The Stanford Drone Dataset Is More Complex Than We Think: An Analysis of Key Characteristics

Josh Andle, Nicholas Soucy, Simon Socolow, and Salimeh Yasaei Sekeh, *Member, IEEE*

***Abstract*—Several datasets exist which contain annotated information of individuals' trajectories. Such datasets are vital for many real-world applications, including trajectory prediction and autonomous navigation. One prominent dataset currently in use is the Stanford Drone Dataset (SDD) (Robicquet *et al.*, 2016). Despite its prominence, discussion surrounding the characteristics of this dataset is insufficient. We demonstrate how this insufficiency reduces the information available to users and can impact performance. Our contributions include the outlining of key characteristics in the SDD, employment of an information-theoretic measure and custom metric to clearly visualize those characteristics, the implementation of the PECNet (Mangalam *et al.*, 2020) and Y-Net (Mangalam *et al.*, 2021) trajectory prediction models to demonstrate the outlined characteristics' impact on predictive performance, and lastly we provide a comparison between the SDD and Intersection Drone (inD) Dataset. Our analysis of the SDD's key characteristics is important because without adequate information about available datasets a user's ability to select the most suitable dataset for their methods, to reproduce one another's results, and to interpret their own results are hindered. The observations we make through this analysis provide a readily accessible and interpretable source of information for those planning to use the SDD. Our intention is to increase the performance and reproducibility of methods applied to this dataset going forward, while also clearly detailing less obvious features of the dataset for new users.**

***Index Terms*—Autonomous vehicles, data analysis, information theoretic measures, pedestrian tracking, stanford drone dataset, trajectory.**

I. Introduction

SEVERAL datasets are available which provide annotated trajectory data of individuals navigating one or more scenes [1], [4]–[7]. Such datasets can be used as tools for an array of problems in Computer Vision and Intelligent Vehicles, including object detection [8], object tracking [9], and trajectory prediction [3], [10], [11]. A popular dataset that has been widely used for comparing recent benchmark methods is the Stanford Drone Dataset (SDD) [1]. The frequent use of the SDD stems in part from its size, the fact that it contains both vehicles and pedestrians in crowded scenes, and its previous use by benchmark methods. While many papers use the SDD to evaluate their performance, often the clear justifications for selecting it over alternative datasets is lacking.

The scarcity of information detailing the SDD may make it harder for users to properly determine if it's the best fit for their research. This scarcity is largely due to the fact that the dataset's accompanying publication and documents do not provide users with a sufficiently comprehensive description of the dataset, nor do the papers which utilize it.

One characteristic which is not addressed is how the videos within the SDD relate to one another in terms of time and location of recording. Another feature which plays a major role in determining the suitability of the SDD for a given model is the actual distribution and behavior of classes within the dataset compared to other, similar dataset options. Understanding differences in distribution and behavior is important when making an informed choice regarding which dataset is most suited to a particular application. Lastly, we discuss properties of the annotation data, including the impact of annotations labeled "lost" and split trajectories that effect tracking and prediction models of agents both directly and indirectly. For methods like trajectory prediction which rely on the annotated coordinates data, properly understanding these characteristics is necessary to ensure that the results are correct and easily reproducible.

We demonstrate each of these characteristics, and showcase how they may impact the accuracy of trajectory prediction applications. In order to compare them to a similar dataset we include the Intersection Drone (inD) dataset within the scope of our investigation. An example of the complexity that these characteristics add to the SDD can be seen in Fig. 1. The characteristics of importance demonstrated in this example are the way in which scenes are oriented relative to one another, as well as the behavior of "lost" annotations. Together, knowledge of these characteristics allows the user to determine the erroneous nature of the annotations in the provided example. This illustrates how understanding the characteristics of the dataset discussed here can provide a more comprehensive understanding of the SDD, and with it an improved ability to interpret observations made about the dataset.

In order to visualize and validate the outlined key characteristics, we implement the pre-existing PECNet [2] and Y-Net [3] trajectory prediction models, as well as a custom Adaptive Interaction Measure (AIM). AIM utilizes the information-theoretic measure of mutual information (MI) [12], which has been used in previous works involving tracking and trajectory prediction problems [13], [14]. We utilize AIM to visualize various dataset characteristics, and apply PECNet and Y-Net to the SDD to

Manuscript received 4 February 2022; revised 14 March 2022; accepted 27 March 2022. Date of publication 12 April 2022; date of current version 20 March 2023. This work was supported in part by NSF DMS under Grant 2053480. *(Corresponding author: Josh Andle.)*

The authors are with the School of Computing and Information Science, University of Maine, Orono, ME 04469 USA (e-mail: joshua.andle@maine.edu; nicholas.soucy@maine.edu; simonsocolow78@gmail.com; salimeh.yasaei@maine.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2022.3166642.

Digital Object Identifier 10.1109/TIV.2022.3166642

2379-8858 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

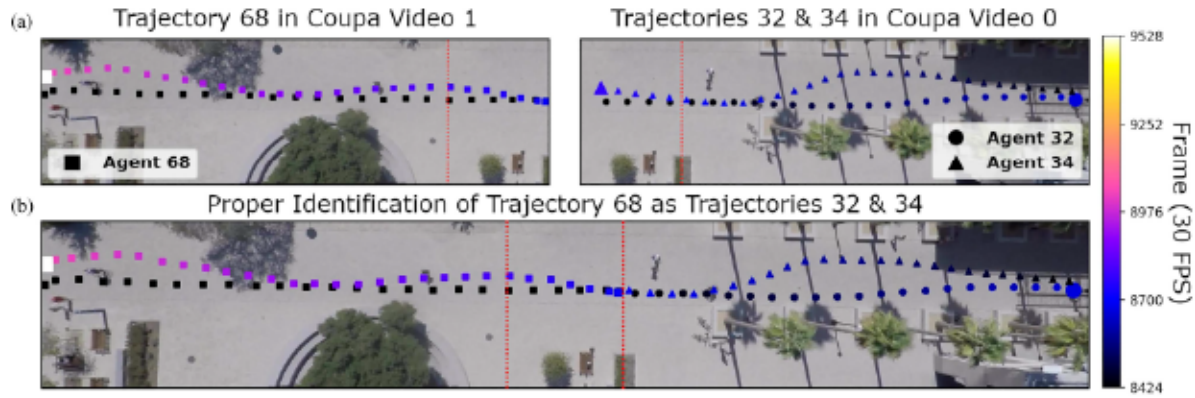

Fig. 1. (a) Three trajectories are shown from the SDD's Coupa videos 1 (left) and 0 (right). Annotations show agent 68 exiting the right side of video 1 and returning shortly after. The intermediate frames are labeled as "lost," indicating that the agent is out of the video's bounds. (b) Knowing that videos 0 and 1 are shot simultaneously and overlap in location, it becomes clear that agent 68 is actually a combination of agent 32 and 34's trajectories. Since agent 34 enters video 1 at the same location as agent 32 leaves, they are mistakenly annotated as the same individual in video 1 with their trajectories connected by "lost" annotations.

demonstrate some of those characteristics' impact on performance when ignored or accounted for. We chose to use these models due to the functional clarity, accurate performance, and readily available code.

