Occupancy Diffusion Model for Trajectory Prediction (OED)

1st Atekha, Omoruyi
Department of Mechanical Engineering
Stanford University
Stanford, California
oatekha@stanford.edu

***Abstract*—In the context of robots integrating into human environments, the ability to anticipate pedestrian trajectories accurately is crucial for seamless and safe interactions. This proposal presents a novel approach to trajectory prediction by integrating visual occupancy and semantic analysis into diffusion models. Traditional methods have advanced trajectory prediction through positional and semantic data but lack in accounting for the pedestrian's visual occupancy and semantic class's impact. Our model proposes to address these limitations by incorporating the visual occupancy space occupied that a given pedestrian sees surrounding them. To quantify the effectiveness of our model, we employ the Frechet Distance to compare the loss between generated trajectories and actual movements, contrasting our results with predictions that overlook visual occupancy and semantic factors. Preliminary outcomes from a single-step predictor showcase a mean-squared error of just 0.04, underscoring the promise of our approach for achieving more precise and holistic trajectory predictions.**

***Index Terms*—Occupancy, Semantic Class, Trajectory Prediction, Diffusion**

I. Introduction

The integration of machine learning methods into computer vision has led to the development of a broad spectrum of models, algorithms, and technologies. These advancements build upon classical geometric approaches, such as photogrammetry, feature detection, and optical flow. However, despite the increase of multi-modal models, a notable gap remains in social navigation contexts [1]. Many trajectory models fail to consider the impact of the robot's navigational behavior on pedestrian trajectories or how the visual engagement and understanding of pedestrians and other navigators influence their navigation patterns. As mobile robots increasingly populate roads, warehouses, and pedestrian crossings, the need for models that accurately predict human behavior is paramount [1] [2].

Efforts to enhance individual tracking and trajectory forecasting have seen the use of sensor fusion techniques and the incorporation of augmented information through learned models for trajectory prediction. Notably, Generative Adversarial Networks (GANs) have been explored for predicting future trajectories with considerations for social compliance and long-term 4D analysis [3]. In [4] [5] and colleagues have introduced a generative model capable of forecasting the trajectories of multiple agents within a spatial-temporal grid, leveraging occupancy encoding and decoding for future trajectory prediction. Additionally, Conditional Variational Auto Encoders (CVAE), as seen in [6], have been extensively employed, encoding the past movements and environmental occupancy around a pedestrian and decoding various potential future paths. Recently, diffusion models have gained traction as a versatile generative technique across numerous learning challenges, including trajectory prediction. [7] [8] illustrate the efficacy of stochastic trajectory prediction models employing diffusion techniques for forecasting future pose trajectories, showcasing superior performance relative to traditional and bench-marked models like Trajectron++, [6]. Nevertheless, these approaches are computationally demanding and, while incorporating a social-temporal state reflecting the agents surrounding a track at a specific time step, they rely on a bird's eye view for tracking position rather than the pedestrian's actual visual perspective. Moreover, these studies do not fully explore how semantic information can enhance understanding and prediction of future trajectories, a critical oversight in addressing the core challenges of social navigation [9]

This paper introduces a novel Trajectory Prediction Diffusion Model, dubbed Occupancy-Trajectory Prediction (OTP). Drawing inspiration from LEDiffusion and MID (motion indeterminacy diffusion) models, OTP aims to refine trajectory diffusion by incorporating encoded occupancy and semantic classifications [7] [8]. Unlike LEDiffusion, which employs Gaussian noise as an initializer for trajectory prediction, OTP utilizes a custom-learned initializer [7]. During the reverse diffusion process, this model leverages the initializer to generate inferred diverse samples, allowing for the bypassing of multiple diffusion steps, thus streamlining the prediction process [4].

In this study, rather than relying solely on visual data from recorded pedestrian datasets, we employ techniques from [10] [11]to generate a Skinned Multi-Person Linear Model (SMPL) of pedestrians. By analyzing key points of joints (such as the head, neck, and shoulders), we can leverage past trajectories to ascertain the orientation and joints of an individual. This, in turn, enables us to deduce their visual gaze. Furthermore, by adhering to constraints based on the human field of view, we can infer an individual's perceived visual occupancy. Additionally, the identification of key points for

Identify applicable funding agency here. If none, delete this.

joints furnishes us with geometric information that assists in categorizing the type of pedestrian detected. This approach marks a significant departure from traditional methods, incorporating a more nuanced understanding of pedestrian behavior and perception.

