

[2] L. M. Galván, E. Fernández-Rodicio, J. S. Salcedo, Á. Castro-González, and M. A. Salichs, "Using deep learning for implementing paraphrasing in a social robot," in *Proc. Int. Symp. Ambient Intell.*, 2022, pp. 1–19.

[3] D. N. Thang, L. A. Nguyen, P. T. Dung, T. D. Khoa, N. H. Son, N. T. Hiep, P. Van Nguyen, V. D. Truong, D. H. Toan, N. M. Hung, T.-D. Ngo, and X.-T. Truong, "Deep learning-based multiple objects detection and tracking system for socially aware mobile robot navigation framework," in *Proc. 5th NAFOSTED Conf. Inf. Comput. Sci. (NICS)*, Nov. 2018, pp. 436–441.

[4] C. Yu and H. Pei, "Face recognition framework based on effective computing and adversarial neural network and its implementation in machine vision for social robots," *Comput. Electr. Eng.*, vol. 92, Jun. 2021, Art. no. 107128.

[5] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human–robot interaction," *Inf. Sci.*, vol. 428, pp. 49–61, Feb. 2018.

[6] K. Zhu and T. Zhang, "Deep reinforcement learning based mobile robot navigation: A review," *Tsinghua Sci. Technol.*, vol. 26, no. 5, pp. 674–691, Oct. 2021.

[7] P. Quan, Y. Lou, H. Lin, Z. Liang, and S. Di, "Research on fast identification and location of contour features of electric vehicle charging port in complex scenes," *IEEE Access*, vol. 10, pp. 26702–26714, 2022.

[8] B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K. R. Choo, and M. Jamshidi, "Toward artificial emotional intelligence for cooperative social human-machine interaction," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 1, pp. 234–246, Feb. 2020.

[9] A. Alam, "Social robots in education for long-term human–robot interaction : Socially supportive behaviour of robotic tutor for creating robo-tangible learning environment in a guided discovery learning interaction," *ECS Trans.*, vol. 107, no. 1, pp. 12389–12403, Apr. 2022.

[10] S. L. Lopes, A. I. Ferreira, and R. Prada, "The use of robots in the workplace: Conclusions from a health promoting intervention using social robots," *Int. J. Social Robot.*, vol. 15, no. 6, pp. 893–905, Jun. 2023.

[11] M. M. Neumann, D. L. Neumann, and L.-C. Koch, "Young children's interactions with a social robot during a drawing task," *Eur. Early Childhood Educ. Res. J.*, vol. 31, no. 3, pp. 421–436, May 2023.

[12] B. Ding, Y. Li, S. Miah, and W. Liu, "Customer acceptance of frontline social robots–human–robot interaction as boundary condition," *Technological Forecasting Social Change*, vol. 199, Feb. 2024, Art. no. 123035.

[13] A. Andriella, C. Torras, C. Abdelnour, and G. Alenyà, "Introducing CARESSER: A framework for in situ learning robot social assistance from expert knowledge and demonstrations," *User Model. User-Adapted Interact.*, vol. 33, no. 2, pp. 441–496, Apr. 2023.

[14] M. Kraus, N. Wagner, N. Untereiner, and W. Minker, "Including social expectations for trustworthy proactive human–robot dialogue," in *Proc. 30th ACM Conf. User Model., Adaptation Personalization*, Jul. 2022, pp. 1–10.

[15] E. Shmueli, V. K. Singh, B. Lepri, and A. Pentland, "Sensing, understanding, and shaping social behavior," *IEEE Trans. Computat. Social Syst.*, vol. 1, no. 1, pp. 22–34, Mar. 2014.

[16] H. B. Barua, P. Pramanick, C. Sarkar, and T. H. Mg, "Let me join you! Real-time F-formation recognition by a socially aware robot," in *Proc. 29th IEEE Int. Conf. Robot. Hum. Interact. Commun.*, Aug. 2020, pp. 371–377.

[17] X.-T. Truong and T.-D. Ngo, "'To approach humans?': A unified framework for approaching pose prediction and socially aware robot navigation," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 3, pp. 557–572, Sep. 2018.

[18] S. K. Pathi, A. Kristofferson, A. Kiselev, and A. Loutfi, "Estimating optimal placement for a robot in social group interaction," in *Proc. 28th IEEE Int. Conf. Robot. Hum. Interact. Commun.*, Oct. 2019, pp. 1–8.

[19] Y. Gao, F. Yang, M. Frisk, D. Hemandez, C. Peters, and G. Castellano, "Learning socially appropriate robot approaching behavior toward groups using deep reinforcement learning," in *Proc. 28th IEEE Int. Conf. Robot. Hum. Interact. Commun.*, Oct. 2019, pp. 1–8.

[20] S. K. Pathi, A. Kristoffersson, A. Kiselev, and A. Loutfi, "F-formations for social interaction in simulation using virtual agents and mobile robotic telepresence systems," *Multimodal Technol. Interact.*, vol. 3, no. 4, p. 69, Oct. 2019.

[21] C. Pedica and H. Vilhjalmsson, "Social perception and steering for online avatars," in *Proc. Int. Conf. Intell. Virtual Agents*, 2008, pp. 104–116.