This paper's aim is to highlight information about the SDD which is not readily apparent or discussed in previous work, as well as to demonstrate the importance of that information. Since there are several applications which the SDD can be used for, an analysis of exactly how those characteristics impact each possible application is outside of the scope of this paper. Instead we aim to utilize only as many experiments and methods as are necessary to provide compelling evidence of these characteristics' importance. Additionally, the custom AIM measure is intended as a metric to visualize and evaluate these characteristics. Analyzing and rigorously investigating the AIM measure is out of our scope in this research. In summary, our contributions in this paper are as follows:

1) We describe the "lost"-labeled annotations and split trajectories within the SDD and their importance. We demonstrate how properly handling these occurrences during preprocessing impacts the resulting model performance.
2) We provide a holistic view of how the videos within a given scene fit together. This includes cases where videos are recorded simultaneously, overlap in location, or both. We propose situations in which understanding how the videos relate to one another would impact the results obtained with the dataset.
3) We demonstrate differences in class distribution and behavior between the SDD and inD dataset which should be taken into consideration when deciding when to use one or the other.

To the best of our knowledge, there is no research which provides sufficient information on how the authors processed the SDD data and how the characteristics we discuss in this work might have impacted their results and model evaluation.

II. Related Works

Two datasets which have previously been used as benchmarks are ETH Zurich's BIWI Walking Pedestrian Dataset [4] and UCY's Crowds by Example [5]. These datasets are often used together as the ETH/UCY dataset. However, the small number of videos and lack of non-pedestrian individuals limit their utility, and differences in annotation formats between the two datasets could make using them together more troublesome for complex models.

These datasets have since been largely replaced in favor of the Stanford Drone Dataset (SDD) [1]. The SDD addresses several of the shortcomings of ETH/UCY, providing 60 videos split across 8 scenes and 6 classes of individuals, all with a single consistent annotation method. More recently the Intersection Drone (inD) dataset [6] has aimed to improve upon the SDD, providing 33 videos split across 4 intersections and 4 classes of individuals.

One prominent application for which the SDD has been used is trajectory prediction. Prominent examples of benchmark methods which use the SDD include Social-GAN [15], SoPhie [16], DESIRE [10], and CAR-Net [17]. PECNet [2] and Y-Net [3] are two other models which use the SDD and report improved performance accuracy over previous baselines.

A notable resource which has helped mitigate inconsistent use of the SDD is TrajNet [18], [19]. This benchmark provides a more uniform method of preprocessing and performing trajectory prediction on various datasets, including SDD. Although TrajNet provides a uniform method of processing the SDD it doesn't provide a detailed analysis of the dataset which would allow users to properly understand it. Additionally, users of the dataset who wish to preprocess the data differently may not be able to rely on TrajNet. For these reasons, we provide a direct analysis of the dataset's characteristics. We focus on providing an intuitive visualization of a set of characteristics which users should be aware of when using the SDD and demonstrating their importance.

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DATASET SPLIT USED FOR THE IND DATASET

Dataset	Videos
Train	0-4, 7-13, 18-25, 30
Validation	5, 14-15, 26-27, 31
Test	6, 16-17, 28-29, 32

III. METHODS OF ANALYSIS

In this section we introduce the methods used in our analysis of the SDD and inD dataset. One method we introduce is our information-theoretic metric to provide an interpretation of pairwise interactions between individuals, which we apply to visualize the characteristics of each dataset. Throughout the paper we use the notations below:

- X^I: Trajectory coordinates for agent I.
- $X^I_{t_1:t_2}$: Trajectory coordinates for agent I in time range $[t_1, t_2]$.
- X^I_t: Trajectory coordinates for agent I at time point t.
- δ: decay hyperparameter.
- T: Last time point of the interaction.
- T': A fixed number of time points after the beginning of the interaction.

A. Datasets

Stanford Drone Dataset (SDD): The SDD is a dataset providing birds-eye view drone recordings of 8 different scenes and 60 videos across a campus setting, with 6 annotated classes of individuals [1]. These classes are pedestrians, skateboarders, bicyclists, carts, cars, and buses. The annotated data contains bounding-box coordinates in pixel values at 30 frames per second, as well as labels indicating if a given coordinate was occluded, "lost" out of the video's bounds, or automatically interpolated.

Intersection Drone (inD) Dataset: The inD dataset also provides birds-eye view drone recordings of 4 intersections across 33 recordings [6]. Only the annotations are provided along with single-frame reference images, while the raw video footage isn't provided. The annotations contain 5 classes: pedestrians, bicyclists, cars, trucks, and buses, however the labels for trucks and buses are grouped as "truck_buses" within the annotations. The annotations provide numerous additional parameters, including bounding box coordinates in meters at 25 frames per second, the necessary values to convert from meters to pixels, the speed limit, time and geographical location, as well as the individuals heading, velocity, and acceleration.

In order to make direct comparisons to the SDD we have converted all inD annotations to pixels using the provided conversion values prior to analysis. For the SDD we use the training/testing split outlined in [18]. Table I outlines our split for the inD dataset, while later on in Section IV Table IV reports the class distribution for each dataset. For both datasets we preprocess the data to 2.5 fps and use the first 8 time-points of each trajectory for observation and the subsequent 12 time points for prediction, corresponding to 3.2 seconds and 4.8 seconds respectively.

B. AIM Definition

In our *Adaptive Interaction Measure* (AIM) we incorporate a physics-based weight function, ρ, as a scalar value to weight mutual information (MI) based interactions [20]–[22]. By summing this scaled information over the duration of a given trajectory pair, we get the cumulative measure, AIM, which describes the overall expected impact that one trajectory has on the other's navigation. The larger values of AIM suggest more significant interactions between agents. Our method is well-suited for this visualization task due in part to its transparent and intuitive incorporation of physical parameters in calculating ρ and subsequently AIM.

Suppose X^I and X^J are the 2-dimensional trajectory coordinates for agents I and J for each time point in their interaction. The interaction between agents I and J is defined to contain all frames in which both individuals are in the scene together. The AIM for an interaction between two agents I and J over the frames $t = T,' T' + 1, \ldots, T$, denoted by $AIM(X^I_{1:T}; X^J_{1:T})$ is defined as follows

$$\sum_{t=T'}^{T} \delta^{T-t} \rho(X^I_{t-N:t}, X^J_{t-N:t})\, MI(X^I_{1:t}; X^J_{1:t}), \qquad (1)$$

We calculate MI and ρ at each time point between T' and T and sum their product over all time points in the interaction. We provide the definitions of MI and ρ in 2 and 6 respectively. At each step in the summation a constant decay term, δ, is applied to the terms of each previous product. The AIM measure (1) shows the result of expanding this function out, in which the decay term has a stacking effect on time points the further they are into the past of the trajectory.

When $\delta = 1.0$, no decay is applied to the past time points in the calculation of AIM, making it possible for high values of AIM to reflect interactions that are no longer relevant to present navigation decisions. When $\delta < 1$, the value of AIM decreases at each time point. This allows more recent interactions to have a larger impact on the overall value of AIM, and in turn how an individual is expected to allocate their attention among neighboring individuals.

To calculate ρ at a given time point t, we consider the previous N time points. As N increases, the relative impact of replacing one frame at each new time point lessens, which has a smoothing effect on the calculated value of ρ. When N is too large this smoothing removes changes of ρ which reflect abrupt, temporary behaviors such as sudden stopping or breaking of a pedestrian or car. We demonstrate the effects of changing δ and N in Fig. 2. In order to ensure that there is sufficient past information to properly calculate MI and ρ we set aside the first T' frames of interaction as a buffer and begin the summation at $t = T'$.

