A Study on Learning Social Robot Navigation with Multimodal Perception

Bhabaranjan Panigrahi, Amir Hossain Raj, Mohammad Nazeri, and Xuesu Xiao

***Abstract*— Autonomous mobile robots need to perceive the environments with their onboard sensors (e.g., LiDARs and RGB cameras) and then make appropriate navigation decisions. In order to navigate human-inhabited public spaces, such a navigation task becomes more than only obstacle avoidance, but also requires considering surrounding humans and their intentions to somewhat change the navigation behavior in response to the underlying social norms, i.e., being socially compliant. Machine learning methods are shown to be effective in capturing those complex and subtle social interactions in a data-driven manner, without explicitly hand-crafting simplified models or cost functions. Considering multiple available sensor modalities and the efficiency of learning methods, this paper presents a comprehensive study on learning social robot navigation with multimodal perception using a large-scale real-world dataset. The study investigates social robot navigation decision making on both the global and local planning levels and contrasts unimodal and multimodal learning against a set of classical navigation approaches in different social scenarios, while also analyzing the training and generalizability performance from the learning perspective. We also conduct a human study on how learning with multimodal perception affects the perceived social compliance. The results show that multimodal learning has a clear advantage over unimodal learning in both dataset and human studies. We open-source our code for the community's future use to study multimodal perception for learning social robot navigation.**[1]

I. INTRODUCTION

Thanks to decades of robotics research [1], [2], autonomous mobile robots can navigate from one point to another in a collision-free manner in many real-world environments, e.g., factories and warehouses. Using onboard sensors, e.g., LiDARs and RGB cameras, those robots can perceive the environments, divide their workspaces into obstacles and free spaces, and then make navigation decisions to avoid obstacles and move towards their goal [3]–[6].

However, when deploying mobile robots in human-inhabited public spaces, the navigation task becomes more complex [7]–[9]: While avoiding any obstacle on the way to the goal is still required, they also need to consider other humans sharing the same environments and adjust their decision-making process to produce new navigation behaviors that respond to the underlying, usually unwritten, social norms.

One avenue to achieve such social compliance is machine learning [10]. Learning approaches allow those complex and subtle human-robot interactions during social navigation to be captured in a data-driven manner and alleviate roboticists from manually designing simplified models [11], [12], crafting cost functions [13], [14], and fine-tuning system parameters [15]–[19]. The development of machine learning infrastructure, e.g., onboard computation devices and an extensive corpus of perception data being generated from robots, also accelerates the adoption of learning methods for social robot navigation. Most current robots have multiple sensors onboard, with LiDARs and RGB cameras as the most common sensing modalities, and are therefore able to perceive complex social interactions from different sources (Fig. 1). While LiDARs have been the main perception modality for mobile robots for decades, recent research has shifted towards visual navigation with RGB input alone, thanks to its cheap cost and wide availability. Intuitively speaking, LiDARs provide high-resolution and high-accuracy geometric information about the environments, while cameras stream in RGB images which contain rich semantics. Both geometric and semantic information play a role in the decision making process of social robot navigation: Geometric structures like obstacles and humans need to be avoided, while semantics including navigation terrain, human gaze [20], [21], gesture, clothing, and body language can shed light on the navigation contexts and other humans' intentions to inform robot navigation decisions.

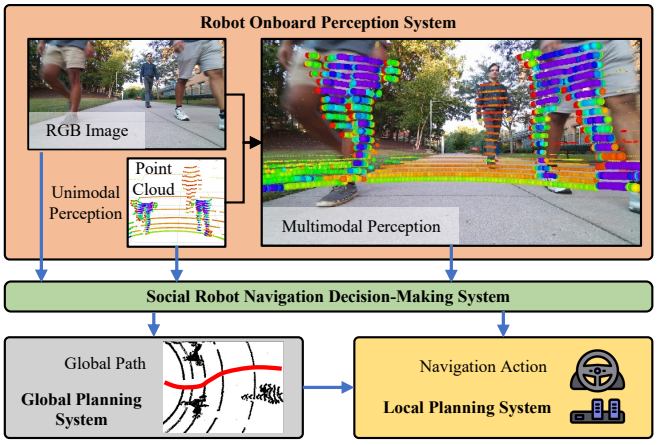

Fig. 1: Social Robot Navigation Decision Making on the Global and Local Level with Multimodal and Unimodal (RGB Image and Point Cloud) Perception Input.

Considering the rich and potentially complementary information provided by multiple available sensor modalities onboard mobile robots and the efficiency of learning methods in enabling emergent social robot navigation behaviors, this paper presents a comprehensive study on using multimodal

All authors are with the Department of Computer Science, George Mason University {bpanigr, araj20, mnazerir, xiao}@gmu.edu

[1]GitHub: https://github.com/RobotiXX/multimodal-fusion-network/

perception of LiDAR and RGB camera inputs, two most common perception modalities of autonomous mobile robots, to learn the robot decision making process during social robot navigation. The study is conducted on a large-scale real-world Socially Compliant Navigation Dataset (SCAND) [22] collected in a variety of natural crowded public spaces on a university campus. From the social robot navigation perspective, we study the decision-making capability of multimodal and unimodal learning on both global and local planning in different social scenarios (e.g., Against Traffic, With Traffic, and Street Crossing); From the machine learning perspective, we study the training and generalizability performance of multimodal and unimodal learning in terms of training time, loss value, generalization accuracy, etc. We also conduct a human study and reveal how social compliance achieved by different sensor modalities can be perceived by humans interacting with the robot. The results show that multimodal learning is more reliable and robust than using unimodal networks in both dataset and human studies. Despite existing work on multimodal perception for, e.g., autonomous driving [23], [24] and trajectory prediction [25], [26], to our best knowledge, our work is the first comprehensive *study* on using multimodal perception to *learn social robot navigation* and the first to empirically show the advantage of multimodal over unimodal learning on a large real-world social robot navigation dataset and in a real-world human study.