[22] F. Yang and C. Peters, "AppGAN: Generative adversarial networks for generating robot approach behaviors into small groups of people," in *Proc. 28th IEEE Int. Conf. Robot. Hum. Interact. Commun.*, Oct. 2019, pp. 1–8.

[23] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of ChatGPT: The history, status quo and potential future development," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1122–1136, May 2023.

[24] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 1–20, Sep. 2023.

[25] I. Goodfellow, P. Pouget-Abadie, P. Mirza, P. Xu, P. Warde-Farley, P. Ozair, P. Courville, and P. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[26] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "SALSA: A novel dataset for multimodal group behavior analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1707–1720, Aug. 2016.

[27] E. Ricci, J. Varadarajan, R. Subramanian, S. R. Bulo, N. Ahuja, and O. Lanz, "Uncovering interactions and interactors: Joint estimation of head, body orientation and F-formations from surveillance videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4660–4668.

[28] C. Shi, S. Satake, T. Kanda, and H. Ishiguro, "How would store managers employ social robots?" in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot. Interact. (HRI)*, Mar. 2016, pp. 519–520.

[29] A. Chella, M. Liotta, and I. Macaluso, "CiceRobot: A cognitive robot for interactive museum tours," *Ind. Robot, Int. J.*, vol. 34, no. 6, pp. 503–511, Oct. 2007.

[30] A. Kanda, M. Arai, R. Suzuki, Y. Kobayashi, and Y. Kuno, "Recognizing groups of visitors for a robot museum guide tour," in *Proc. 7th Int. Conf. Hum. Syst. Interact.*, Jun. 2014, pp. 123–128.

[31] S. Pekkarinen, L. Hennala, O. Tuisku, C. Gustafsson, R.-M. Johansson-Pajala, K. Thommes, J. A. Hoppe, and H. Melkas, "Embedding care robots into society and practice: Socio-technical considerations," *Futures*, vol. 122, Sep. 2020, Art. no. 102593.

[32] A. Sharkey and N. Sharkey, "Granny and the robots: Ethical issues in robot care for the elderly," *Ethics Inf. Technol.*, vol. 14, no. 1, pp. 27–40, Mar. 2012.

[33] T. Alhmiedat, A. M. Marei, W. Messoudi, S. Albelwi, A. Bushnag, Z. Bassfar, F. Alnajjar, and A. O. Elfaki, "A SLAM-based localization and navigation system for social robots: The pepper robot case," *Machines*, vol. 11, no. 2, p. 158, Jan. 2023.

[34] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "DS-SLAM: A semantic visual SLAM towards dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1168–1174.

[35] H. Kuzuoka, Y. Suzuki, J. Yamashita, and K. Yamazaki, "Reconfiguring spatial formation arrangement by robot body orientation," in *Proc. 5th ACM/IEEE Int. Conf. Hum.-Robot. Interact. (HRI)*, Mar. 2010, pp. 285–292.

[36] M. A. Yousuf, Y. Kobayashi, Y. Kuno, A. Yamazaki, and K. Yamazaki, "Development of a mobile museum guide robot that can configure spatial formation with visitors," in *Proc. Int. Conf. Intell. Comput.*, 2012, pp. 1–16.

[37] S.-A. Yang, E. Gamborino, C.-T. Yang, and L.-C. Fu, "A study on the social acceptance of a robot in a multi-human interaction using an F-formation based motion model," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 2766–2771.

[38] P. Althaus, H. Ishiguro, T. Kanda, T. Miyashita, and H. I. Christensen, "Navigation for human–robot interaction tasks," in *Proc. IEEE Int. Conf. Robot. Autom., Proceedings. ICRA.*, Aug. 2004, pp. 1–14.

[39] M. Svenstrup, T. Bak, and H. J. Andersen, "Trajectory planning for robots in dynamic human environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 4293–4298.

[40] C. Rösmann, M. Oeljeklaus, F. Hoffmann, and T. Bertram, "Online trajectory prediction and planning for social robot navigation," in *Proc. IEEE Int. Conf. Adv. Intell. Mechatronics (AIM)*, Jul. 2017, pp. 1255–1260.

[41] S. Mishra, P. K. Rajendran, and D. Har, "Socially acceptable route planning and trajectory behavior analysis of personal mobility device for mobility management with improved sensing," in *Proc. Int. Conf. Robot. Intell. Technol. Appl.*, 2021, pp. 1–17.

[42] M. Shiomi, F. Zanlungo, K. Hayashi, and T. Kanda, "Towards a socially acceptable collision avoidance for a mobile robot navigating among pedestrians using a pedestrian model," *Int. J. Social Robot.*, vol. 6, no. 3, pp. 443–455, Aug. 2014.

[43] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[44] S. P. H. Boroujeni and A. Razi, "IC-GAN: An improved conditional generative adversarial network for RGB-to-IR image translation with applications to forest fire monitoring," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121962.

[45] G. Tong, W. Shao, and Y. Li, "ReverseGAN: An intelligent reverse generative adversarial networks system for complex image captioning generation," *Displays*, vol. 82, Apr. 2024, Art. no. 102653.