1) Mutual Information: Let P_{XY} be a probability measure on the Euclidean space $\mathcal{X} \times \mathcal{Y}$. Here, P_X and P_Y define the marginal probability measures. The MI reflects the information that each of a pair of variables gives about the other, and is defined as

$$MI(X;Y) = \mathop{\mathbb{E}}_{P_X P_Y}\left[g\left(\frac{dP_{XY}}{dP_X P_Y}\right)\right], \quad g(t) = \frac{(t-1)^2}{2(t+1)}, \qquad (2)$$

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

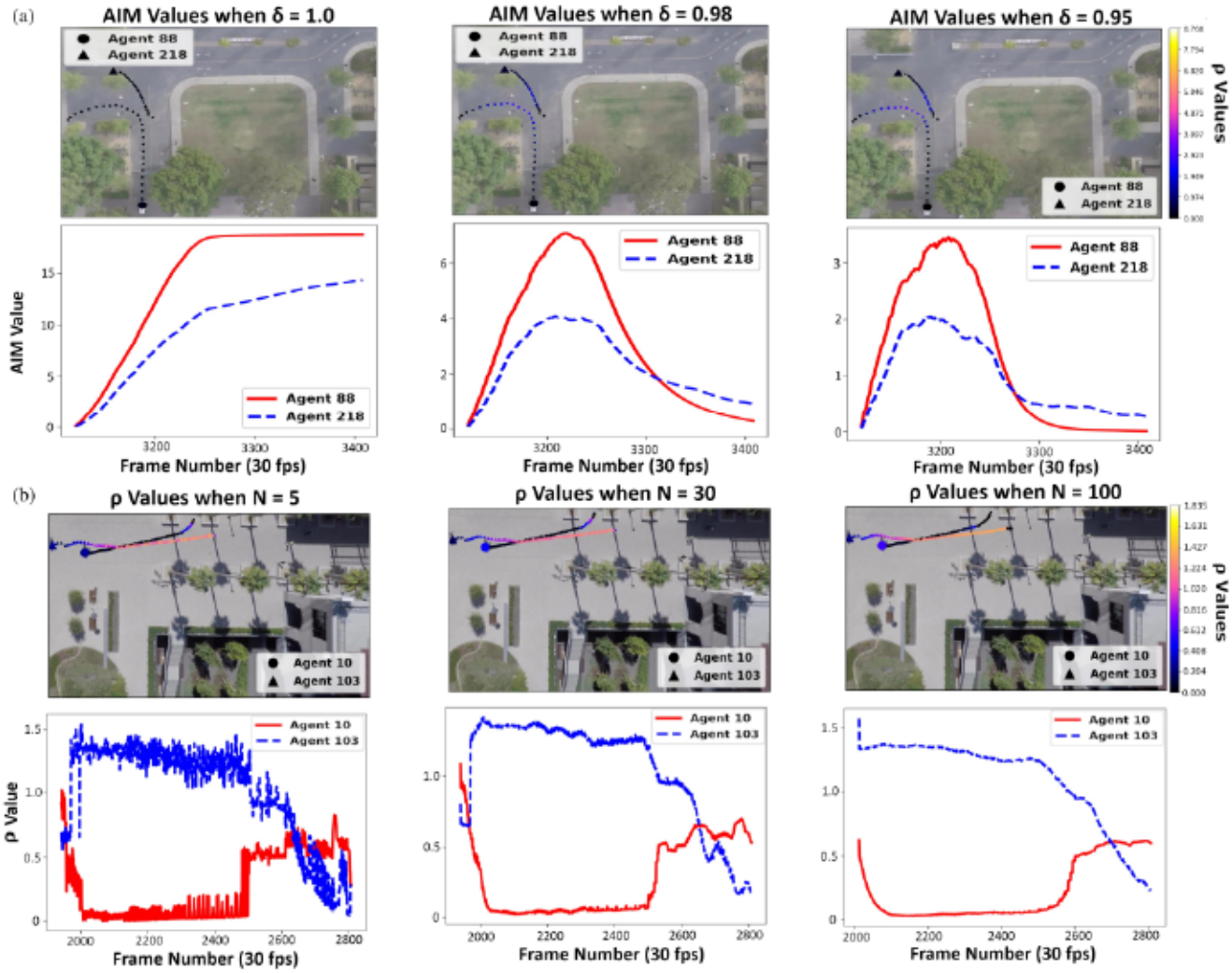

Fig. 2. (a) We show the different values of AIM obtained for a single interaction when $\delta = 1.0, 0.98$, and 0.95. When $\delta = 1.0$, AIM doesn't decay and the summation is monotonically increasing. Introducing 2% decay with $\delta = 0.98$ results in a constant decay of the value of AIM. The overall value of AIM still increases when newly added time points outweigh the decrease caused by decay. Further decreasing to $\delta = 0.95$ increases the rate of decay and observed AIM values, but doesn't substantially change behavior of AIM compared to $\delta = 0.98$. (b) When $N = 5$, ρ varies erratically since $\frac{1}{5}$ of the input values are changed at each new time point. For $N = 30$, only $\frac{1}{30}$ of the inputs change at each time point, which reduces the rapid changes seen when $N = 5$. When $N = 100$, the smoothing fails to capture the brief changes that occur in the pedestrians' trajectories after frame 2400.

where $\frac{dP_{XY}}{dP_X P_Y}$ is the Radon-Nikodym derivative. In this case, the variables are the coordinates along the trajectories of two individuals. High values of MI typically occur when agents are stationary or have been moving in a constant manner. MI does not noticeably change when the individuals are near or far apart, or based on the directions the individuals are moving in. To account for this we include our physics-based weighting measure ρ when deriving AIM.

Examples of MI estimators include Kraskov Stogbauer Grassberger (KSG) [23], Kernel Density Estimation (KDE) [24], Nearest Neighbor Ratio (NNR) [25], and Minimal Spanning Tree (MST) [26]. The MI estimation process is computationally intensive, e.g. the complexity of the KDE is $O(n^2)$, while the KSG takes $O(kn\log(n))$ (k is a parameter). In this paper, we use a hash-based MI estimator called the ensemble dependency graph estimator (EDGE) [27] due to its linear complexity and optimal mean squared error convergence rate.

1) Weight Function ρ: We utilize the handcrafted function ρ in (1) to provide a contextual interpretation of MI. We selected potential parameters of this function with the goal of reflecting the likelihood and severity of potential collisions between individuals. This is intended to approximate how attention is allocated by a navigating individual (e.g. avoiding oncoming cars, letting others pass on the sidewalk, and stopping at a crosswalk when people are crossing). For the function's parameters, we use the *velocity* (V), *distance* (D), and the relative *heading* (H) of the two individuals. V is the sum of both agents' velocities averaged over N frames, D is the average distance between the two agents, and H is asymmetric and is calculated as the average angle between an individual's current heading and the other individual's current position. These parameters are calculated