II. RELATED WORK

We review related work in social robot navigation, machine learning for navigation, and multimodal learning.

A. Social Robot Navigation

While collision-free navigation has been investigated by the robotics community for decades [1]–[6], roboticists have also built mobile robots that navigate around humans since the early museum tour-guide robots RHINO [27] and MINERVA [28]. Going beyond simply treating humans as dynamic, non-reactive obstacles [4], researchers have also modeled the uncertainty of human movements [29]–[33] or prescribed social norms for navigating agents [34]–[36], and then devised navigation planners that can take such uncertainty into account or abide such selected rules. These physics-based models [37]–[40] consider humans' behavior features, such as proxemics [41]–[44], intentions [45], [46], and social formations and spaces [12], [20], [31], [47]. However, prescribing a simple model is usually not sufficient to capture complex human behaviors in the wild. For example, pedestrians move differently during rush hours or on weekends, within formal or informal contexts. Furthermore, such a plethora of factors to be considered during social robot navigation all have to be processed from raw perceptual data from onboard sensors, e.g., LiDARs and RGB cameras, and set challenges for onboard perception algorithms, e.g., human tracking, motion prediction, and intention detection. Along with the recent success in machine learning, both these challenges led to the recent adoption of data-driven approaches for social robot navigation [10].

B. Machine Learning for Navigation

As a potential solution to the aforementioned challenges, machine learning approaches have been leveraged to implicitly encode the complexities and subtleties of human social behaviors in a data-driven manner [10] and also address other challenges in navigation, e.g., off-road navigation [48]–[53]. These data-driven approaches include learning representations or costmaps [13], [14], [54]–[56], parameterizations of navigation planners [15]–[19], [57], local planners [58]–[62], or end-to-end navigation policies that map directly from raw or pre-processed perceptions of the humans in the scene to motor commands that drive the robot [63]–[65]. From the perspective of machine learning methods, reinforcement learning [19], [61], [62], [66], [66] and imitation learning [13], [16]–[18], [67]–[69] depend on training data from mostly simulated trial-and-error experiences and either human or artificial expert demonstrations respectively. Considering the difficulty in producing high-fidelity perceptual data and natural human-robot interactions in simulation, this study adopts an imitation learning setup, in particular, Behavior Cloning (BC) [68], [69], with a large-scale social robot navigation demonstration dataset.

C. Multimodal Learning

Recent research has shown that combining data from different modalities in a multimodal learning framework can lead to promising results in solving downstream tasks [70]. For autonomous mobile robot navigation, researchers have tried sensor fusion by combining RGB cameras, LiDARs, and robot odometry with a multimodal graph neural network to navigate unstructured terrain including bushes, small trees, and grass regions of different heights and densities [71]. Furthermore, they have demonstrated the robustness of the network towards partial occlusion and unreliable sensor information in challenging outdoor environments. Other researchers have also combined laser, RGB images, point cloud, and distance map to learn navigation in time-sensitive scenarios such as disaster response or search and rescue, which include constrained narrow passages, pathways with debris, and irregular navigation scenarios [72]. Additionally, they have demonstrated that multimodal networks outperformed models that only utilized RGB images and distance maps. Multimodal perception has been shown to be valuable in addressing different challenges during real-world navigation tasks, but to the best of our knowledge, investigation into how multimodal perception can affect decision making during social robot navigation is still very limited, which is the focus of this study. Notice that we are interested in learning social robot navigation with multimodal *perception* as input [70], rather than learning models with multimodal *distribution*, which has a relatively richer literature [73]–[75].

D. Socially Compliant Robot Navigation Dataset (SCAND)

Our study is based on an open-source, large-scale, real-world social robot navigation dataset, SCAND [22], of 8.7 hours, 138 trajectories, 40 kilometers of socially compliant, human teleoperated driving demonstrations that comprise

multimodal data streams including 3D LiDAR, visual and inertial information, robot odometry, and joystick commands, collected on two morphologically different mobile robots—a Boston Dynamics Spot and a Clearpath Jackal—by four different human demonstrators in both indoor and outdoor environments. Due to its rich social interactions and multimodal perception-to-action navigation decisions, SCAND is suitable for studying social robot navigation learning with multimodal perception. Specifically, we study the effect of both point cloud data from a 3D LiDAR and RGB images from a camera, the most commonly available perception modalities onboard mobile robots, considering the geometric and semantic information provided by the point cloud data and RGB images can complement each other to assist decision making during social robot navigation in human-inhabited public spaces.

III. MULTIMODAL LEARNING FOR SOCIAL ROBOT NAVIGATION

We adopt an imitation learning approach, i.e., BC, to learn socially compliant navigation decisions using multimodal perception from SCAND. Similar to classical navigation systems with a global and a local planning system, we design our multimodal learning framework so that it will produce both global and local plans and study how multimodal and unimodal learning can imitate the navigation decisions made by the human demonstrator on both global and local levels.

A. Problem Formulation

Specifically, at each time step t of each trial in SCAND, the robot receives onboard perceptual input, including a sequence of 3D LiDAR point cloud data L and RGB images I, and a goal G it aims to reach, which is taken as a waypoint 2.5m away from the robot on the future robot odometry. We denote all these inputs necessary to inform the decision-making process during social robot navigation as a navigation input: $\mathcal{I}_t^D = \{L_k^D, I_k^D, G_t^D\}_{k=t-N+1}^t$, where N denotes the history length included in the navigation input at t and D denotes that the data is from the SCAND demonstrations.

