

# T2FPV: Dataset and Method for Correcting First-Person View Errors in Pedestrian Trajectory Prediction

Benjamin Stoler<sup>1</sup>, Meghdeep Jana<sup>2</sup>, Soonmin Hwang<sup>3</sup>, and Jean Oh<sup>3</sup>

**Abstract**—Predicting pedestrian motion is essential for developing socially-aware robots that interact in a crowded environment. While the natural visual perspective for a social interaction setting is an egocentric view, the majority of existing work in trajectory prediction therein has been investigated purely in the top-down trajectory space. To support first-person view trajectory prediction research, we present T2FPV, a method for constructing high-fidelity first-person view (FPV) datasets given a real-world, top-down trajectory dataset; we showcase our approach on the ETH/UCY pedestrian dataset to generate the egocentric visual data of all interacting pedestrians, creating the T2FPV-ETH dataset. In this setting, FPV-specific errors arise due to imperfect detection and tracking, occlusions, and field-of-view (FOV) limitations of the camera. To address these errors, we propose CoFE, a module that further refines the imputation of missing data in an end-to-end manner with trajectory forecasting algorithms. Our method reduces the impact of such FPV errors on downstream prediction performance, decreasing displacement error by more than 10% on average. To facilitate research engagement, we release our T2FPV-ETH dataset and software tools<sup>§</sup>.

## I. INTRODUCTION

As more and more autonomous robots are anticipated to interact with people in shared environments, trajectory prediction in robotics has become increasingly popular in the research community, as well as among various industry and military stakeholders. In particular, predicting pedestrian motion is essential for developing socially-aware robots that interact in a crowded environment [1], [2], [3], [4], [5]. Existing state-of-the-art (SOTA) trajectory prediction algorithms leverage datasets such as the ETH/UCY pedestrian dataset that provide full trajectory information of all pedestrians in a bird’s-eye view (BEV) scene [6]. However, bird’s-eye view is an unrealistic view for agents navigating in the real-world; agents generally rely on egocentric, first-person view (FPV) sensing for these tasks. A realistic setting also includes limited field-of-view (FOV), occlusions, and changes in perspective and orientation of the ego-agent.

While collecting top-down data using an overhead camera is relatively convenient, creating a first-person view counterpart is far more challenging for several reasons. To begin with, all participants in the scene would need to wear a camera sensor to record their egocentric views, as well as a location-recording sensor to establish their ground truth

locations. Furthermore, such a setting is subject to psychological issues such as the observer (or Hawthorne) effect [7], where people’s behaviors in these experiments may not be entirely representative of natural social interaction.

Therefore, to promote research on first-person view trajectory prediction, we propose T2FPV, a method for constructing an FPV version of data from a trajectory-only dataset by simulating the agents in high fidelity. The FPV data is collected by having each agent follow their recorded trajectory with a simulated camera attached to them. We perform extensive annotation and post-processing to provide unique information beyond existing FPV datasets as follows: 1) we conduct SOTA detection-and-tracking, giving realistic partial perception of trajectories and enforcing data imputation as a core aspect of the task (in contrast with prior works which simply re-provide ground-truth BEV trajectories [4], [8]); 2) our approach utilizes SEANavBench [9], a high-fidelity simulation environment, to provide realistic synthetic images; and 3) we additionally provide the corresponding ground truth of all observed and missed points of each trajectory, for its history and future (compared to only having information perceived from the camera as in [3], [10], [11]). An overview of our approach is shown visually in Figure 1.

To showcase our approach, we construct the T2FPV-ETH dataset based on the ETH/UCY trajectory dataset [6]. In this realistic FPV setting, we observe that a new class of errors is present compared to in BEV. These “FPV errors” arise from occlusion and field-of-view (FOV) limitations of robot sensing, combined with imperfect detection and tracking, resulting in missing observations. When performing trajectory prediction with various SOTA approaches, these errors caused our observed metrics to be significantly worse than what was reported in the BEV setting in the literature\*.

Prior work in pedestrian prediction has largely ignored FPV errors, either throwing out incomplete tracks or relying on simple interpolation over the missing points [13] [14]. Recent work has made significant advancement in data imputation; however, the vast majority of these works focus on artificially missing data [15], [16], [17]. Additionally, these works only focus on imputation as an independent task without considering how it affects prediction performance. Hence, to reduce the FPV errors for improved prediction, we propose Correction of FPV Errors (CoFE) that can refine initial imputations via end-to-end training with a trajectory prediction approach. We find that our approach decreases

<sup>1</sup>Computer Science Dept., Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA bstoler@cs.cmu.edu

<sup>2</sup> Mechanical Engineering Dept., Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA mjana@andrew.cmu.edu

<sup>3</sup>Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA {soonminh, jeanoh}@cmu.edu

<sup>§</sup><https://github.com/cmubig/T2FPV>

\*For instance, Average Displacement Error (ADE) / Final Displacement Error (FDE) performance increased from 0.44m / 0.89m in BEV to 1.51m / 2.08m in FPV for VRNN [12]