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

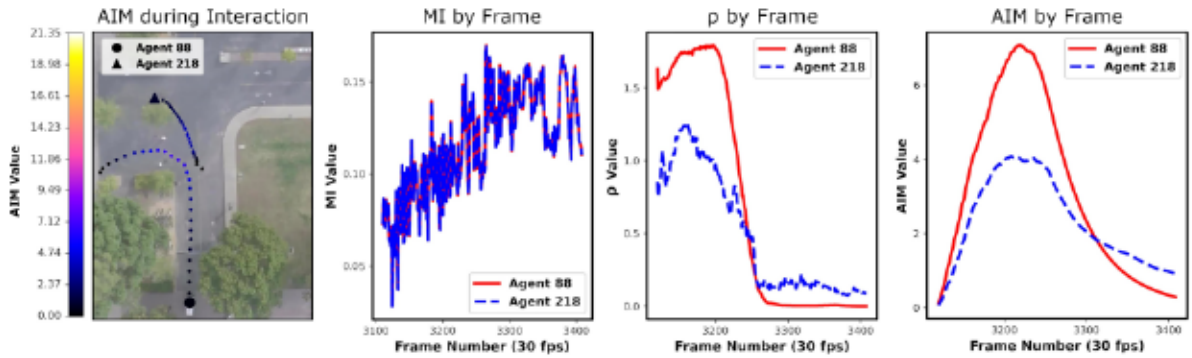

Fig. 3. For this example, we track MI, ρ, and AIM values for each frame in the trajectory pair. The car (agent 88) shows a higher initial ρ value than the pedestrian, which is reflected by the fact that its AIM increases faster than the pedestrian's. Once the car has passed by the pedestrian, both ρ values decrease and as a result AIM begins to decrease. This demonstrates that the expected attention is temporary in the absence of further relevant interactions.

as follows:

$$V_t = \frac{1}{N}\sum_{n=t-N}^{t}\left(\sqrt{\|X_n^I - X_{n-1}^I\|^2} + \sqrt{\|X_n^J - X_{n-1}^J\|^2}\right), \tag{3}$$

$$D_t = \frac{1}{N}\sum_{n=t-N}^{t}\sqrt{\|X_n^I - X_n^J\|^2}, \tag{4}$$

and H_t is defined as

$$\frac{1}{N}\sum_{n=t-N}^{t}\left|arctan\left(\frac{\left\|(X_{n-1}^J - X_{n-1}^I)\times(X_n^I - X_{n-1}^I)\right\|}{<(X_{n-1}^J - X_{n-1}^I),(X_n^I - X_{n-1}^I)>}\right)\right|, \tag{5}$$

where $\|.\|$ is Euclidean distance, the operation $\times$ is the cross product, and $<.,.>$ is the dot product. Prior to calculating ρ, V_t, D_t, and H_t are normalized denoted by V_{t^*}, D_{t^*}, and H_{t^*} respectively such that V_{t^*}, $D_{t^*} \in (0,1)$ while $H_{t^*} \in (-1,1)$. The scale function which was used for ρ is

$$\rho(X_{t-N:t}^I, X_{t-N:t}^J) := (\alpha + V_{t^*}) \cdot (D_{t^*}(1 + H_{t^*})). \tag{6}$$

The constant α in (6) ensures that when $V = 0$ then ρ can still have a non-zero value based on D_{t^*} and H_{t^*}; and when V increases, so does ρ. By using $(D_{t^*}(1 + H_{t^*}))$, when one individual is moving toward the other's position ρ increases. This reflects an increased likelihood of collision as the individuals move nearer to each other but also gives more attention to individual J when they are in front of agent I, reflecting that agent J would be readily visible to agent I compared to neighbors outside of their field of view (assuming individuals look in the direction they're moving).

In Fig. 4, we demonstrate the impact of removing different parameters from the ρ function. To do this, we compare two similar interactions between agents 80 (pedestrian), 71 (pedestrian), and 72 (bicyclist) as well as a third interaction between agents 88 (car) and 218 (pedestrian). We show that removing the velocity, distance, or heading parameters negatively impacts the intuitive interpretability of values of ρ calculated for the interactions.

C. Experimental Models

For our experiments we use both the PECNet [2] and Y-Net [3] models. PECNet is a trajectory prediction model which implements a variational auto-encoder and social pooling to predict trajectory endpoints. PECNet encodes the observed portions of trajectories and their ground-truth endpoints. This encoded information is used to predict end-points for each trajectory. The K closest predictions to the goal are then used along with the observed portion of the trajectory to estimate the future coordinates of the trajectory. We use the author's provided code. Preprocessing was done directly with the SDD's annotations, using the PECNet model. Y-Net instead utilizes a set of subnetworks similar to U-net [28] to handle both the uncertainty in an agent's goal, as well as the uncertainty in how they reach a given goal. Encoding is performed for the past trajectory data and scene segmentation information, prior to predicting the future trajectory.

D. AIM Implementation

In this section we outline our hyperparameter settings for AIM and demonstrate how it can be applied to interaction data and interpreted.

AIM Function: In (1), we set hyperparameter $\delta = 0.98$, resulting in a decay of 2% at each time point.

ρ Function: In our ρ function (6) we set $\alpha = 0.3$. We set $N = 30$ for SDD and $N = 25$ for inD. These values were selected based on the two dataset's differing frame rates, such that ρ covers the past one second of trajectory data for each dataset. Each normalizing function for the parameters of ρ was tuned with respect to the values of their parameter within the SDD.

We use MI, ρ, and AIM to visualize the characteristics of the SDD and inD dataset. Through the rest of the paper we present these measures as heatmaps overlaying the corresponding trajectories. The color of a given individual's trajectory indicates their value of the specified measure at that time point. The associated color bar reflects the corresponding values of the given measure among all individuals in a given recording. Accompanying these heatmaps are graphs tracking the values of the measure over the course of the trajectory for both interacting individuals. Fig. 3 demonstrates this format while giving an example of how MI,