Facing a social navigation input $\mathcal{I}_t^D$, the SCAND demonstrator shows the desired, socially compliant navigation decision $\mathcal{D}_t$ on both global and local levels: P_t is the demonstrated global plan, recorded as the human-driven future robot odometry starting from time t, and takes the form of a sequence of 2D waypoints $P_t^D = \{(x_i^D, y_i^D)\}_{i=t}^{t+M-1}$; A_t is the demonstrated local plan represented as a sequence of joystick action commands $A_t^D = \{(v_i^D, \omega_i^D)\}_{i=t}^{t+K-1}$, where v and ω is the linear and angular velocity respectively. M and K denote the length of the navigation decision on the global and local plan level respectively. The demonstrated navigation decision is therefore defined as $\mathcal{D}_t^D = \{P_t^D, A_t^D\}$.

Producing the navigation decision $\mathcal{D}_t^D$ based on $\mathcal{I}_t^D$ as input, a navigation system is defined as a combination of two functions, $\mathcal{F}^g(\cdot)$ and $\mathcal{F}^l(\cdot)$, responsible of generating the global plan P_t^D and local plan (action) A_t^D:

$$P_t^D = \mathcal{F}^g(\mathcal{I}_t^D),$$
$$A_t^D = \mathcal{F}^l(\mathcal{I}_t^D, P_t^D).$$

In a data-driven manner, we instantiate both global and local planners by learning $\mathcal{F}_\theta^g(\cdot)$ and $\mathcal{F}_\phi^l(\cdot)$ as deep neural networks with learnable parameters θ and ϕ respectively. In particular, we aim to learn the parameters to minimize a BC loss:

$$\theta^*, \phi^* = \underset{\theta,\phi}{\operatorname{argmin}} \sum_{P_t^D, A_t^D, \mathcal{I}_t^D \in \text{SCAND}} \left[||P_t^D - \mathcal{F}_\theta^g(\mathcal{I}_t^D)|| + \lambda ||A_t^D - \mathcal{F}_\phi^l(\mathcal{I}_t^D, \mathcal{F}_\theta^g(\mathcal{I}_t^D))|| \right], \quad (1)$$

where the first term is the difference between demonstrated and learned global plan, while the second term is for the local plan, with λ as a weight between them.

In this study, we are interested in studying the effect of including different perception modalities in $\mathcal{I}_t$ on making socially compliant navigation decisions P_t and A_t. We study three scenarios, i.e., multimodal perception $\mathcal{I}_t^{\text{MM}} = \{L_k, I_k, G_t\}_{k=t-N+1}^t$, unimodal LiDAR (point cloud) perception $\mathcal{I}_t^{\text{LiDAR}} = \{L_k, G_t\}_{k=t-N+1}^t$, and unimodal vision (RGB image) perception $\mathcal{I}_t^{\text{Vision}} = \{I_k, G_t\}_{k=t-N+1}^t$. For simplicity and consistency, we keep $N = 1$ for all three cases in this study and leave an investigation into different history lengths as future work.

B. Unimodal Perception

1) Point Cloud Modality: We take points that are within the range of 8 meters in front, 3 meters on either side and within 2.5 meters of height from the robot as perceived by the 3D LiDAR. All points are placed into their respective voxel inside a 3D voxel grid with $5 \times 5 \times 5$cm voxels, resulting in a $160 \times 120 \times 50$ voxel representation for L_k. We use a 3D Convolution Neural Network (CNN) [76] to process the voxel representation to extract meaningful information for our downstream social robot navigation task. The point cloud encoder is shown as the green trapezoid in the red box at the bottom of Fig. 2.

2) RGB Modality: For RGB images, we take a $224 \times 224 \times 3$ image from the camera as input. We use ResNet-18 [77] to extract features for our social robot navigation task. The image encoder is shown as the green trapezoid in the yellow box at the top of Fig. 2.

Both RGB and point cloud inputs have their own unimodal decision making modules, shown in the upper yellow and lower red box in Fig. 2 respectively. For a fair comparison, we enforce the same architecture, the only difference is the different input modalities. To be specific, we concatenate the embeddings from the corresponding input encoders with the local goal (2.5m away), and feed them into a Recurrent Neural Network (RNN) to capture history information (blue ellipsoids in Fig. 2). Then we use a Multi-Layer Perceptron (MLP) (yellow boxes in Fig. 2) to produce global plan in the form of a sequence of 2D waypoints (red dots in Fig. 2), which are further fed into another MLP. Concatenating the MLP output with the RNN output, a transformer, and another MLP at the end produces local plan, i.e., actions of linear and angular velocities.