While the diffusion model is still in development, considerable effort has been devoted to data processing to distill Occupancy and Semantic information. To validate the performance and extract necessary insights, we conducted tests using the Oxford Town Center Dataset [12]. These early findings underscore the potential of our approach to enhance trajectory prediction by integrating complex, real-world pedestrian dynamics into the model's framework.

A. Trajectory Prediction

Trajectory Prediction entails forecasting the potential future paths of objects or entities based on their historical movements. This process diverges from Tracking and State Estimation by focusing on predicting a range of future positions rather than a single next step. Traditionally rooted in Bayesian models, trajectory prediction relies on historical data to inform updated projections. Recent advancements have shifted towards machine learning models, particularly Generative Adversarial Networks (GANs) and Variational Auto Encoders (VAEs), to forecast a spectrum of possible future trajectories [7]. Works like [6] and [13], and others, leverage VAEs not only to predict diverse future trajectories but also to incorporate dynamic constraints and contextual environmental data, marking a significant evolution in the methodology of trajectory prediction.

B. Diffusion Models

Diffusion Models have become increasingly applied to a plethora generative tasks, such as synthesizing images from text or predicting trajectory [3] [14]. Grounded in the pioneering work of Ho et al., who further developed the foundational principles of Diffusion Models introduced in [Deep unsupervised learning using nonequilibrium thermodynamics], Denoising Diffusion Probabilistic Models (DDPMs) harness the principles of non-equilibrium thermodynamics. They methodically reconstruct data distributions through a sequence of steps, effectively 'diffusing' them into a specific state. Works [7] and [8] have adapted diffusion models for trajectory prediction through innovations such as the LeapFrog Diffusion Model and the MID Model. Despite these advancements, both models similarly rely on historical trajectory data and a comprehensive social-temporal encoding, which may not fully reflect the pedestrian's perspective.

C. Human Robot Collaboration

Trajectory Prediction is imperative for robust Human-Robot Collaboration in social navigation tasks, where an autonomous system must navigate in crowded environments with many possible pedestrians. Traditionally, many formulations of Trajectory Prediction for social navigation have decoupled prediction and planning of robot movement. Rather than coupled prediction and planning, which accounts for how humans react to a moving robot or other agents in a given environment. [1] describes that many methods assume a naive generalization of human motion, where the individual follows a "consistently goal-directed, compliant motion" [1]. However, in reality, pedestrians move differently depending on whether they are on a bike or in a wheelchair, pushing a stroller, staring at their phone, walking in a particular way, or having visual impairments. To effectively navigate diverse environments, autonomous systems need a more nuanced understanding of the type of pedestrian and their visual perception of a given environment.

II. Proposed Method

The objective of this project is to develop a model capable of predicting future trajectories based on past trajectory data, current visual occupancy, and the semantic classification of a pedestrian, using 4D Human Trajectory information. The diffusion policy is formalized in equations (1) and (2). The loss function, $\mathcal{L}$, calculates the Fréchet Loss between the predicted trajectory $X_{n,t+1}^{\text{predicted}}$ and the ground truth trajectory $X_{n,t+1}^{\text{true}}$ at time step t. N represents the normal (Gaussian) distribution, and $\mathbf{\Sigma}_\theta$ are the mean and covariance parameters of the normal distribution, respectively.

$$p(X_{n,t+1}|X_{n,t}, S_{n,t}, O_{n,t}) \tag{1}$$

$$\begin{aligned} p(X_{n,t}|X_{n,t+1}, S_{n,t}, O_{n,t+1}) = \\ \mathcal{N}(X_{n,t}; \mu_\theta(X_{n,t+1}, S_{n,t}, O_{n,t}), \\ \Sigma_\theta(X_{n,t+1}, S_{n,t}, O_{n,t})) \end{aligned} \tag{2}$$

$$\mathcal{L} = \text{FréchetLoss}(X_{n,t+1}^{\text{predicted}}, X_{n,t+1}^{\text{true}}) \tag{3}$$

Additionally, this project aims to create an initialization model that predicts one future time step of the trajectory for the diffusion model. $X_{n,t}$ denotes the future trajectory pose, $S_{n,t}$ the semantic classification of the trajectory, $O_{n,t}$ the occupancy, indicating the state of objects visible to $X_{n,t}$, and $X_{n,t}$ the previous poses of the object. Unlike conventional approaches that generate initial guesses from Gaussian noise, the designed initializer is trained on all examples and employs an MSME loss between $X_{n,t+1}^{\text{predicted}}$ and $X_{n,t+1}^{\text{true}}$.

$$p(X_{n,t+1}|X_{n,t}, S_{n,t}, O_{n,t}) \tag{4}$$

$$\mathcal{L} = \text{MSME}(X_{n,t+1}^{\text{predicted}}, X_{n,t+1}^{\text{true}}) \tag{5}$$

III. Proposed Model

A. Dataset

For this experiment, the Oxford Town Center Dataset was used [**?**]. The dataset is a 6-minute CCTV recording of Oxford Town Center of Pedestrians (Cyclists and Walking Individuals Individuals Pushing Strollers) navigating a busy

street. Training portions of this experiment utilized a small portion of data, reduced to 6 seconds, due to the computation cost of processing it into SMPL [11].