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

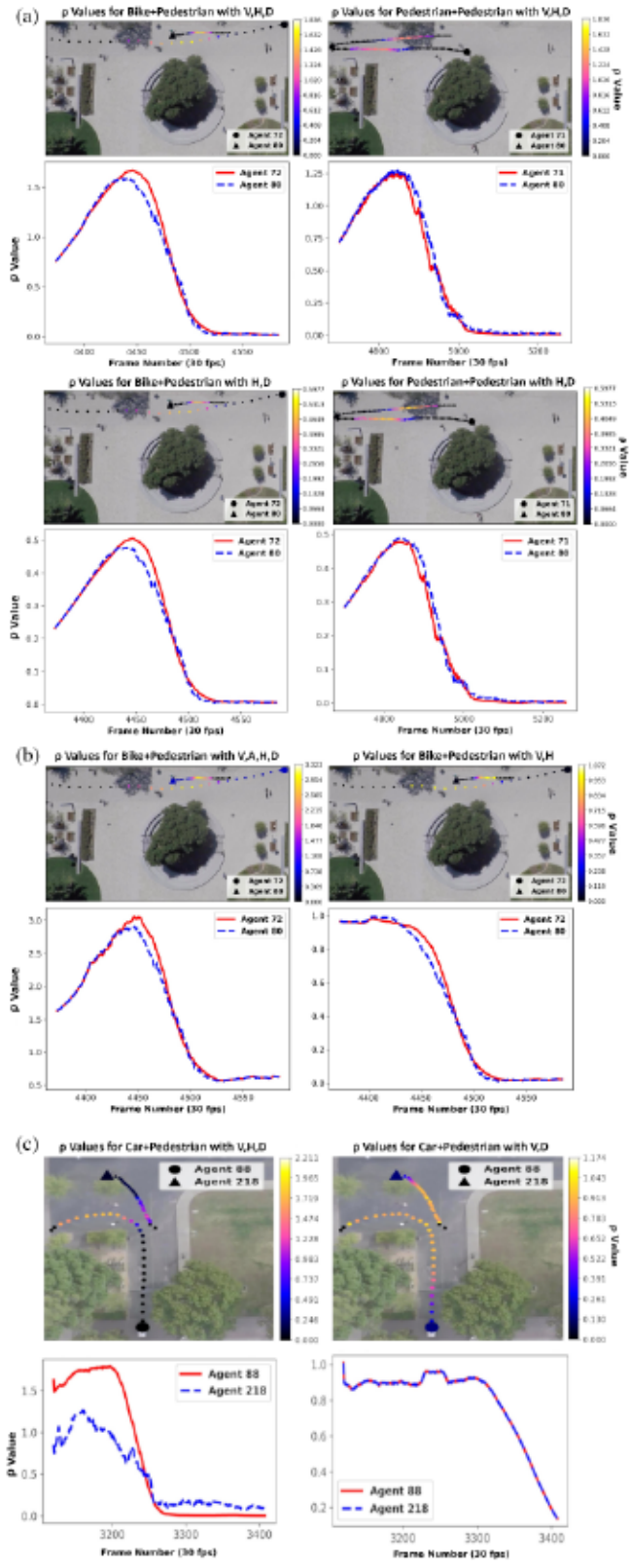

Fig. 4. (a) Using *velocity* V to calculate ρ results in higher values for the interaction with agent 72, as they are moving faster than agent 71. Removing V produces nearly the same values for both interactions, which is a less intuitive result. (b) Introducing *acceleration* (A) does not significantly affect the behavior of ρ, while removing D provides the same ρ values regardless of distance, which is counter-intuitive for an attention measure. (c) Removing *heading* H from ρ results in agents paying equal attention when either in front of or behind one another.

TABLE II
TOTAL NUMBER OF TRAJECTORIES AND FREQUENCY OF LOST ANNOTATIONS IN THE SDD

Scene	Number of Trajectories	Start(%)	Middle(%)	End(%)
Bookstore	1645	78.1	7.4	65.2
Coupa	425	74.4	5.6	69.2
DeathCircle	2830	69.3	8.1	53.8
Gates	1249	68.9	10.6	54.8
Hyang	1980	68.7	9.9	56.2
Little	656	90.9	5.8	84.0
Nexus	1456	68.1	7.6	46.8
Quad	59	25.4	13.6	16.9

ρ, and AIM vary over the course of a pair of trajectories. The primary role of these measures within this paper is to provide readily interpretable methods of visualizing the characteristics of the SDD and inD datasets, rather than as tools for prediction, tracking or detection.

IV. KEY CHARACTERISTICS

A. Annotations

The SDD contains a label for frames that are "lost" out of the bounds of the video. This description alone doesn't give an adequate description of these annotations. When plotted, the actual behavior of these points becomes more apparent. Lost coordinates often occur either at the start or end of a trajectory, and less often in the middle for those agents that leave and reenter a scene. These annotations either remain stationary at the point where the individual will enter or leave the scene, or alternatively move linearly back into the scene. Fig. 5 uses MI and ρ to visualize some of the behavior and effects of these lost coordinates.

We report the frequency of trajectories containing these annotations in Table II. Additionally, we distinguish between how frequently they occur at the start of the trajectory, the end, or in the middle. This is important because simple approaches of filtering may properly remove the annotations at the start and end of the trajectory, but may have different effects on the middle occurrences. Some possible effects may be to leave a gap in the trajectory data with no coordinate data for those frames, or to split the trajectory into multiple new trajectories before and after the lost annotations occurred. For this reason it is important for users of the dataset to describe how they process the data they use. Omitting this information leads to difficulties when reproducing other researchers' work without explicit instructions on how they addressed such decisions. For this work we filter out the lost annotations, and then if this splits the trajectory we keep only the first portion.

This characteristic is of particular importance for researchers using the SDD for trajectory prediction, as many benchmark methods take ~ 3.2 seconds of each trajectory as their observed portion of the trajectory and then attempt to predict the next ~ 4.8 seconds [2], [10], [16]. If the lost coordinates are left included at the start of the trajectories, then this could lead to a large bias towards observing and predicting stationary trajectories.

[10], [29] are the only works in which the lost coordinates are briefly mentioned, however the authors do not describe them. To demonstrate the importance of properly filtering these

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

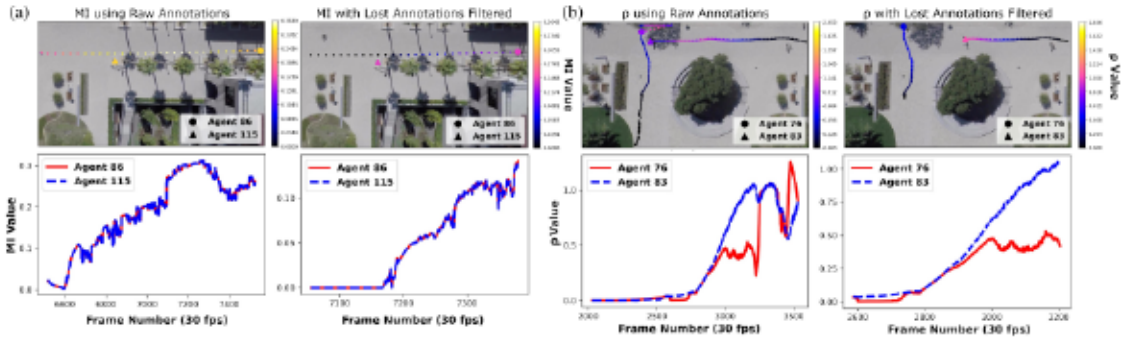

Fig. 5. (a) The graphs reflect that nearly half of the trajectory's frames are removed when filtering "lost" annotations. As MI tends to increase during stationary trajectories, the additional frames inflate its value. (b) Here the lost annotations wrongly suggest that the individual who has left the scene instead reenters and begins converging with the other individual. A predictive model may incorrectly predict that agent 83 will move to avoid a collision, when in fact there's no one else around them.

TABLE III
TESTING ACCURACY ON THE IND DATASET UNDER DIFFERENT TRAINING CONDITIONS

Model	Dataset	All		Pedestrians		Bicycles		Cars	
		ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE
PECNet	SDD w/ Lost	21.82	35.26	7.77	13.95	26.87	44.11	83.25	140.32
	SDD w/o Lost	20.18	35.08	7.16	12.46	21.8	33.11	65.38	112.33
	inD	12.32	20.98	6.74	12.54	14.02	24.28	19.35	35.25
Y-Net	SDD w/ Lost	14.22	22.61	8.76	13.51	20.80	29.00	72.10	64.55
	SDD w/o Lost	13.84	22.28	7.49	12.37	19.61	26.80	82.75	41.96
	inD	6.16	10.23	3.48	5.67	7.58	12.26	7.36	12.23

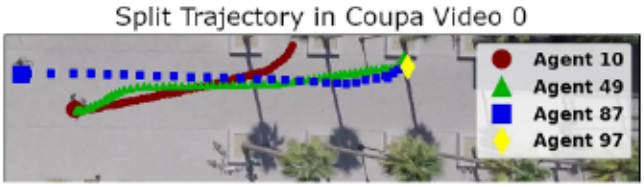

Fig. 6. In this example four unique track IDs from video 0 in the Coupa scene of the SDD are shown. The provided video and annotations confirm that these four tracks belong to a single individual.

coordinates, we show how removing them can improve a trained trajectory prediction model's ability to generalize to a new dataset in Table III. This table reports the testing accuracy on the inD dataset using models trained either on the inD dataset, or on the SDD dataset with or without the lost annotations filtered out.