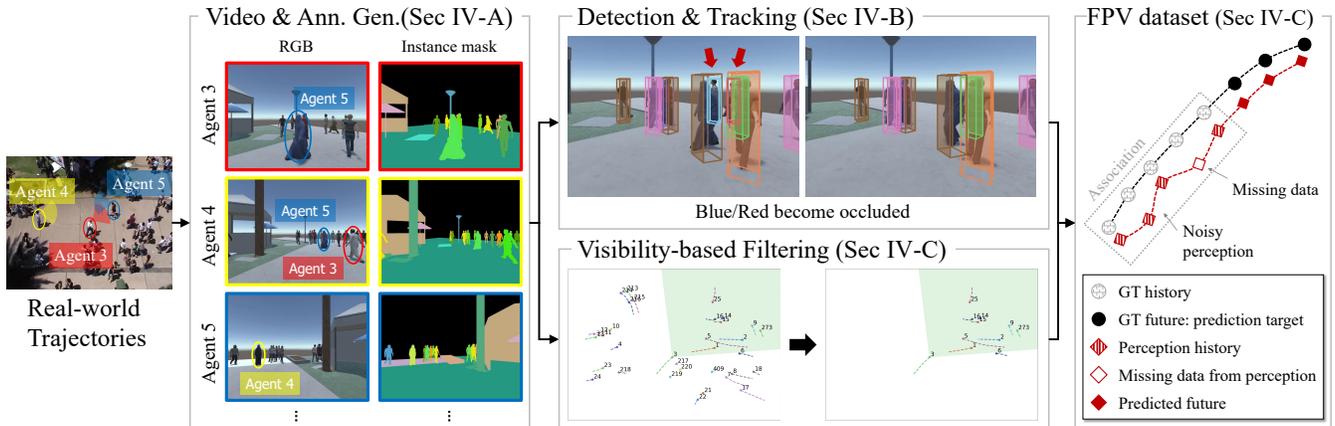


Fig. 1: **T2FPV Overview**: We generate filtered ground truth tracks and the corresponding D&T tracks from a real-world pedestrian dataset. The downstream task is to predict the future path (black circles) for a set of perceived track histories (striped red diamonds).

prediction displacement errors by more than 10% on average when compared to all tested imputation and forecasting combinations.

Our main contributions are: 1) we propose a method for creating an egocentric view for each agent given a set of trajectories; 2) we generate the T2FPV-ETH dataset, a new first-person view dataset that corresponds to the ETH/UCY dataset; 3) we propose and evaluate CoFE, an end-to-end learned input correction module, which reduces the impact of FPV errors beyond SOTA imputation approaches; and 4) we release our dataset and software tools to promote research in first-person view trajectory prediction.

## II. RELATED WORK

**Real-World First-Person Datasets.** Various large-scale datasets provide video footage from an ego agent’s perspective. [18] is a large-scale first-person view video dataset, with over 3500 hours of footage collected from various sources around the world. Egocentric Basketball Motion Planning [19] provides a wearable camera perspective from multiple people in the scene. However, neither of these datasets are focused on social navigation. They feature many instances of the ego agent walking by themselves or performing an unrelated task (such as carpentry, basketball, etc.) that have inherently different social contexts than navigating in public.

[10] is a dataset of an egocentric pedestrian video stream, providing pose, acceleration, and orientation information. Similarly, [3] uses human camera wearers and pose estimation to create a dataset of 2D targets to predict. However, these approaches only provide a single perspective of an ego agent in each scenario, limiting the diversity of ego behaviors. Also, both lack the ground truth pose information of other agents in the scene, especially due to occlusion and FOV limits.

Self-driving datasets, such as [20], [21], suffer the same problem of not having full ground truth information for training and evaluation. Furthermore, a car’s ego-motion is incomparable to a pedestrian’s ego-motion in terms of

physical characteristics. Additionally, the social interactions and roles between the ego and detected agents are vastly different between the two fields.

**Synthetic Pedestrian Datasets and Simulation.** Several recent works have generated synthetic data in simulations based on a corresponding real-world dataset. FvTraj [4] uses Unity to render FPV images from ground truth trajectory data [22], but these rendered images consist only of a flat ground plane with no corresponding environment modeled. DeepSocNav [8] generates ego view depth images from ETH/UCY, with a low-fidelity environment model. However, they do not include images from RGB cameras, which are far more common than depth sensors. Furthermore, DeepSocNav [8] and FvTraj [4] do not release any generated images or their in-house simulators, limiting reproducibility and engagement within the research community. Most importantly, both works only use the generated images for augmenting ground truth trajectories when performing prediction; no perception or detection and tracking is used, keeping the task less realistic.

[23] and [9] are relatively high-fidelity simulation environments with scene constructions of ETH/UCY, but both lack first-person views. Furthermore, these approaches also have the same aforementioned limitations as [8], [4] regarding perception and task settings.

**Pedestrian Trajectory Prediction.** Recent work on trajectory prediction and forecasting has mostly focused on top-down trajectory datasets such as ETH/UCY [6], SDD [24], and inD [25]. [1] uses LSTMs to jointly predict trajectories of all agents, incorporating pooled hidden-state information from neighbors as a social cue. Some approaches, such as AC-VRNN [12], use generative models within a VRNN [26], incorporating social interactions via attentive hidden state refinement. Several works also leverage top-down images explicitly, whether in an RGB form or with added semantic segmentation [5], [27], [28]. SNet [29] generates coarse step-wise goals to assist trajectory prediction sequentially. [30] incorporates agent dynamics and environment informa-

tion and forecasts using a graph-structured recurrent model.

Fewer works have focused on the FPV setting for pedestrians. [3] utilizes FPV to model and predict the trajectory of a single agent directly in pixel-space. [31] creates a spatial visual distribution of objects from FPV, and applies perception and ego-agent trajectory planning in a 2.5D coordinate system. [11] uses a transformer-based architecture, with a graph scene encoding to forecast the camera wearer’s trajectory with nearby agents as a cue. Still, none of these works deal with FPV errors when providing the trajectories of detections.

**Trajectory Robustness.** The field of sequence imputation has had much success with deep learning recently. NAOMI [16] uses a non-autoregressive approach at multiple step sizes to impute missing data in the context of basketball players and billiard ball trajectories. [17] trains imputation and prediction together but still only evaluates them separately rather than as an end-to-end pipeline. [32] evaluates the end-to-end task but only focuses on low-resolution, long-term GPS data, dissimilar to the fine-grained social navigation task. Also, these approaches only deal with artificial missing-completely-at-random (MCAR) data rather than dealing with pathologically missing data due to FPV errors.