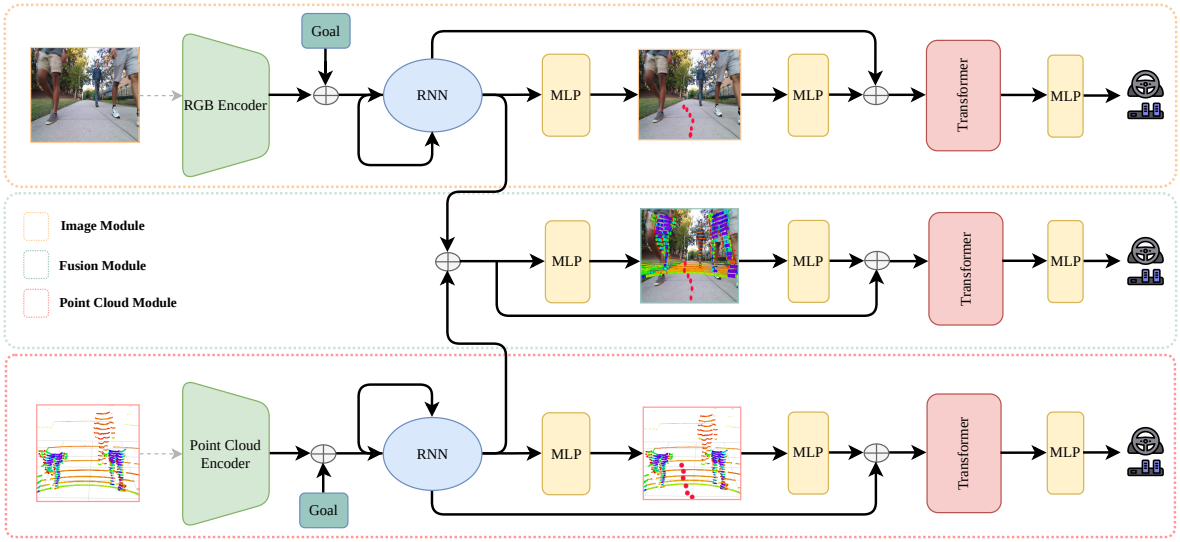

Fig. 2: Image Module, Fusion (Multimodal) Module, and Point Cloud Module Architecture for Social Robot Navigation.

C. Multimodal Fusion

For multimodal fusion, the outputs of the RNNs from the point cloud and image modules are concatenated and passed through the fusion process, shown in Fig. 2 middle. Similar to the unimodal modules, our feature fusion also happens at two different places in our multimodal network. Each fusion caters to different downstream tasks, i.e., producing both global and local plans.

D. Navigation Decisions and Loss Functions

The global navigation decisions are instantiated as a sequence of five future waypoints ahead of the robot, i.e., $P_t^D = \{(x_i^D, y_i^D)\}_{i=t}^{t+4}$ $(M = 5)$, each of which is 0.5m apart taken from the future robot odometry. The local navigation decisions take the form of the current linear and angular velocity commands, i.e., $A_t^D = \{(v_t^D, \omega_t^D)\}$ $(K = 1)$.

For the first and second loss terms in Eqn. 1, we use $L2$-norm of the five future waypoints and $L1$-norm of the current angular and linear velocity. We set $\lambda = 1$.

E. Design Choices

Notice that all aforementioned design choices with respect to neural network hyper-parameters and architecture are made after extensive trial-and-error and careful fine-tuning to ensure the different modalities can achieve the best learning performance for a fair comparison. All detailed hyper-parameters and design choices can be found in our open-source implementation for the future use of the community.

We have experimented with PointNet [78] and PointNet++ [79] for the point cloud encoder, which does not perform well on SCAND social navigation scenarios: PointNet encodes individual point and relies on the global pooling layers to extract effective features. However, encoding points for highly diverse indoor and outdoor SCAND scenarios is not effective. Unlike closed, and small-scale indoor objects, point clouds collected during real-world robot navigation contain significantly more variation in terms of the number and distribution of points. Our further investigation into the point cloud encoder reveals that converting them to a voxelized grid and then processing them through a 3D CNN network results in a significant performance gain.

We also try to learn local planner using simple MLP, but it fails to capture the variations in SCAND. For instance, for the same global path, there can be different velocities: If humans are nearby the linear velocity will be slower, in contrast to a scenario where they are far apart. Transformer can achieve significant performance gain because of the attention modules which can decide which features it should attend to in order to capture these variations.

IV. SCAND STUDY RESULTS

We first present our study results on all the social scenarios in SCAND before presenting our human study results. We divide the SCAND trials into 18 for training and 8 for testing. We analyze the learning results on the test data from both the machine learning and social robot navigation perspectives. The training loss curves for the global planner in terms of L1 loss on the eight SCAND ROSBAGS are shown in Fig. 3, while the local planner loss in Fig. 4. We also plot the performance of a variety of classical social robot navigation planners using the same loss function between their output and the SCAND demonstration to compare against end-to-end learned policies.

A. Multimodal Learning Performance

The results of the eight test SCAND ROSBAGS are ordered roughly according to increasing performance discrepancy among different modalities in Fig. 3, which can also be treated as an approximate representation of the "difficulty" level in social robot navigation decision making. For example, the loss values of most modalities converge faster and to a lower point in the earlier "easy" trials (upper left), compared to the later "difficult" ones (lower right).

It is clear that in terms of test loss for global planning, learning with multimodal perception significantly outperforms both unimodal perception modalities. The multimodal test loss shown by the green curves drops faster, converges