[46] B. Cao, H. Zhang, N. Wang, X. Gao, and D. Shen, "Auto-GAN: Self-supervised collaborative learning for medical image synthesis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10486–10493.

[47] F. Haghighi, M. R. Hosseinzadeh Taher, M. B. Gotway, and J. Liang, "Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial?" *Med. Image Anal.*, vol. 94, May 2024, Art. no. 103086.

[48] Z. Anastasakis, D. Mallis, M. Diomataris, G. Alexandridis, S. Kollias, and V. Pitsikalis, "Self-Supervised learning for visual relationship detection through masked bounding box reconstruction," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 1206–1215.

[49] B.-J. Krings and N. Weinberger, "Assistant without master? Some conceptual implications of assistive robotics in health care," *Technologies*, vol. 6, no. 1, p. 13, Jan. 2018.

[50] T. Iachini, Y. Coello, F. Frassinetti, and G. Ruggiero, "Body space in social interactions: A comparison of reaching and comfort distance in immersive virtual reality," *PLoS One*, vol. 9, no. 11, Nov. 2014, Art. no. e111511.

[51] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Towards safe and trustworthy social robots: Ethical challenges and practical issues," in *Proc. Int. Conf. Social Robot. (ICSR)*, 2015, pp. 1–15.

[52] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human–robot interaction," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 53, no. 5, pp. 517–527, Oct. 2011.

[53] M. Filipenko and I. Afanasyev, "Comparison of various SLAM systems for mobile robot in an indoor environment," in *Proc. Int. Conf. Intell. Syst.*, Sep. 2018, pp. 400–407.

[54] K. Ohno, T. Nomura, and S. Tadokoro, "Real-time robot trajectory estimation and 3D map construction using 3D camera," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2006, pp. 5279–5285.

[55] Z. Chen, J. Wang, C. Jia, and X. Ye, "Pathological image super-resolution using mix-attention generative adversarial network," *Int. J. Mach. Learn. Cybern.*, vol. 15, no. 1, pp. 149–159, Jan. 2024.

[56] Y. Kumar, A. Ilin, H. Salo, S. Kulathinal, M. K. Leinonen, and P. Marttinen, "Self-supervised forecasting in electronic health records with attention-free models," *IEEE Trans. Artif. Intell.*, pp. 1–17, Jan. 2024.

[57] Y. Zheng, M. Jin, Y. Liu, L. Chi, K. T. Phan, and Y. P. Chen, "Generative and contrastive self-supervised learning for graph anomaly detection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12220–12233, Dec. 2023.

[58] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[59] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[61] C. Xiao, P. Zhong, and C. Zheng, "BourGAN: Generative networks with metric embeddings," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1–18.

[62] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4431–4440.

[63] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "PD-GAN: Probabilistic diverse GAN for image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9367–9376.

[64] *Technical Specifications of Nvidia Jetson Modules*. Accessed: Jun. 18, 2024. [Online]. Available: https://developer.nvidia.com/embedded/jetson-modules

[65] M. Smakman, J. Berket, and E. A. Konijn, "The impact of social robots in education: Moral considerations of Dutch educational policymakers," in *Proc. 29th IEEE Int. Conf. Robot. Human Interact. Commun.*, Aug. 2020, pp. 647–652.

[66] A. Amelia, C. Mathies, and P. G. Patterson, "Customer acceptance of frontline service robots in retail banking: A qualitative approach," *J. Service Manage.*, vol. 33, no. 2, pp. 321–341, Feb. 2022.

[67] Y. Li and C. Wang, "Effect of customer's perception on service robot acceptance," *Int. J. Consum. Stud.*, vol. 46, no. 4, pp. 1241–1261, Jul. 2022.

[68] A. Milman, A. Tasci, and T. Zhang, "Perceived robotic server qualities and functions explaining customer loyalty in the theme park context," *Int. J. Contemp. Hospitality Manage.*, vol. 32, no. 12, pp. 3895–3923, Nov. 2020.

[69] D. L. Johanson, H. S. Ahn, and E. Broadbent, "Improving interactions with healthcare robots: A review of communication behaviours in social and healthcare contexts," *Int. J. Social Robot.*, vol. 13, no. 8, pp. 1835–1850, Dec. 2021.

SEVENDI ELDRIGE RIFKI POLUAN received the B.Sc. degree in computer science and engineering (CSE) from Universitas Klabat, Airmadidi, Manado, Indonesia, in 2015, and the M.S. degree in CSE from Yuan Ze University, Taoyuan, Taiwan, in 2019, where he is currently pursuing the Ph.D. degree in CSE with the Mobile Sensing and Systems Laboratory. His research interests include machine learning, deep learning, computer vision, generative AI, and information fusion, with a focus on their application in various domains.

YAN-ANN CHEN (Member, IEEE) received the Ph.D. degree in computer science and engineering from National Chiao Tung University, Taiwan, in 2016. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Yuan Ze University, Taiwan. He has published about 20 research papers in international journals and conferences. His research interests include edge artificial intelligence, the Internet of Things, mobile/wearable sensing, and cyber-physical systems.