[15] focuses on real vehicle and human motion trajectories, by transforming existing forecasting challenges into imputation ones. However, similar to the above approaches, they still only apply MCAR masks to ground truth. Furthermore, they only deal with imputation between ground truth points, rather than points which themselves may be erroneous from the perception model.

There have been several recent works in improving the robustness of trajectory forecasting to perception errors. [13] combines refinement via exponential smoothing with trajectory prediction to iteratively re-match observed trajectories with ground truth. [14] reframes the perception pipeline to remove tracking altogether, instead operating directly on detections and affinity matrices. However, both approaches still rely only on simple linear interpolation and extrapolation for missing data. While these are interesting approaches that we believe to be complementary to ours, we leave them as future work as they primarily focus on tracking and data association errors.

### III. PROBLEM FORMULATION

A trajectory prediction problem using complete information is defined as follows: for  $N$  pedestrians in a scene, we denote the position of each agent  $i$  in the  $xy$  ground-plane at time-step  $t$  as  $\mathbf{X}_i^t = (x_i^t, y_i^t)$ . Given the observed track histories,  $\mathbf{X}_i^{\text{hist}} = \{\mathbf{X}_i^t | t = 1, 2, \dots, T_{\text{obs}}\}$ , the task is to predict the future paths  $\mathbf{X}_i^{\text{fut}} = \{\mathbf{X}_i^t | t = T_{\text{obs}}+1, T_{\text{obs}}+2, \dots, T_{\text{pred}}\}$  for all agents  $i$  in a given scene, including the ego agent.

In this paper, we introduce a trajectory prediction task where each agent is to predict the trajectories of all agents in their view only using their egocentric information. Hence, observed non-ego track histories may be erroneous compared to the ground truth (GT). Thus, given that  $\mathbf{X}_i^t$  denotes the

ground truth position of a given agent  $i$  at time  $t$ , we will define  $\tilde{\mathbf{X}}_i^t$  to be the estimated position of agent  $i$  at time  $t$ . Similarly,  $\tilde{\mathbf{X}}_i^{\text{hist}}$  will represent an estimated history portion of the agent, where missing points have been imputed by some method. If the scene has  $N$  agents, note that there is a single ego agent  $e$ , and  $N - 1$  detected agents. Thus, the FPV trajectory prediction problem involves predicting  $\mathbf{X}_i^{\text{fut}}$  for each agent  $i$  in a scene, given  $\tilde{\mathbf{X}}_{d_i}^{\text{hist}}$  for each detected agent  $d_i$  and  $\mathbf{X}_e^{\text{hist}}$  for the ego agent,  $e$ .

### IV. TRAJECTORIES TO FIRST-PERSON VIEW

We describe our T2FPV method, demonstrating how we construct first-person view data from an example trajectory dataset, namely the ETH/UCY dataset. This dataset consists of five “folds” of recorded data, in different locations and times: ETH, Hotel, Univ, Zara1, and Zara2.

#### A. Video and Annotation Generation

Our approach for creating FPV datasets from real-world trajectory datasets begins with generating videos and ground-truth annotations. We use the SEANavBench [9] simulation environment as a starting point for our simulation. SEANavBench consists of high-fidelity pre-modeled scenes for each location within ETH/UCY. We leave these scenes as unchanged as possible, for consistency with prior works using SEANavBench.

As in [4], we enforce a number of assumptions when rendering these tracks. For instance, we orient each pedestrian’s gaze with the direction they are traveling in, with spherical linear interpolation for smoother angle changes. Additionally, we mount a camera on each pedestrian at a fixed height of  $1.6m$  from their base and assign the following physical characteristics to the camera:  $18mm$  focal length,  $36 \times 24mm$  sensor, and zero lens shift for the principal point. When rendered at our  $640 \times 480$  resolution, this results in a vertical FOV of approximately  $67^\circ$ .

Using the above assumptions, we then render the first-person videos for every person following their track from the original dataset, as well as output an annotation for each agent at every frame. The videos consist of the RGB render, as well as an instance segmentation render, as shown in Figure 1, where each object in the scene has been given a unique color. The annotations consist of the agent’s ID, pose information, and a list of what other agents can be seen in the camera’s view, i.e., the poses of all visible agents in both the camera and world reference frame. This detection list is generated by utilizing the aforementioned segmentation mask to determine agent visibility.

#### B. Perception: Detection and Tracking

To perform trajectory prediction in a realistic setting, we employed an off-the-shelf object detector and tracker to produce the observations required. We used a 3D object detector [33] which is SOTA among recent image-only methods which do not require depth information [34], and a simple but effective probabilistic tracker [35]. We made the

TABLE I: Detection and tracking performance.

Fold	Detection		Tracking	
	$AP_{2D}$	$AP_{BEV}$	AMOTA	AMOTP
ETH	96.50	44.10	0.384	1.262
Hotel	94.24	42.56	0.361	1.325
Univ	90.65	67.56	0.318	1.465
Zara1	97.29	90.22	0.709	0.610
Zara2	94.67	73.78	0.517	1.000

following changes to both approaches to produce reasonable detection and tracking results.

In DD3D [33], we set the parameters of feature map assignment to use thresholds that fit our ground truths appropriately. We also only used instances that are “visible” (as defined in Section IV-C), which helps to filter out heavily occluded instances. For the tracker [35], we changed the matching metric to use BEV IoU (Intersection-over-Union in top-down view) from Mahalanobis distance [36] to associate detections to tracks. We also applied the Kalman filter only to each instance’s 3D location and orientation and used state and observation noise covariances calculated from our ground truth data.

Following the common evaluation procedure as in the ETH/UCY trajectory prediction task, we trained one model for each of the five folds, using the other four folds as the training and validation sets respectively. We then produced tracking results on all ego videos from each fold’s test-set.