B. Data Processing

We utilized SLAHMR [15] to analyze the video. This model constructs global human trajectories on "in-wild" videos and produces 3D SMPL reconstructions of individuals. This provides valuable information such as 3D pose and joint information (73 joints, head to toe) [11] for each pedestrian in the video.

Fig. 1. *

(b) Detailed pose estimation across multiple individuals.

Fig. 2. Comparative visualization of SLAHMR outputs: (a) demonstrates the SLMP models' capability in pedestrian detection, while (b) showcases the detailed pose estimation across multiple individuals.

C. Occupancy

Occupancy, $O_{n,t}$ represents what the pedestrian in track n can see at a given time step t. To capture a true sense of the pedestrian's occupancy, their relative orientation and the tilt of their head must be captured. The output of SLAHMR gives us the complete pose of a given individual, the translation $X = (x, y, z)$ along with the root orientation $\Phi_{n,t} = (\theta, \phi, \psi)$, and 23 body joint poses $\beta_{n,t}$. Utilizing $\beta_{n,t}$ and $\Phi_{n,t}$ we can determine the heading angle of the individual. Given $\Phi_{n,t}$ we know the horizontal direction and can easily determine the horizontal Field of View (FOV) of human eyes. Using $\beta_{n,t}$ we can form the vertical from looking at the relative relationship between the shoulders, neck, and head joint. A plane is formed between the shoulder and joints, subsequently the head joint is then used to form a plane. The angle between the head and shoulder neck plane, in the case one is looking straight is 20-35 degrees.

Given the current position of the observer $c_{pose} = (x, y)$ and the positions of N pedestrians $P = \{(x_i, y_i)\}_{i=1}^{N}$, the heading vector $\vec{h}$ is defined by the observer's heading angle θ. The displacement vector to pedestrian i is $\vec{d_i}$, and the angle α_i between $\vec{h}$ and $\vec{d_i}$ is calculated in equation 6.

$$\alpha_i = \arccos\left(\frac{\vec{h} \cdot \vec{d_i}}{\|\vec{h}\|\|\vec{d_i}\|}\right) \tag{6}$$

The resolution R_i for each pedestrian is determined by their angular position relative to the observer's FOV. Therefore, the occupancy O_n for each pedestrian is defined as:

$$O_n = \{(x_i, y_i, R_i)\}_{i=1}^{N} \tag{7}$$

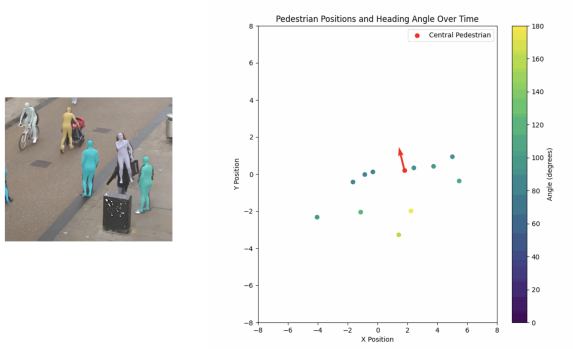

Fig. 3. Illustration of the Single Step Pose Model Pipeline.

D. Semantic Information

Given the pose of each trajectory projected onto the camera frame, we know the pixel position u and v of every object, given this position we perform semantic segmentation and utilize the Segment Anything Model (SAM) to produce the confidence of the semantic class of the selected point is threshold-ed into one of three categories which are Pedestrian, Cyclist, Stroller, as a 3x1 binary vector. The representation is shown in figure 7.

$$\begin{aligned} S_{n,\text{pedestrian}} &= [0, 0, 1], \\ S_{n,\text{cyclist}} &= [0, 1, 0], \\ S_{n,\text{stroller}} &= [1, 0, 0] \end{aligned} \tag{8}$$