These results show that when training on the SDD and testing on the inD dataset, training on data which has the lost annotations filtered out (SDD w/o lost) improves both of the resulting performance measures (we use the standard trajectory prediction measures of Average Displacement Error (ADE) and Final Displacement Error (FDE) [2], [16], [19]) compared to including them (SDD w/ lost). We show that this remains true when training/testing on all classes, pedestrians, bicycles, or cars, for both the PECNet and Y-Net models.

In addition to lost annotations, another behavior of the trajectory data in the SDD is that multiple trajectory IDs may correspond to a single individual. We demonstrate a prominent example of this behavior in Fig. 6. In these cases the individuals full trajectory is split into multiple partial trajectories, each of

TABLE IV
CLASS DISTRIBUTION IN THE SDD (TOP) AND THE UNIQUE INTERSECTIONS OF THE IND DATASET (BOTTOM)

Dataset	Scene	Pedestrians	Bicyclists	Cars	Buses	Skaters	Carts
SDD	Bookstore	63.9	32.9	0.83	0.37	1.63	0.34
	Coupa	80.6	18.9	0.17	0	0.17	0.17
	DeathCircle	33.1	56.3	4.71	0.42	2.33	3.1
	Gates	43.3	51.9	1.08	0.78	2.55	0.29
	Hyang	70.0	27.7	0.5	0.09	1.29	0.43
	Little	42.5	56.0	0.17	0.67	0.67	0
	Nexus	64.0	4.22	29.5	1.25	0.6	0.4
	Quad	87.5	12.5	0	0	0	0
inD	0-6	6.95	3.66	79.9	9.5	0	0
	7-17	21.4	11.6	65.4	1.58	0	0
	18-29	33.7	27.3	38.8	0.26	0	0
	30-32	3.44	3.05	88.9	4.61	0	0

which is given a different ID. Unlike with lost annotations, there is no clear indicator of when this occurs, and confirming it requires manual cross-checking of each partial trajectory's annotation.

B. Data Diversity and Scene Feature Adherence

Previous works have commented on the SDD's class [7] and scene [30] diversity. Vehicles make up a significant minority, and many of them are parked. In comparison, the inD dataset contains a larger percentage of moving cars but fewer pedestrians and classes.This suggests that the SDD may be more suited for applications that focus on pedestrians [3], [31] while the inD dataset is suited for models intended for use in environments with both pedestrians and cars [32] This idea is further supported by differences in scene diversity and navigational behavior.

We demonstrate the percent distribution of classes in each dataset in Table IV. The values for the SDD are provided by the datasets authors, while the values for inD were determined by checking the number of occurrences of each class in the metadata files and summing across all videos in a given intersection. These values show the prominence of foot traffic in the SDD, and of vehicles in the inD dataset.

While the SDD has multiple scenes, all of these scenes are located in a campus setting where the predominance of foot traffic impacts navigation behavior. This behavior can be seen in Fig. 7, which shows the frequency at which individuals in the SDD walk or bike through the streets rather than sidewalks or

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

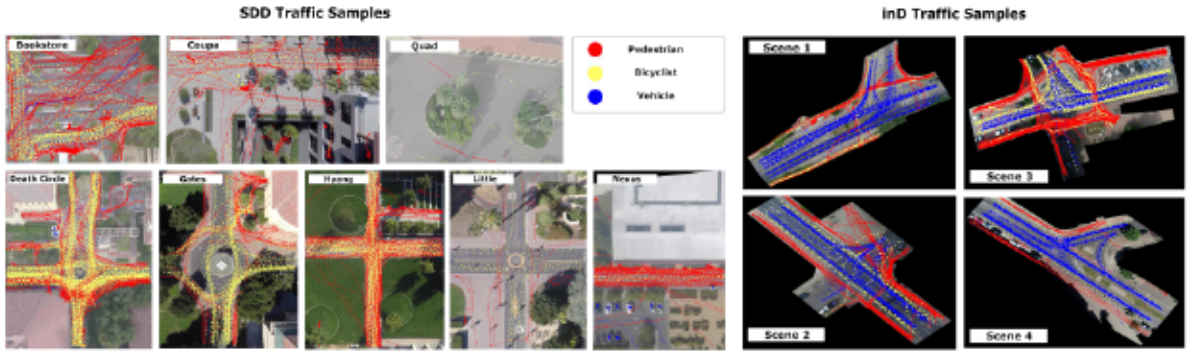

Fig. 7. The trajectories from a single video in each scene within the SDD and inD dataset are plotted. The behavior of foot traffic varies from scene to scene within the SDD. The pedestrians and bicyclists in the SDD often enter or cross the road, and vehicles are infrequent. In the inD dataset there are numerous vehicles, so bicyclists primarily remain on the shoulders of the road and pedestrians more consistently use crosswalks to cross.

shoulders of the road. This indicates that certain scene features such as sidewalks and roads have a less predictable impact on navigation than when vehicles are more prevalent (as in the inD dataset). As such, the SDD may not be well suited to methods which learn to rely on such features.

In contrast to the SDD, the inD dataset has scenes located at four intersections in public roads. As with the SDD, there are some behaviors typical of intersections which are readily apparent in this dataset. For instance, cars are much more prone to stopping and turning than they would be in the absence of stop lights or intersections. Similarly, compared to areas such as the campus setting of SDD, cars are more prominent in the dataset, and as a result foot traffic has to adhere more closely to certain scene features than in SDD, as in Fig. 7. For this reason, it's possible that the inD dataset is less suited for models which do not account for scene features, as these features add a constraint to navigational behavior.

An impact of the different behaviors and class distributions in both datasets can be seen in Table III. PECNet performed similarly well for pedestrians in the inD dataset whether it was trained on the SDD or the inD dataset, however the performance for cars is significantly worse when trained on the SDD, indicating that the cars in the SDD are insufficient for predicting car behavior in the inD dataset. This is possibly because many of the cars in the SDD are parked, or their driving is otherwise different due to the presence of more pedestrians in the road and lack of intersections with stoplights. We observe similar results with Y-Net where the gap in performance between pedestrians when trained on the SDD or inD dataset is much smaller than for cars.

C. Scene Layout

For a given scene in the SDD, the region covered by each video may overlap completely, partially, or not at all. Similarly, the time of recording may be the same or different. An example in which both the location and time overlap is shown in Fig. 8.

When both time and location overlap, interactions that occur within the overlapping regions appear in multiple videos. This leads to redundancy in the observed interactions depending on the methods used. Fig. 9 uses MI and ρ to demonstrate how this

Fig. 8. Here the same frame is taken from each video in the Coupa scene. The overlap in location is reflected by dashed lines, and the simultaneous recording time can be seen from the four pedestrians who are in the same locations in each video.

TABLE V
OUTLINE OF OVERLAPPING AND SPLIT VIDEOS IN THE SDD

Overlap	Bookstore	Coupa	DeathCircle	Gates	Hyang	Little	Nexus	Quad
Location	Partial	Partial	Full	Partial	Partial	Partial	Partial	Partial
Time	Partial	Full	None	Partial	Partial	Partial	Partial	Full
Split Videos	1-6	1-4	None	0-2	6,10-14	1-3	0-2	0-3
				4-7	7-9		3-5	
				5,6	2,3		6-8	
							9-11	

overlap may affect different measures. The MI values within the overlapping regions differ between the two videos, while the values of ρ are highly similar.