### C. FPV Dataset Creation

In transitioning from *bird’s-eye-view* (BEV) to *first-person view* (FPV), given a scene with  $N$  agents, we now construct  $N$  variations of the original scene, i.e., from each agent’s perspective. We begin with the same pre-processing popularized in Social GAN [2], only considering scenes with at least two concurrent agents in a sliding window consisting of  $T_{obs} = 8$  and  $T_{pred} = 12$  time steps. Then, to account for FPV errors, we redesign the scene as follows, for each agent’s perspective.

First, we consider the set of observed tracks from the detection and tracking module. We filter out tracks that are seen by the ego agent for fewer than  $k$  of the first  $T_{obs}$  time steps. Next, we perform an initial imputation on missing values using linear interpolation (as in [14]). This creates a D&T set of tracks, consisting of  $\tilde{\mathbf{X}}_{d_i}^{\text{hist}}$  for each detection.

We then consider the set of ground truth tracks from BEV. We filter out tracks that are impossible to have been seen by the ego agent for fewer than  $k$  of the first  $T_{obs}$  time steps (i.e. by having fewer than  $P$  pixels visible from instance segmentation). Furthermore, we filter out tracks for which the ground truth is missing pieces of data for any of  $T_{obs}$  or  $T_{pred}$ . In our creation of T2FPV-ETH, we used  $k = 3$  and  $P = 100$ . This step then creates a GT set of tracks, consisting of  $\mathbf{X}_{d_i}^{\text{hist}}$  as well as  $\mathbf{X}_{d_i}^{\text{fut}}$  for each agent which is feasibly visible to the ego agent.

For each scene, the GT and detected sets of tracks are associated together by performing Hungarian matching, as in [13], [37], [14], based on the mean squared error (MSE)

TABLE II: T2FPV-ETH statistics.

Fold	Num Ego	Num Dets	Det MSE	FPV Err. Rate
ETH	181	60	2.05	0.44
Hotel	1053	449	2.03	0.51
Univ	24,334	120,072	1.13	0.45
Zara1	5,939	3,686	0.64	0.28
Zara2	17,608	11,775	1.05	0.32

between each  $\mathbf{X}_{d_i}^{\text{hist}}$  and  $\tilde{\mathbf{X}}_{d_k}^{\text{hist}}$ . Finally, each scene has the corresponding ego agent  $\mathbf{X}_e^{\text{hist}}$  and  $\mathbf{X}_e^{\text{fut}}$  appended.

### D. Dataset Statistics

We measure the detection and tracking performances of the SOTA methods we employed in Table I. For detection performance, we measure the standard average precision ( $AP_{2D}$ ) in 2D image space and observe that it performs well. Also, we measure the localization quality of detected objects in 3D space by calculating IoU-based average precision in the top-down view ( $AP_{BEV}$ ). Both metrics use the same IoU threshold of 0.5. The  $AP_{BEV}$  performance is worse than  $AP_{2D}$ , which shows the challenge of image-based 3D detection. For tracking, we adopt two popular metrics from [38], Average Multi-Object Tracking Accuracy (AMOTA) and Precision (AMOTP). AMOTA combines false positives, missed targets, and identity switches, and AMOTP measures the misalignment between prediction and ground truth. Although “Univ” shows the worst performance because of the pedestrian density (Table II), the detector and tracker perform reasonably well, as shown qualitatively in Figure 1.

Table II provides a high-level overview of the number of scenes and detections, as created in Section IV-C. We note that this table demonstrates a data augmentation effect, as there is now a one-to-one correspondence between each ego agent and a scene; a single ground-truth track is often observed by multiple other agents at once, although with different possible FPV errors and 3D detection locations. These statistics indicate the diversity between the different folds as testing sets, as they have significantly varying scene densities (i.e. detections per ego agents), as well as rate of FPV errors (i.e. number of points needing to be imputed downstream) and difficulty of imputation.

## V. PROPOSED METHOD: CoFE

### A. Motivation

As noted in Section II, existing imputation approaches have two primary deficits when being applied to the field of human trajectory prediction. First, approaches largely use a missing-completely-at-random (MCAR) treatment of the points to be imputed. This assumption does not hold in a setting with FPV errors, as data is missing in a manner pathological to the detection and tracking approach being used as well as compulsory to occlusion and FOV limitations from the ego camera. Second, approaches have full trust in the accuracy of the points around the missing data. This assumption clearly also does not hold in the FPV setting, as

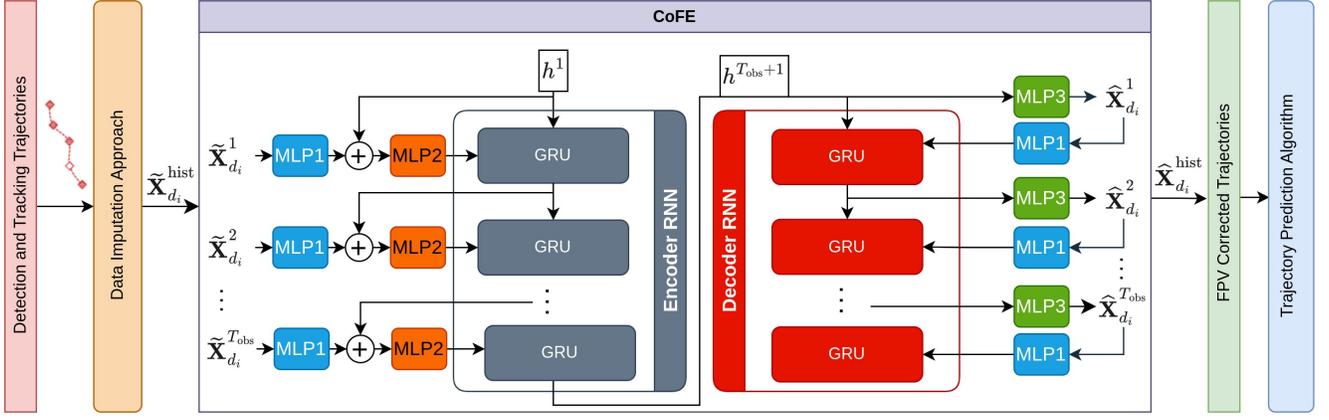


Fig. 2: **CoFE Approach**: CoFE has an encoder-decoder architecture that refines the imputed trajectories to better account for FPV errors. The corrected trajectories are then passed into a trajectory prediction algorithm.

the positions of observations are estimated from our approach described in Section IV-B.