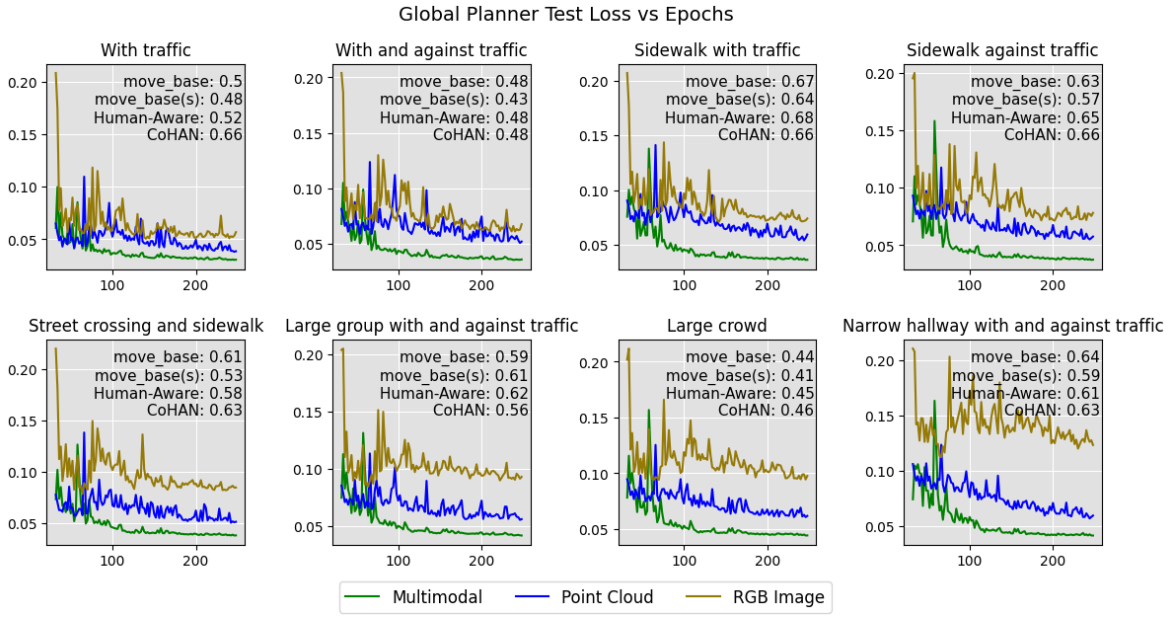

Fig. 3: Test Loss on Eight SCAND ROSBAGS with Multimodal, Point Cloud, and RGB Image Input (Averaged Over Three Training Runs with Negligible Variances Invisible in the Figures). Losses of four Classical Approaches are for Comparison.

at a smaller epoch number, and reduces to a lower value compared to both the yellow and blue curves for RGB image and point cloud respectively. It is also worth noticing that the green multimodal learning curves are similar and consistent across all eight test SCAND ROSBAGS with different social interactions in different social scenarios, showing the advantage of multimodal learning from both point cloud and RGB image.

Another very clear trend is that for the two unimodal perception types, point cloud perception consistently outperforms RGB image in all test trials, despite underperforming multimodal learning. In the earlier "easier" trials, point cloud performs slightly better than RGB image and has a relatively larger discrepancy compared to multimodal learning. For the later "difficult" trials, such trend is reversed, with the point cloud blue curves come closer to the multimodal green curves, compared to the RGB yellow curves.

Considering that there is no significant difference on the local planning loss curves across the eight test SCAND ROSBAGS, for the sake of space, we combine all eight curves into one for each modality and show them in Fig. 4. We observe a similar trend in learning local planning from all three perception modalities: Multimodal learning can achieve slightly better performance at imitating the SCAND demonstrations than learning with the point cloud, which further outperforms learning with RGB image.

B. Multimodal Social Compliance

In addition to the pure machine learning statistics, we also discuss how each perception modality performs with different social interactions in different social scenarios. As discussed above, the learning performance of RGB image decreases from the first to the last test trial and results in a more-than-doubled loss value in Fig. 3, while multimodal learning and point cloud learning consistently maintain similar performance. We also list the majority of the social scenarios presented in each test SCAND ROSBAG at the top of each subfigure in Fig. 3. We observe that the increasing "difficulty" level (mostly for RGB images) directly corresponds to increased human density caused by more confined social spaces and larger number of humans in the crowd. While learning with RGB image produces performance only slightly worse than point cloud and multimodal learning in the first "with traffic" scenario, which is a relatively simple scenario on a wide open walkway on the UT Austin campus, including "against traffic" human crowds and constraining navigation on a sidewalk instead of an open walkway deteriorates the performance of learning with RGB image only (first rown in Fig. 3). When the "difficulty" level keeps increasing by adding more complex social scenarios such as "street crossing", "large group/crowd", and "narrow hallway", RGB image's performance keeps degrading. We posit that such performance degradation is caused by the increased complexity and variance in the RGB input, which prevent learning with RGB image only from generalizing to unseen data in challenging social scenarios. Furthermore, considering the lack of direct and explicit geometric information from RGB images, operating mobile robots in confined social spaces with large human crowds is also less safe compared to point cloud, whose geometric information can be utilized to assure safety, i.e., asserting a safe stopping behavior when the distance between the robot and the humans in the scene is too close. Such a lack of safety by relying only on RGB images is also apparent in our human study (see details in Sec. V).

The obvious gap between multimodal and point cloud learning is also of interest. While both of them are able to perform similarly across all eight test SCAND ROSBAGS, multimodal learning maintains a very consistent advantage over point cloud alone in terms of a lower converged loss

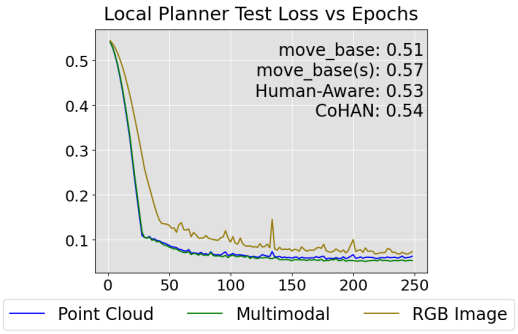

Fig. 4: Average Test Loss on All SCAND ROSBAGS with Multimodal, Point Cloud, and RGB Image Input (Three Training Runs).