The discrepancy between MI values is explained since MI relies on the full past of the trajectories, leading to higher MI values in video 1 during the overlapping frames as there is more past trajectory data providing information about the current interaction. Since ρ only relies on the past one second of information, the interactions result in nearly identical values of ρ. This demonstrates how measures such as our ρ, which care only about local information, lead to redundant values in overlapping sections of different videos.

Several videos in the SDD overlap both in location and recording time, which may influence how certain applications such as object tracking or trajectory prediction are used on the dataset (e.g. users may assess the accuracy and consistency of tracking a single individual across two videos, while trajectory

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

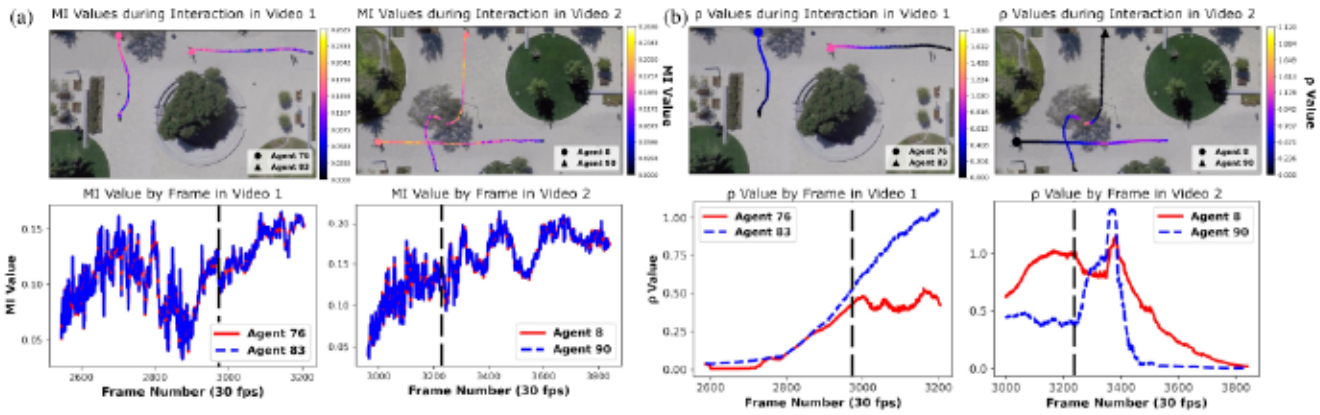

Fig. 9. The trajectories of the same two individuals are plotted from two videos in the SDD's Coupa scene. A portion of this interaction is included in both videos approximately between frames 3000 and 3200. MI takes different values over these frames for each video. In contrast, the ρ values which only rely on the most recent 30 frames at each time point show nearly identical values for each video between these overlapping frames. This demonstrates how the video overlaps may impact results differently depending on the methods being used.

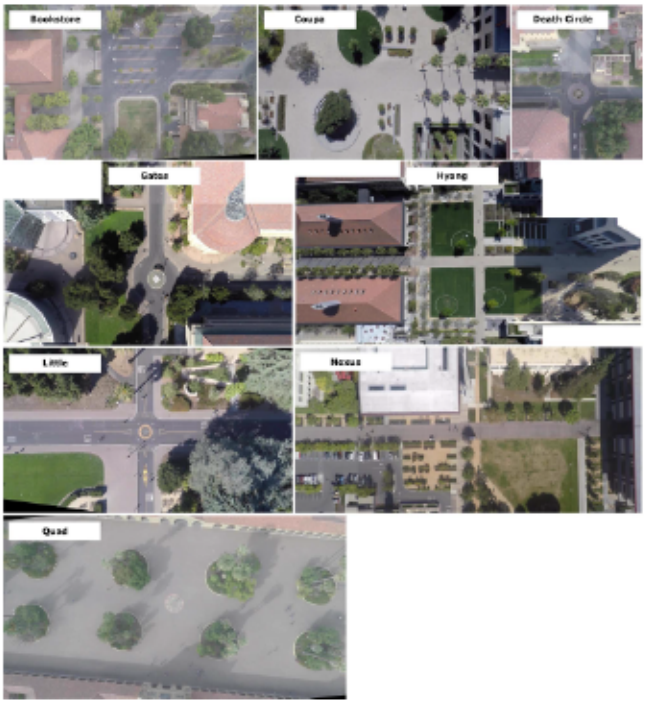

Fig. 10. Here we provide composite images of each scene as an aid for users of the SDD to more easily understand where each video is in relation to the others within a given scene.

prediction models may need to split their training and test data accordingly).

When only the time of recording is the same, the effects are less predictable. If the scenes are nearby (for instance within a college campus), then behaviors associated with a given time would affect multiple recordings such as rush hour traffic or students going to class. In the inD dataset, all recordings of a given intersection record the same location at different times. Because of this, the features within that location are present in each of the recordings belonging to that scene. For features that significantly impact behavior, such as bus stops or benches, this redundancy could disproportionately emphasize those features.

The groups of videos in each scene which overlap in time or location for each scene in the SDD are listed in Table V. In this table, "partial" overlap indicates that only a subset of the videos within the scene overlap either in time or location. We list which videos were shot simultaneously in the "Split Videos" data. The cause for such videos appears to be that certain videos are shot over a larger area prior to being broken into multiple videos, each covering a smaller subset of the area from the original recording. This is most clear with Coupa as seen in Fig. 8. We additionally provide the composite images of each scene in Fig. 10 for reference. These composite images allow users of the dataset to better visualize how individual videos relate to one another within a scene to identify where these characteristic overlaps occur.

V. Discussion and Conclusion

In this work, we have demonstrated several key characteristics of the SDD which have not been properly addressed before. These characteristics are: the properties of the lost annotations, differences in class diversity and behavior compared with the inD dataset, and the scene layout. We have provided clear and intuitive visualizations and descriptions of each, including the potential impacts they have in different applications.

For those that have previously used the SDD without accounting for these characteristics, it is possible that the accuracy of their model was worsened by erroneous preprocessing or because the SDD was not well suited to their methods. In this paper, we have shown the potential for the former case by leaving the lost annotations in during preprocessing for the trajectory prediction models PECNet and Y-Net (Table III), which resulted in a worsened ability of the networks to generalize to the inD dataset when trained on the SDD compared to models trained without the lost annotations being included. We have also provided a comparison to the inD dataset to help future researchers make informed decisions regarding which data best fits their

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

methods. We believe that these contributions will not only make it easier to work with the SDD, but also to do so properly while being able to better interpret the results obtained regardless of the application.

Understanding the properties of the SDD's annotated data is vital when using it, as detailing how your data is preprocessed (whether or not you have removed lost annotations, how you handle trajectories which become split as a result, etc), significantly improves others' ability to reproduce your results. Despite this, information surrounding these steps is frequently left out of publications. While the presence of split trajectories is worth noting and more difficult to identify, we suspect that it is also less impactful than the lost coordinates. Any model which is not expected to recognize that all four partial trajectories in Fig. 6 belong to the same pedestrian (whose class changes from pedestrian to biker mid-trajectory) may train just as well by treating them as four unique individuals with non-overlapping frames.