Therefore, we propose to incorporate a correction module, **Correction of FPV Errors (CoFE)**, between existing imputation approaches and downstream trajectory prediction approach, as shown in Figure 2. To make the correction consistent with the trajectory prediction task, we thus train a neural network for both the imputation refinement and the trajectory prediction algorithm itself in an end-to-end (E2E) manner.

### B. CoFE Architecture

As shown in Figure 2, our architecture is similar to many RNN-based trajectory forecasting approaches, such as VRNN [26]. Our main insight is to use an encoder RNN and decoder RNN back to back. The accumulated hidden state after processing the input then feeds into the decoder to output a “corrected” version of the input. We utilize two different Gated Recurrent Unit (GRU) modules for this purpose.

For a detected agent  $d$ , each missing point of  $\tilde{\mathbf{X}}_{d_i}^{\text{hist}}$  is imputed via the given approach, e.g., NAOMI [16]. Then, each point is transformed into its relative motion from the previous point, in order to be agnostic to absolute coordinates in the given scene. Next, these relative motions are feature extracted with a Multilayer Perceptron (MLP), labeled MLP1 in our diagram. These features are concatenated with the hidden state  $h^t$  at each time step in  $T_{\text{obs}}$ , then fed into MLP2 and the encoding GRU cell to obtain the next hidden state,  $h^{t+1}$ . After processing the entire input, we then switch to decoding with the last hidden state,  $h^{T_{\text{obs}}+1}$ . We output  $\hat{\mathbf{X}}_{d_i}^1$  as a prediction for  $\mathbf{X}_{d_i}^1$ , via MLP3. We then apply the same feature extractor (i.e., same weights) MLP1 to this prediction to then feed back into the decoding GRU cell, and repeat this process for all  $T_{\text{obs}}$  points. Finally, we convert the predicted points back into absolute coordinates. The exact details of this architecture, including the number of layers and hidden units in each MLP, can be seen in our open-sourced implementation.

### C. End-to-End (E2E) Training

We introduce a simple MSE Loss objective to train CoFE itself, between the ground truth  $\mathbf{X}_{d_i}^{\text{hist}}$  and corrected estimations  $\hat{\mathbf{X}}_{d_i}^{\text{hist}}$ . Given the original estimated points with imputation  $\tilde{\mathbf{X}}_{d_i}^{\text{hist}}$  and also the refinements  $\hat{\mathbf{X}}_{d_i}^{\text{hist}}$ , we then update the points in  $\hat{\mathbf{X}}_{d_i}^t$  with  $\tilde{\mathbf{X}}_{d_i}^t$  for timesteps  $t$  where imputation is not required. This final  $\hat{\mathbf{X}}_{d_i}^t$  is then used to train the downstream prediction method (e.g., SGNet [29]) in an end-to-end (E2E) manner, where the Loss function being optimized is the sum of the CoFE Loss objective and the prediction method’s original objective.

## VI. EXPERIMENTS

### A. Experimental Setup

We implemented several representative approaches on the ETH/UCY trajectory prediction task. We selected these algorithms as they stood out along several key techniques common in human trajectory prediction: variational prediction (VRNN [26]), social awareness (A-VRNN [12]), and goal conditioning (SGNet [29]).

For data imputation, we incorporated three commonly used approaches. We selected linear interpolation (“Linear-interp”), a simple but powerful approach used as part of many recent works, such as [14]. We also selected double exponential smoothing (“Smooth”), used in [13], a more complex baseline that better handles dynamic trends in the sequence. Finally, we incorporated NAOMI [16], which is a recent SOTA approach leveraging deep learning.

### B. Evaluation Procedure

As in Social GAN [2], we evaluate trajectory predictions using a leave-one-out approach. For each of the five folds, models are trained and validated on data from four of them at a time. Then, the best model according to validation performance is tested on the entirety of the held-out fold.

We train NAOMI [16] separately, following the author’s procedure, once per fold. Then, for each combination of imputation techniques and prediction algorithms, we train