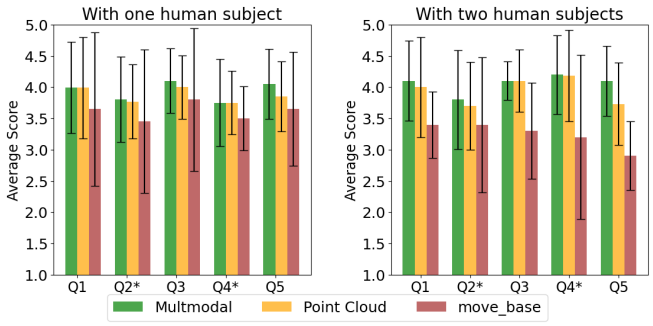

Fig. 5: Human Study Results.

value and fewer epochs until convergence. We posit that the additional semantic information provided by the RGB image in addition to the pure geometric data from point cloud can provide extra relevant social cues to inform social navigation decision making. Such an empirical gap reveals the necessity of including semantic information in the social robot navigation decision making process, compared to traditional autonomous mobile robot navigation, for which avoiding obstacles is the only concern.

V. HUMAN STUDY RESULTS

We conduct a human study to test whether the findings from our SCAND study can translate to real-world social robot navigation. We use a Clearpath Jackal robot with a Velodyne VLP-16 LiDAR and a ZED2 RGB-D camera for the point cloud and RGB image input respectively. We recruit eight human subjects for our human study.

Two sets of experiments are designed according to a previous protocol to evaluate social robot navigation [80]: frontal approach of the robot with one and two human participants in a public outdoor space (Fig. 6). In the one-human study, participants are instructed to take a natural path towards the robot; Participants in the two-human study are instructed to take three different approaches to initiate social interactions: move directly towards the robot, move forward then diverge, and move towards one side of the robot. After deploying the RGB module, we found that the robot may move dangerously close to the human subjects. Therefore, we exclude the RGB module in the human study.

After each human-robot interaction, we ask the participant

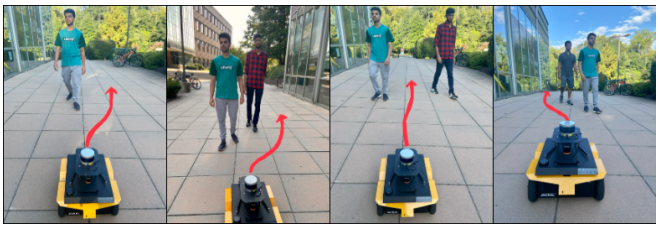

Fig. 6: Human Study with Different Social Scenarios.

to fill in a standard questionnaire [80] with five questions[2]: *1. The robot moved to avoid me*, *2. The robot obstructed my path**, *3. The robot maintained a safe and comfortable distance at all times*, *4. The robot nearly collided with me**, and *5. It was clear what the robot wanted to do*.

The per-question average along with error bars are plotted in Fig. 5 for both the one-person (left) and two-person scenarios (right). For all five questions, the multimodal learning approach is able to consistently achieve higher social compliance scores with smaller variance, compared to `move_base`, the best classical planner according to the loss values in the SCAND study. Compare the left and right figures, the difference between multimodal learning and `move_base` increases with more humans, showing multimodal learning's potential to enable socially compliant navigation with higher human density in public spaces, which is consistent with the results we observe in terms of test loss values in the SCAND study (Fig. 3). For our curated human study, we do not observe a significant advantage of multimodal learning in comparison to point cloud only. We posit that it is because our curated social scenarios do not contain sufficiently rich semantic social cues to showcase the necessity of using RGB images.

VI. CONCLUSIONS

We present a study on learning social robot navigation with multimodal (and unimodal) perception conducted on both a large-scale real-world social robot navigation dataset and in a human study with a physical robot, in comparison to a set of classical approaches. Our study results indicate that multimodal learning has clear advantage over either unimodal counterpart by a large margin in both the dataset and human studies, especially in difficult situations with increasing human density. In terms of unimodal learning, point cloud input is superior compared to RGB input, but it can be improved by utilizing the extra semantic information provided by the camera. Despite the found superiority of multimodal learning, the current study only remains in pre-recorded dataset and curated social scenarios. How multimodal learning will perform in real-world, large-scale, long-term social robot navigation tasks remains unclear and may require extra research and engineering effort.

REFERENCES

[1] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 4, no. 1, pp. 23–33, 1997.

[2]* denotes negatively formulated questions, for which we reverse-code the ratings to make them comparable to the positively formulated ones.

[2] S. Quinlan and O. Khatib, "Elastic bands: Connecting path planning and control," in *[1993] Proceedings IEEE International Conference on Robotics and Automation*. IEEE, 1993, pp. 802–807.

[3] D. Perille, A. Truong, X. Xiao, and P. Stone, "Benchmarking metric ground navigation," in *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2020, pp. 116–121.

[4] A. Nair, F. Jiang, K. Hou, Z. Xu, S. Li, X. Xiao, and P. Stone, "Dynabarn: Benchmarking metric ground navigation in dynamic environments," in *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2022, pp. 347–352.

[5] X. Xiao, Z. Xu, Z. Wang, Y. Song, G. Warnell, P. Stone, T. Zhang, S. Ravi, G. Wang, H. Karnan *et al.*, "Autonomous ground navigation in highly constrained spaces: Lessons learned from the benchmark autonomous robot navigation challenge at icra 2022 [competitions]," *IEEE Robotics & Automation Magazine*, vol. 29, no. 4, pp. 148–156, 2022.

[6] X. Xiao, Z. Xu, G. Warnell, P. Stone, F. G. Guinjoan, R. T. Rodrigues, H. Bruyninckx, H. Mandala, G. Christmann, J. L. Blanco-Claraco *et al.*, "Autonomous ground navigation in highly constrained spaces: Lessons learned from the 2nd barn challenge at icra 2023," *arXiv preprint arXiv:2308.03205*, 2023.