E. Single Step Pose Initialization

A GRU model is used to perform single step pose initiation, that predicts the future position X_{t+1} of an agent given its current position X_t, semantic class S_t, and the occupancy grid O_t of other agents. The model is designed to process inputs $X \in \mathbb{R}^{N\times T\times 4}$, representing positions and orientations of N agents over T time steps, $S \in \mathbb{R}^{N\times T\times 1}$ for semantic classes, and $O \in \mathbb{R}^{N\times T\times (N-1)\times 3}$ for the occupancy information. It aims to estimate the conditional probability $p(X_{n-1}|X_n, S_n, O_n)$. The architecture combines an MLP Encoder to transform O_t into a latent space, with a TrajectoryEncoder that leverages spatial convolutions and GRU layers to capture spatial-temporal dynamics. The GRU

model uses a MSME loss between the predicted $X_{t+1}^{predicted}$ and true X_{t+1}^{true}. Training was done over 10 epochs for the singular example.

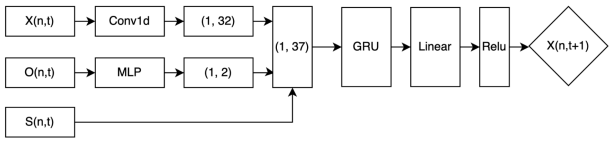

Fig. 4. Illustration of the Single Step Pose Model Pipeline. This graphic depicts the model's pipeline with which the model forecasts the next position in a trajectory, utilizing only .

IV. Results

A. Experiment

The experiment focused on training the Single Step Pose Initialization model using 6 seconds of processed data. Within this timeframe, we captured 148 frames, each containing 12 pedestrian tracks, resulting in a total of 1,776 unique training examples. These examples were organized into a PyTorch data loader and structured into four categories for each track across the 148 frames: X, S, O, and X_{pred}. The objective was to predict X_{pred} by utilizing the variables X, S, and O across the 1,776 unique state examples. Owing to limitations in time and computational resources, we did not partition the data into test and validation sets. However, future experiments will include these datasets to provide a more comprehensive evaluation of the model's performance.

B. Performance

After undergoing 10 epochs of training, the Single Step Pose Initialization model achieved a Mean Squared Mean Error (MSME) of 0.04. The effectiveness of this model was visually demonstrated, with numerous single-step prediction examples showing outputs that closely approximated the actual positions within the trajectories. As shown in figures 5 and 6, the model comes extremely close to predicting the correct pose.

V. Discussion

This research project demonstrates considerable potential. The initializer's low error rate is promising, particularly for generating diverse predictions within a diffusion pipeline akin to the one described in Leapfrog Diffusion. However, the computational demand for data processing and derivation of necessary information, such as the Field of View (FOV) and individual occupancy, is substantial. Currently, deploying such a system in real-time within an autonomous stack seems impractical. Future efforts might explore simplifying the computational process by selecting a reduced set of specific joints for faster processing times.

The project's success in deriving Occupancy from the SLAHMR model represents a significant achievement, paving the way for novel pedestrian prediction algorithms. Moreover, while the model currently categorizes semantic examples into three distinct classes, there is potential to expand this classification to include more nuanced categories, such as age, to better predict behavior patterns.

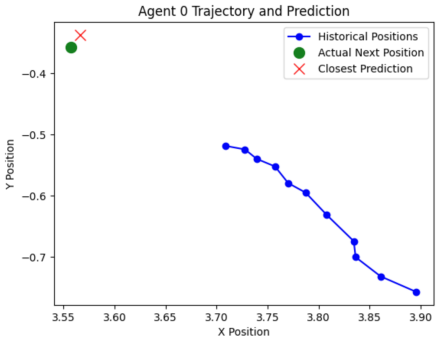

Fig. 5. (a) Trajectory Prediction for Agent 0.

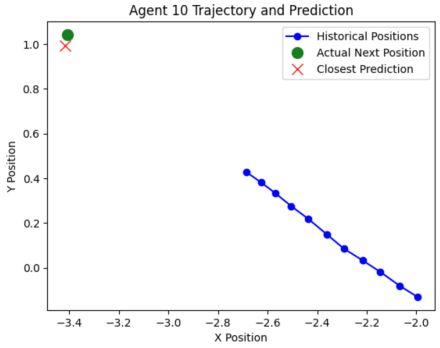

Fig. 6. (b) Trajectory Prediction for Agent 10.

Fig. 7. Performance evaluation of the Single Step Predictor on four pedestrian trajectories. Subfigures (a) and (b) present comparisons between the predicted and actual next states for Pedestrians 0 and 10 at a given time step.