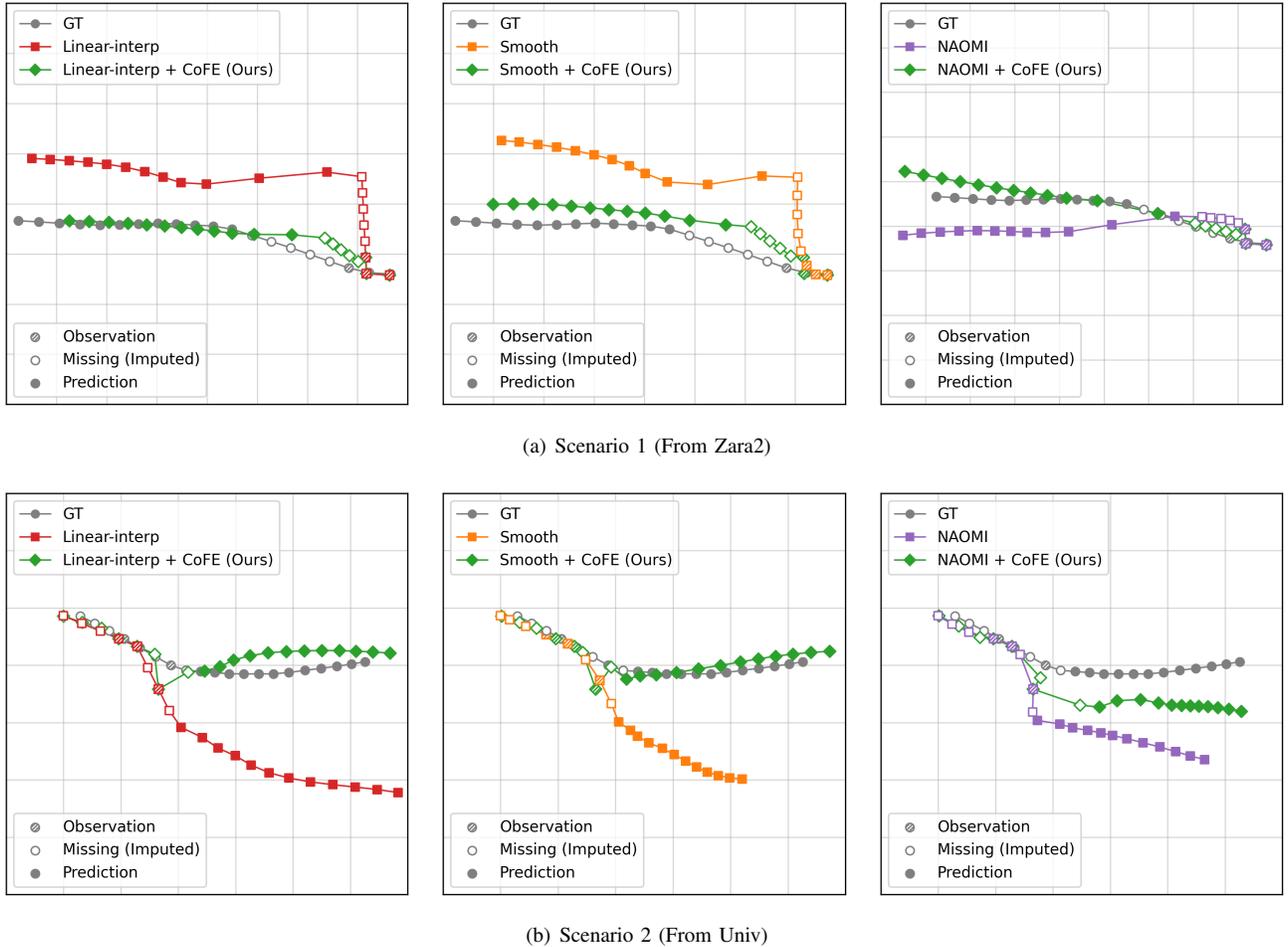


Fig. 3: **Qualitative Results:** We demonstrate the effectiveness of CoFE, when applied to different imputation methods, using SNet [29] for prediction. Grid lines represent  $1m$  in distance in ground plane.

one prediction model utilizing CoFE and one model without (i.e., just using the  $\hat{X}_i^{\text{hist}}$  outputs from the imputation).

In the field of trajectory prediction, and especially for ETH/UCY, the most commonly used metrics are Average Displacement Error (ADE) and Final Displacement Error (FDE). These metrics can be easily computed on a per-agent basis, for ground truth future track  $X_i^{\text{fut}}$  and predicted future track  $\hat{X}_i^{\text{fut}}$ , for each agent  $i$  in the scene. The  $L_2$ -distance at each time  $t$  from  $t = T_{\text{obs}+1}$  to  $t = T_{\text{pred}}$  is taken between  $X_i^t$  and  $\hat{X}_i^t$ ; ADE is the average of these distances, while FDE is the final distance.

We note that there are other metrics which could be utilized, including collision rate, social comfort level, path complexity, and many more [39]. We chose to focus on the core metrics of the tasks at hand, but suggest that future work in applying these metrics could provide helpful new insight.

### C. Results

We conducted extensive experimentation to assess CoFE’s performance when combined with the various imputation and prediction approaches. As shown in Table III, adding the CoFE module is quite effective. When considering the

average performance over all dataset folds, all combinations of imputation and prediction algorithms which use CoFE are better than the corresponding versions which bypass it. Furthermore, without CoFE, the average performance is highly variable, dependent on the choice of imputation approach and prediction method. However, with CoFE, these differences become much less pronounced. Note that although performance on most folds in most cases is quite good, CoFE appears to not be as effective on ETH. We suspect that this is because, as shown in Table II, there are only a small number of detected tracks in the first place (60), so performance is more variable and sensitive to any individual prediction’s error.

To gain further insight into CoFE’s performance, we performed qualitative analyses. As seen in Figure 3-(b), the imputation approach (NAOMI [16]) trusts surrounding points in the data, performing an extrapolation and thus, does not effectively capture the FPV errors. When paired with CoFE, the approach is more effective at capturing underlying temporal and spatial patterns in the data, correcting the FPV errors which results in better downstream prediction.

TABLE III: ADE / FDE for each fold and approach tested on T2FPV-ETH dataset. The better result between using CoFE and not is **bolded** and the overall best performance for each prediction method is **green**. Lower is better.