[7] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core challenges of social robot navigation: A survey," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[8] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Conflict avoidance in social navigation–a survey," *arXiv preprint arXiv:2106.12113*, 2021.

[9] A. Francis, C. Pérez-d'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *arXiv preprint arXiv:2306.16740*, 2023.

[10] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Motion planning and control for mobile robot navigation using machine learning: a survey," *Autonomous Robots*, vol. 46, no. 5, pp. 569–597, 2022.

[11] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[12] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics Research: The 14th International Symposium ISRR*. Springer, 2011, pp. 3–19.

[13] X. Xiao, T. Zhang, K. M. Choromanski, T.-W. E. Lee, A. Francis, J. Varley, S. Tu, S. Singh, P. Xu, F. Xia, S. M. Persson, L. Takayama, R. Frostig, J. Tan, C. Parada, and V. Sindhwani, "Learning model predictive controllers with real-time attention for real-world navigation," in *Conference on robot learning*. PMLR, 2022.

[14] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.

[15] X. Xiao, Z. Wang, Z. Xu, B. Liu, G. Warnell, G. Dhamankar, A. Nair, and P. Stone, "Appl: Adaptive planner parameter learning," *Robotics and Autonomous Systems*, vol. 154, p. 104132, 2022.

[16] X. Xiao, B. Liu, G. Warnell, J. Fink, and P. Stone, "Appld: Adaptive planner parameter learning from demonstration," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4541–4547, 2020.

[17] Z. Wang, X. Xiao, G. Warnell, and P. Stone, "Apple: Adaptive planner parameter learning from evaluative feedback," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7744–7749, 2021.

[18] Z. Wang, X. Xiao, B. Liu, G. Warnell, and P. Stone, "Appli: Adaptive planner parameter learning from interventions," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 6079–6085.

[19] Z. Xu, G. Dhamankar, A. Nair, X. Xiao, G. Warnell, B. Liu, Z. Wang, and P. Stone, "Applr: Adaptive planner parameter learning from reinforcement," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 6086–6092.

[20] S. M. Fiore, T. J. Wiltshire, E. J. Lobato, F. G. Jentsch, W. H. Huang, and B. Axelrod, "Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior," *Frontiers in psychology*, vol. 4, p. 859, 2013.

[21] J. Hart, R. Mirsky, X. Xiao, S. Tejeda, B. Mahajan, J. Goo, K. Baldauf, S. Owen, and P. Stone, "Using human-inspired signals to disambiguate navigational intentions," in *International Conference on Social Robotics*. Springer, 2020, pp. 320–331.

[22] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. Stone, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, 2022.

[23] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 421–10 434, 2022.

[24] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[25] A. J. Sathyamoorthy, U. Patel, M. Paul, N. K. S. Kumar, Y. Savle, and D. Manocha, "Comet: Modeling group cohesion for socially compliant robot navigation in crowded scenes," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1008–1015, 2021.

[26] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, "Covernet: Multimodal behavior prediction using trajectory sets," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 074–14 083.

[27] J. Buhmann, W. Burgard, A. B. Cremers, D. Fox, T. Hofmann, F. E. Schneider, J. Strikos, and S. Thrun, "The mobile robot rhino," *Ai Magazine*, vol. 16, no. 2, pp. 31–31, 1995.

[28] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Haehnel, C. Rosenberg, N. Roy *et al.*, "Probabilistic algorithms and the interactive museum tour-guide robot minerva," *The International Journal of Robotics Research*, vol. 19, no. 11, pp. 972–999, 2000.

[29] J. Joseph, F. Doshi-Velez, A. S. Huang, and N. Roy, "A bayesian nonparametric approach to modeling motion patterns," *Autonomous Robots*, vol. 31, no. 4, pp. 383–400, 2011.

[30] M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun, "Learning motion patterns of people for compliant robot motion," *The International Journal of Robotics Research*, vol. 24, no. 1, pp. 31–48, 2005.

[31] 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," *International Journal of Social Robotics*, vol. 6, no. 3, pp. 443–455, 2014.

[32] V. V. Unhelkar, C. Pérez-D'Arpino, L. Stirling, and J. A. Shah, "Human-robot co-navigation using anticipatory indicators of human walking motion," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 6183–6190.

[33] P. Xu, J.-B. Hayet, and I. Karamouzas, "Socialvae: Human trajectory prediction using timewise latents," in *European Conference on Computer Vision*. Springer, 2022, pp. 511–528.

[34] R. A. Knepper and D. Rus, "Pedestrian-inspired sampling-based multi-robot collision avoidance," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 94–100.

[35] E. A. Sisbot, L. F. Marin-Urias, R. Alami, and T. Simeon, "A human aware mobile robot motion planner," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 874–883, 2007.

[36] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras, "People tracking with human motion predictions from social forces," in *2010 IEEE international conference on robotics and automation*. IEEE, 2010, pp. 464–469.

[37] H. Gupta, B. Hayes, and Z. Sunberg, "Intention-aware navigation in crowds with extended-space pomdp planning," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022, pp. 562–570.

[38] P. Xu and I. Karamouzas, "Pfpn: Continuous control of physically simulated characters using particle filtering policy network," in *Proceedings of the 14th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2021, pp. 1–12.

[39] ——, "Human-inspired multi-agent navigation using knowledge distillation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8105–8112.

[40] X. T. Truong, Y. S. Ou, and T.-D. Ngo, "Towards culturally aware robot navigation," in *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*. IEEE, 2016, pp. 63–69.