VI. Future Directions

Given more time, the diffusion model of this quarter's ambitious project would first be completed. In future iterations of this project, we aim to conduct a rigorous comparison between our diffusion model initializer and established traditional methods, focusing on performance and efficiency. To enhance the model's predictive capabilities, we plan to significantly expand our dataset, ensuring it encompasses a wider variety of scenarios and interactions. Concurrently, we will refine the diffusion model's formulation to strike an optimal balance between computational efficiency and predictive accuracy.

A pivotal future direction of our work involves leveraging Zoe Depth [16] to obtain 3D information about individuals. This approach will allow us to form a 3D semantic visual occupancy, capturing all the objects within a pedestrian's view in a semantic and 3D embedded latent space. This novel strategy transcends the limitations of mere occupancy mapping, by revealing how an individual's actions may be influenced by the semantic class of observed objects and their surrounding context, including elements like sidewalks. Utilizing Zoe Depth in combination with CLIP will facilitate the extraction of both depth and semantic information from the

scene, which, along with the SLAHMR model's capabilities in determining head joint positions, will enhance our understanding of pedestrian dynamics.

To tackle the challenge of noise in root orientation data derived from the SLAHMR model, the integration of a Kalman Smoother is anticipated. Furthermore, we propose a novel approach to trajectory prediction that shifts from a direct reliance on raw occupancy information to the adoption of a semantic embedding of agents' recent perceptual experiences. This strategic shift aims to minimize the impact of transient visual disturbances, such as head movements, on prediction accuracy. By doing so, the predicted trajectories will not only become more precise but also more reflective of the agents' contextual understanding and the dynamic nature of their environment.

REFERENCES

[1] 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. 36:1–36:39, Apr. 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3583741

[2] N. S. Selby, J. Ng, G. S. Stump, G. Westerman, C. Traweek, and H. H. Asada, "TeachBot: Towards teaching robotics fundamentals for human-robot collaboration at work," *Heliyon*, vol. 7, no. 7, p. e07583, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844021016868

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[4] L. Tai, J. Zhang, M. Liu, and W. Burgard, "Socially Compliant Navigation through Raw Depth Inputs with Generative Adversarial Imitation Learning," Feb. 2018, arXiv:1710.02543 [cs]. [Online]. Available: http://arxiv.org/abs/1710.02543

[5] R. Mahjourian, J. Kim, Y. Chai, M. Tan, B. Sapp, and D. Anguelov, "Occupancy Flow Fields for Motion Forecasting in Autonomous Driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5639–5646, Apr. 2022, arXiv:2203.03875 [cs]. [Online]. Available: http://arxiv.org/abs/2203.03875

[6] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-Feasible Trajectory Forecasting With Heterogeneous Data," Jan. 2021, arXiv:2001.03093 [cs]. [Online]. Available: http://arxiv.org/abs/2001.03093

[7] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, "Leapfrog Diffusion Model for Stochastic Trajectory Prediction," Mar. 2023, arXiv:2303.10895 [cs]. [Online]. Available: http://arxiv.org/abs/2303.10895

[8] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu, "Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion," Mar. 2022, arXiv:2203.13777 [cs]. [Online]. Available: http://arxiv.org/abs/2203.13777

[9] C. Mavrogiannis, P. Alves-Oliveira, W. Thomason, and R. A. Knepper, "Social Momentum: Design and Evaluation of a Framework for Socially Competent Robot Navigation," *ACM Transactions on Human-Robot Interaction*, vol. 11, no. 2, pp. 14:1–14:37, Feb. 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3495244

[10] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and Tracking Humans with Transformers," Aug. 2023, arXiv:2305.20091 [cs]. [Online]. Available: http://arxiv.org/abs/2305.20091

[11] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: a skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2816795.2818013

[12] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR 2011*, Jun. 2011, pp. 3457–3464, iSSN: 1063-6919. [Online]. Available: https://ieeexplore.ieee.org/document/5995667

[13] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It Is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction," Jul. 2020, arXiv:2004.02025 [cs]. [Online]. Available: http://arxiv.org/abs/2004.02025

[14] Y. Zheng, Y. Yang, K. Mo, J. Li, T. Yu, Y. Liu, C. K. Liu, and L. J. Guibas, "GIMO: Gaze-Informed Human Motion Prediction in Context," in *Computer Vision – ECCV 2022*, ser. Lecture Notes in Computer Science, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 676–694.

[15] V. Ye, G. Pavlakos, J. Malik, and A. Kanazawa, "Decoupling Human and Camera Motion from Videos in the Wild," Mar. 2023, arXiv:2302.12827 [cs]. [Online]. Available: http://arxiv.org/abs/2302.12827

[16] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth," Feb. 2023, arXiv:2302.12288 [cs]. [Online]. Available: http://arxiv.org/abs/2302.12288