Unit: meter								
Traj. Prediction	Imputation	CoFE (ours)	ETH	Hotel	Univ	Zara1	Zara2	Avg
VRNN [26]	Linear-interp	-	<b>1.35 / 2.00</b>	1.30 / 1.73	2.24 / 2.89	1.14 / 1.68	1.54 / 2.10	1.51 / 2.08
	Linear-interp	✓	1.52 / 2.35	<b>1.06 / 1.53</b>	<b>1.65 / 2.10</b>	<b>1.06 / 1.63</b>	<b>1.27 / 1.62</b>	<b>1.31 / 1.84</b>
	Smooth [13]	-	2.25 / 3.75	1.53 / 2.36	2.17 / 3.00	1.26 / 1.95	1.58 / 2.22	1.76 / 2.65
	Smooth [13]	✓	<b>1.57 / 2.46</b>	<b>1.08 / 1.55</b>	<b>1.77 / 2.31</b>	<b>1.00 / 1.44</b>	<b>1.31 / 1.68</b>	<b>1.35 / 1.89</b>
	NAOMI [16]	-	<b>1.46 / 2.29</b>	1.59 / 2.17	1.83 / 2.31	0.96 / 1.57	1.13 / 1.49	1.39 / 1.97
	NAOMI [16]	✓	1.54 / 2.34	<b>1.09 / 1.55</b>	<b>1.58 / 1.97</b>	<b>0.92 / 1.39</b>	<b>1.11 / 1.39</b>	<b>1.25 / 1.73</b>
A-VRNN [12]	Linear-interp	-	<b>1.39 / 2.04</b>	1.31 / 1.75	2.26 / 3.00	1.04 / 1.40	1.47 / 1.93	1.49 / 2.03
	Linear-interp	✓	1.47 / 2.18	<b>1.16 / 1.72</b>	<b>1.54 / 1.88</b>	<b>1.03 / 1.48</b>	<b>1.31 / 1.69</b>	<b>1.30 / 1.79</b>
	Smooth [13]	-	1.77 / 3.11	1.36 / 1.81	2.27 / 3.34	1.18 / 1.72	1.59 / 2.07	1.63 / 2.41
	Smooth [13]	✓	<b>1.69 / 2.71</b>	<b>1.10 / 1.59</b>	<b>1.76 / 2.38</b>	<b>1.06 / 1.59</b>	<b>1.28 / 1.62</b>	<b>1.38 / 1.98</b>
	NAOMI [16]	-	<b>1.44 / 2.17</b>	1.66 / 2.30	1.82 / 2.25	<b>0.83 / 1.24</b>	<b>1.09 / 1.39</b>	1.37 / 1.87
	NAOMI [16]	✓	1.49 / 2.18	<b>1.16 / 1.66</b>	<b>1.54 / 1.91</b>	0.88 / 1.31	1.16 / 1.49	<b>1.25 / 1.71</b>
SGNet [29]	Linear-interp	-	1.43 / 1.97	0.72 / 1.00	1.48 / 1.73	0.58 / 0.79	0.78 / 0.91	1.00 / 1.28
	Linear-interp	✓	<b>0.98 / 1.32</b>	<b>0.59 / 0.76</b>	<b>1.23 / 1.48</b>	<b>0.55 / 0.76</b>	<b>0.73 / 0.86</b>	<b>0.82 / 1.04</b>
	Smooth [13]	-	1.06 / 1.45	0.73 / 1.04	1.45 / 1.68	0.57 / 0.78	0.79 / 0.93	0.92 / 1.18
	Smooth [13]	✓	<b>1.03 / 1.41</b>	<b>0.61 / 0.80</b>	<b>1.28 / 1.54</b>	<b>0.56 / 0.77</b>	<b>0.74 / 0.86</b>	<b>0.84 / 1.08</b>
	NAOMI [16]	-	<b>0.90 / 1.28</b>	0.78 / 0.97	1.21 / 1.43	<b>0.50 / 0.69</b>	0.72 / 0.84	0.82 / 1.04
	NAOMI [16]	✓	0.99 / 1.40	<b>0.59 / 0.78</b>	<b>1.16 / 1.39</b>	0.51 / 0.70	<b>0.67 / 0.79</b>	<b>0.78 / 1.01</b>

TABLE IV: Ablation study on CoFE applied to SGNet with linear interpolation.

Algorithm	CoFE	Train E2E	Impute Only	ADE / FDE
	-	-	-	1.00 / 1.28
SGNet [29]	✓	No	No	1.11 / 1.43
	✓	No	Yes	0.98 / 1.27
	✓	Yes	No	0.94 / 1.22
	✓	Yes	Yes	<b>0.82 / 1.04</b>

We also performed an ablation study, shown in Table IV, to assess aspects of our design choices. We focused on SGNet [29] combined with linear interpolation for the study, and found clearly that the E2E training was vital in obtaining top performance. We further find that focusing on imputation only in the prediction phase (i.e., replacing non-imputed points in  $\hat{\mathbf{X}}_{d_i}^{\text{hist}}$  as described in Section V-C) also has a significant effect in improving performance.

## VII. FUTURE WORK

Although SEANavBench is a high-fidelity environment, we do note that further effort in improving its realism could be useful. Realism could be enhanced not just by increasing the 3D-modeling asset and animation qualities, but also by further improving alignment between the reproduced scenery and the original locations.

Additionally, for associating D&T tracks with their corresponding GT tracks, we relied on Hungarian matching on our tracking output directly. This decreased the number of correctly matched trajectories, due to identity association errors of detections. Incorporating affinity-based techniques from [14] or performing the full re-tracking algorithm from

[13] could be a promising way to even further reduce FPV errors.

## VIII. CONCLUSION

In existing work, pedestrian trajectory prediction has been mainly studied under a complete information assumption. In this paper, we introduce a first-person view trajectory prediction problem where agents need to make predictions based on partial, imprecise information. To promote this research direction, we present T2FPV, a method for generating high-fidelity egocentric datasets for pedestrian navigation by leveraging existing real-world trajectory datasets. In this setting, FPV-specific errors arise due to imperfect detection and tracking, occlusions, and FOV limitations of the camera. To address these errors, we propose CoFE, a module that further refines imputation of missing data in an end-to-end manner with trajectory forecasting algorithms. Our method reduces the impact of such FPV errors on downstream prediction performance, decreasing displacement error by 11.73%, averaging over all combinations of imputation techniques and prediction approaches tested. We also show that E2E training of CoFE is essential in achieving this performance increase. Our constructed T2FPV-ETH dataset provides a benchmark for human trajectory prediction from detection and tracking results, which is a more natural and realistic setting. Therefore, we argue that incorporating such realism throughout the perception pipeline is an important direction to move toward in enabling robots to navigate in the real world.