[41] E. T. Hall, *The hidden dimension*. Garden City, NY: Doubleday, 1966, vol. 609.

[42] R. Kirby, R. Simmons, and J. Forlizzi, "Companion: A constraint-optimizing method for person-acceptable navigation," in *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2009, pp. 607–612.

[43] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: improving robot readability with animation principles," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2011, pp. 69–76.

[44] E. Torta, R. H. Cuijpers, and J. F. Juola, "Design of a parametric model of personal space for robotic social navigation," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 357–365, 2013.

[45] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[46] C. I. Mavrogiannis, W. B. Thomason, and R. A. Knepper, "Social momentum: A framework for legible navigation in dynamic multi-agent environments," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 361–369.

[47] M. Vázquez, E. J. Carter, J. A. Vaz, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Social group interactions in a role-playing game," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 2015, pp. 9–10.

[48] X. Xiao, J. Biswas, and P. Stone, "Learning inverse kinodynamics for accurate high-speed off-road navigation on unstructured terrain," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 6054–6060, 2021.

[49] H. Karnan, K. S. Sikand, P. Atreya, S. Rabiee, X. Xiao, G. Warnell, P. Stone, and J. Biswas, "Vi-ikd: High-speed accurate off-road navigation using learned visual-inertial inverse kinodynamics," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3294–3301.

[50] P. Atreya, H. Karnan, K. S. Sikand, X. Xiao, S. Rabiee, and J. Biswas, "High-speed accurate robot control using learned forward kinodynamics and non-linear least squares optimization," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 789–11 795.

[51] K. S. Sikand, S. Rabiee, A. Uccello, X. Xiao, G. Warnell, and J. Biswas, "Visual representation learning for preference-aware path planning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 303–11 309.

[52] A. Datar, C. Pan, and X. Xiao, "Learning to model and plan for wheeled mobility on vertically challenging terrain," *arXiv preprint arXiv:2306.11611*, 2023.

[53] A. Datar, C. Pan, M. Nazeri, and X. Xiao, "Toward wheeled mobility on vertically challenging terrain: Platforms, datasets, and algorithms," *arXiv preprint arXiv:2303.00998*, 2023.

[54] B. Kim and J. Pineau, "Socially adaptive path planning in human environments using inverse reinforcement learning," *International Journal of Social Robotics*, vol. 8, no. 1, pp. 51–66, 2016.

[55] D. Vasquez, B. Okal, and K. O. Arras, "Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1341–1346.

[56] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 3931–3936.

[57] J. Liang, U. Patel, A. J. Sathyamoorthy, and D. Manocha, "Crowd-steer: Realtime smooth and collision-free robot navigation in densely crowded scenarios trained using high-fidelity simulation," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 4221–4228.

[58] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Toward agile maneuvers in highly constrained spaces: Learning from hallucination," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1503–1510, 2021.

[59] X. Xiao, B. Liu, and P. Stone, "Agile robot navigation through hallucinated learning and sober deployment," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 7316–7322.

[60] Z. Wang, X. Xiao, A. J. Nettekoven, K. Umasankar, A. Singh, S. Bommakanti, U. Topcu, and P. Stone, "From agile ground to aerial navigation: Learning from learned hallucination," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 148–153.

[61] A. Francis, A. Faust, H.-T. L. Chiang, J. Hsu, J. C. Kew, M. Fiser, and T.-W. E. Lee, "Long-range indoor navigation with prm-rl," *IEEE Transactions on Robotics*, vol. 36, no. 4, pp. 1115–1134, 2020.

[62] Z. Xu, X. Xiao, G. Warnell, A. Nair, and P. Stone, "Machine learning methods for local motion planning: A study of end-to-end vs. parameter learning," in *2021 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2021, pp. 217–222.

[63] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1343–1350.

[64] L. Tai, J. Zhang, M. Liu, and W. Burgard, "Socially compliant navigation through raw depth inputs with generative adversarial imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1111–1117.

[65] M. Pfeiffer, M. Schaeuble, J. Nieto, R. Siegwart, and C. Cadena, "From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots," in *2017 ieee international conference on robotics and automation (icra)*. IEEE, 2017, pp. 1527–1533.

[66] Z. Xu, B. Liu, X. Xiao, A. Nair, and P. Stone, "Benchmarking reinforcement learning techniques for autonomous navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9224–9230.

[67] D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar, and X. Xiao, "Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.

[68] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[69] M. H. Nazeri and M. Bohlouli, "Exploring reflective limitation of behavior cloning in autonomous vehicles," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1252–1257.

[70] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[71] K. Weerakoon, A. J. Sathyamoorthy, J. Liang, T. Guan, U. Patel, and D. Manocha, "Graspe: Graph based multimodal fusion for robot navigation in unstructured outdoor environments," 2023.

[72] A. Nguyen, N. Nguyen, K. Tran, E. Tjiputra, and Q. D. Tran, "Autonomous navigation in complex environments with deep multimodal fusion network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5824–5830.

[73] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Advances in neural information processing systems*, vol. 25, 2012.

[74] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," 2018.

[75] K. Li, M. Shan, K. Narula, S. Worrall, and E. Nebot, "Socially aware crowd navigation with multimodal pedestrian trajectory prediction for autonomous vehicles," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.

[76] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 922–928.

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[78] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[79] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[80] S. Pirk, E. Lee, X. Xiao, L. Takayama, A. Francis, and A. Toshev, "A protocol for validating social navigation policies," *arXiv preprint arXiv:2204.05443*, 2022.