The differences in classes between the two datasets has implications for when each dataset is best suited for a given application. Agents navigating the SDD are less adherent to scene features such as sidewalks, and there are far fewer moving cars. This suggests that the SDD would be well suited for applications intended for use in areas such as parks or shopping centers, where pedestrians make up most of the traffic and are less observant of scene features. By contrast, the inD dataset may be better suited for street traffic where the prevalence of moving cars forces foot-traffic to more strictly adhere to sidewalks.

Lastly our demonstration of the overlap between videos in the SDD is useful for applications which can take advantage of this characteristic. An example of this is for tracking, where the overlapping locations could be used to test a method's ability to track an individual's trajectory across multiple adjacent videos. This information can be used by those looking to utilize different training/testing splits for the SDD data. Most trajectory prediction papers using the SDD utilize the Trajnet split outlined in [18]. For the most part this split accounts for overlapping scene locations and simultaneous recordings, however we provide information in Table V regarding which scenes contain overlapping locations or are recorded simultaneously in case other users of the dataset want to investigate alternative splits. This information is supplemented by the composite images of each scene provided in Fig. 10, which as best we could determine have yet to be published or provided anywhere. While simple, this information is not readily clear when using the SDD, and we've included it in keeping with this work's goal of improving the usability of the SDD.

Acknowledgment

The findings are those of the authors only and do not represent any position of funding bodies.

References

[1] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.

[2] K. Mangalam *et al.*, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 759–776.

[3] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15 233–15 242.

[4] S. Pellegrini, A. Ess, K. Schindler, and L. V. Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 261–268.

[5] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer Graphics Forum*, vol. 26, Hoboken, NJ, USA: Wiley, 2007, pp. 655–664.

[6] 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," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2020, pp. 1929–1934.

[7] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems," in *Proc. 21st Int. Conf. Intell. Transp. Syst.*, 2018, pp. 2118–2125.

[8] J. Yang, X. Xie, and W. Yang, "Effective contexts for UAV vehicle detection," *IEEE Access*, vol. 7, pp. 85 042–85 054, 2019.

[9] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 300–311.

[10] N. Lee *et al.*, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 336–345.

[11] H. Zhao *et al.*, "TNT: Target-driven trajectory prediction," in *Proc. Conf. Robot Learn.*, 2021, pp. 895–904.

[12] T. Cover and Joy A. Thomas, *Elements of Information Theory*, 1st ed. Chichester, U.K.: Wiley, 1991.

[13] M. Xu, Y. Wu, P. Lv, H. Jiang, M. Luo, and Y. Ye, "miSFM: On combination of mutual information and social force model towards simulating crowd evacuation," *Neurocomputing*, vol. 168, pp. 529–537, 2015.

[14] Y. Zhang, Z. Chen, C. Wu, J. Jiang, and B. Ran, "Vehicle trajectory analysis system via mutual information and sparse reconstruction," *Transp. Res. Rec.*, vol. 2645, no. 1, pp. 195–202, 2017.

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

[16] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1349–1358.

[17] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese, "Car-Net: Clairvoyant attentive recurrent network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 151–167.

[18] A. Sadeghian, V. Kosaraju, A. Gupta, S. Savarese, and A. Alahi, "TrajNet: Towards a benchmark for human trajectory prediction," 2018.

[19] P. Kothari, S. Kreiss, and A. Alahi, "Human trajectory forecasting in crowds: A deep learning perspective," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2021.3069362.

[20] I. Stuhl, M. Kelbert Y. Suhov, and S. Yasaei Sekeh, "Weighted Gaussian entropy and determinant inequalities," *Aequationes Mathematicae*, vol. 96, pp. 85–114, 2022.

[21] Y. Suhov, I. Stuhl, Y. Salimeh Sekeh, and M. Kelbert, "Basic inequalities for weighted entropies," *Aequationes Mathematicae*, vol. 90, no. 4, pp. 817–848, 2016.

[22] B. Oselio, A. Hero, A. Sadeghian, and S. Savarese, "Time-varying interaction estimation using ensemble methods," in *Proc. IEEE Data Sci. Workshop*, 2019, pp. 69–75.

[23] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, 2004, Art. no. 066138.

[24] Young-Il Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Phys. Rev. E*, vol. 52, no. 3, 1995, Art. no. 2318.

[25] M. Noshad, K. R. Moon, Y. Salimeh Sekeh, and A. O. Hero, "Direct estimation of information divergence using nearest neighbor ratios," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017, pp. 903–907.

[26] Y. Salimeh Sekeh and Alfred O. Hero, "Geometric estimation of multivariate dependency," *Entropy*, vol. 21, no. 8, 2019, Art. no. 787.

[27] M. Noshad and A. Hero, "Scalable hash-based estimation of divergence measures," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1877–1885.

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.
[29] D. Ridel, N. Deo, D. Wolf, and M. Trivedi, "Scene compliant trajectory forecast with agent-centric spatio-temporal grids," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 2816–2823, Apr. 2020.
[30] R. Krajewski, T. Moers, J. Bock, L. Vater, and L. Eckstein, "The round dataset: A drone dataset of road user trajectories at roundabouts in Germany," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–6.
[31] P. Dendorfer, A. Osep, and L. Leal-Taixé, "Goal-GAN: Multimodal trajectory prediction based on goal position estimation," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 405–420.
[32] H. Cheng, W. Liao, X. Tang, M. Y. Yang, M. Sester, and B. Rosenhahn, "Exploring dynamic context for multi-path trajectory prediction," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 12 795–12 801.

Josh Andle received the B.S. degree in biochemistry from the University of Maine, Orono, ME, USA, in 2017. After two years working as a Research Assistant with the Joslin Diabetes Center in Boston, Massachusetts, he returned to the University of Maine to pursue the Ph.D. degree in computer science. His research interests include machine learning, artificial intelligence, novel neural network designs, and the various applications of machine learning.

Nicholas Soucy received the B.A. degree in physics in 2020 from the University of Maine, Orono, ME, USA, where he is currently working toward the M.Sc. degree in computer science. His concentration is in Artificial Intelligence and Machine Learning. His research interest include semantic segmentation, classification, intelligent systems, image recognition, and intelligent control.

Simon Socolow is currently working toward the graduation degree with Bangor High School, Bangor, ME, USA. He was an Intern with the University of Maine, Orono, ME, USA, Sekeh AI Laboratory, from June 2020 to June 2021. His research interests include artificial intelligence, blockchains, and cryptography.

Salameh Yasaei Sekeh (Member, IEEE) received the Ph.D. degree in inferential statistics from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2013. She is currently an Assistant Professor of computer science with the School of Computing and Information Science (SCIS), University of Maine, Orono, ME, USA. Prior to University of Maine, she was a Postdoctoral Research Fellow with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI, USA, working with Alfred O. Hero. She held CAPES-PNPD funder Postdoctoral Fellow appointment with the Federal University of Sao Carlos (UFSCar), Sao Carlos, Brazil, in 2014 and 2015, and she was a Visiting Scholar with the Polytechnic University of Turin, Turin, Italy, between 2011 and 2013. She is the Director of Sekeh Laboratory. Her recent research interests include machine learning, large scale data science, and statistical signal processing. Her primary research interests include design, improvement, and analysis of deep learning techniques with emphasis on deep network compression, multi-class classification problems, graph-based learning, data mining, high-dimensional network structure learning, practical applications of machine learning in real-time problems, and network interaction analysis.

Authorized licensed use limited to: University of Maine. Downloaded on July 01,2023 at 02:19:21 UTC from IEEE Xplore. Restrictions apply.