## ACKNOWLEDGMENT

This work is in part funded by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute of Advancement of Technology (KIAT) through the International

Cooperative R&D program: P0019782, Embedded AI Based fully autonomous driving software and Maas technology development.

## REFERENCES

- [1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
- [2] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: socially acceptable trajectories with generative adversarial networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
- [3] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
- [4] H. Bi, R. Zhang, T. Mao, Z. Deng, and Z. Wang, "How can i see my future? fvtraj: Using first-person view for pedestrian trajectory prediction," in *European Conference on Computer Vision*, 2020.
- [5] J. Yue, D. Manocha, and H. Wang, "Human trajectory prediction via neural social physics," in *European Conference on Computer Vision*, 2022.
- [6] A. Ess, B. Leibe, K. Schindler, and L. van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2008.
- [7] B. Irfan, J. Kennedy, S. Lemaignan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Social psychology and human-robot interaction: An uneasy marriage," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2018.
- [8] J. P. de Vicente and A. Soto, "Deepsoconv: Social navigation by imitating human behaviors," *RSS Workshop on Social Robot Navigation* 2021, 2021.
- [9] N. Tsoi, M. Hussein, O. Fugikawa, J. D. Zhao, and M. Vázquez, "An approach to deploy interactive robotic simulators on the web for hri experiments: Results in social robot navigation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.
- [10] A. A. E. Krishna Kumar Singh, Kayvon Fatahalian, "Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks," in *IEEE Winter Conference on Applications of Computer Vision*, 2016.
- [11] J. Qiu, L. Chen, X. Gu, F. P.-W. Lo, Y.-Y. Tsai, J. Sun, J. Liu, and B. Lo, "Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion," *IEEE Robotics and Automation Letters*, 2022.
- [12] A. Bertugli, S. Calderara, P. Coscia, L. Ballan, and R. Cucchiara, "Acvrnn: Attentive conditional-vrnn for multi-future trajectory prediction," *Computer Vision and Image Understanding*, 2021.
- [13] R. Yu and Z. Zhou, "Towards robust human trajectory prediction in raw videos," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.
- [14] X. Weng, B. Ivanovic, K. Kitani, and M. Pavone, "Whose track is it anyway? improving robustness to tracking errors with affinity-based trajectory prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [15] C. Zhao, A. Song, Y. Du, and B. Yang, "Trajgat: A map-embedded graph attention network for real-time vehicle trajectory imputation of roadside perception," *Transportation research part C: emerging technologies*, vol. 142, p. 103787, 2022.
- [16] Y. Liu, R. Yu, S. Zheng, E. Zhan, and Y. Yue, "Naomi: Non-autoregressive multiresolution sequence imputation," *Advances in neural information processing systems*, vol. 32, 2019.
- [17] M. Qi, J. Qin, Y. Wu, and Y. Yang, "Imitative non-autoregressive modeling for trajectory forecasting and imputation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [18] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4d: Around the World in 3,000 Hours of Egocentric Video," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [19] G. Bertasius, A. Chan, and J. Shi, "Egocentric basketball motion planning from a single first-person image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018.
- [20] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.
- [21] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [22] A. Juliani, V. Berges, E. Vckay, Y. Gao, H. Henry, M. Mattar, and D. Lange, "Unity: A general platform for intelligent agents," *arXiv preprint arXiv:1809.02627*, 2018.
- [23] J. Liang, L. Jiang, K. P. Murphy, T. Yu, and A. G. Hauptmann, "The garden of forking paths: Towards multi-future trajectory prediction," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [24] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European Conference on Computer Vision*, 2016.
- [25] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein, "The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections," *IEEE Intelligent Vehicles Symposium*, 2020.
- [26] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Advances in neural information processing systems*, vol. 28, 2015.
- [27] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," *IEEE/CVF International Conference on Computer Vision*, 2020.
- [28] C. Wong, B. Xia, Z. Hong, Q. Peng, and X. You, "View vertically: A hierarchical network for trajectory prediction via fourier spectrums," in *European Conference on Computer Vision*, 2022.
- [29] K. Li, N. Y. Wang, Y. Yang, and G. Wang, "Sgnet: A superclass guided network for image classification and object detection," *Conference on Robots and Vision*, 2021.
- [30] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control," *European Conference on Computer Vision*, 2020.
- [31] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
- [32] K. K. Qin, Y. Ren, W. Shao, B. Lake, F. Privitera, and F. D. Salim, "Multiple-level point embedding for solving human trajectory imputation with prediction," *ACM Trans. Spatial Algorithms Syst.*, 2023, just Accepted.
- [33] D. Park, R. Amrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-lidar needed for monocular 3d object detection?" in *IEEE/CVF International Conference on Computer Vision*, 2021.
- [34] X. Ma, W. Ouyang, A. Simonelli, and E. Ricci, "3d object detection from images for autonomous driving: a survey," *arXiv preprint arXiv:2202.02980*, 2022.
- [35] H. kuang Chiu, A. Prioletti, J. Li, and J. Bohg, "Probabilistic 3d multi-object tracking for autonomous driving," *arXiv*, vol. abs/2001.05673, 2020.
- [36] G. J. McLachlan, "Mahalanobis distance," *Resonance*, 1999.
- [37] X. Weng, B. Ivanovic, and M. Pavone, "Mtp: multi-hypothesis tracking and prediction for reduced error propagation," in *IEEE Intelligent Vehicles Symposium*, 2022.
- [38] X. Weng, J. Wang, D. Held, and K. Kitani, "3d multi-object tracking: A baseline and new evaluation metrics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.
- [39] C. I. 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*, 2023.