Learning a Pedestrian Social Behavior Dictionary

Faith Johnson*
faith.johnson@rutgers.edu
Kristin Dana
kristin.dana@rutgers.edu

Electrical and Computer Engineering Department
Rutgers University
New Brunswick, NJ, USA

Abstract

Understanding pedestrian behavior patterns is key for building autonomous agents that can navigate among humans. We seek a learned dictionary of pedestrian behavior to obtain a semantic description of pedestrian trajectories. Supervised methods for dictionary learning are often impractical since pedestrian behaviors may be unknown a priori and manually generating behavior labels is prohibitively time consuming. We utilize a novel, unsupervised framework to create a taxonomy of pedestrian behavior observed in a specific space. First, we learn a trajectory latent space that enables unsupervised clustering to create an interpretable pedestrian behavior dictionary. Then, we show the utility of this dictionary for building pedestrian behavior maps to visualize space usage patterns and for computing distributions of behaviors in a space. We demonstrate a simple but effective trajectory prediction by conditioning on these behavior labels. While many trajectory analysis methods rely on RNNs or transformers, we develop a lightweight, low-parameter approach and show results outperforming SOTA on the ETH and UCY datasets.

1 Introduction

The success of computer vision in robust recognition has paved the way for vision-guided autonomous agents in real world environments. For embodied agents, an understanding of pedestrian behavior is important for navigating in a non-disruptive manner. Trajectory prediction of pedestrians has received significant attention in recent years, with algorithms that learn time-series representations of trajectories while taking into account both scene context and other nearby pedestrians. While trajectory prediction algorithms are powerful, they output sequences of x-y coordinates which do not provide high level, interpretable knowledge of the scene or explain scene dynamics in terms of human behavior.

How do pedestrians use this space? Where do people congregate and socialize? What are the dominant patterns in pedestrian behavior? Addressing these questions enables identification of social waypoints and a spatial mapping of the space in terms of social behavior. This mapping can be used in robot motion planning. For example, just as a robot should avoid running into a mailbox, it should also be cognizant of social waypoints such as patios, bus-stops, map kiosks, and other places where people congregate and socialize.

* Corresponding Author

We create *PT-net*, a network that learns a *pedestrian behavior dictionary* in an unsupervised, data-driven manner to provide explainable, semantic social behavior labels for pedestrian trajectories in a scene. These can be used to characterize global trends in space usage and pedestrian behavior habits, as well as assist in downstream tasks like trajectory prediction, so that very shallow networks can be used to predict accurate paths. Trajectories are already low-dimensional as a sequence of x-y values with limited long-range dependencies. Consequently, the computational tools used in past trajectory prediction such as RNNs with attention [34, 41] and without [1, 47], transformers [15, 49], and spatio-temporal graph networks [19, 49] may be unnecessarily complex for the task.

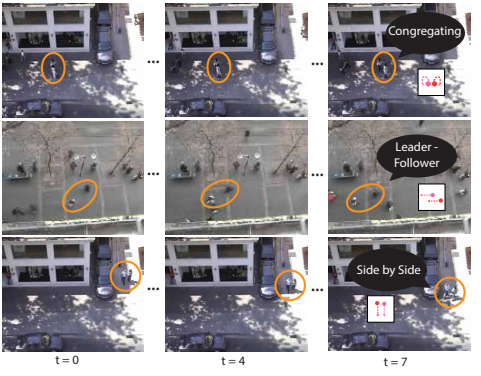

Figure 1: Understanding the underlying social behaviors inherent in pedestrian trajectories makes predicting them easier. We create PT-Net to predict social behaviors, like those shown above, given historical pedestrian locations.

Unsupervised methods are integral to our framework because they do not require prohibitively time consuming labelling of large datasets like their supervised counterparts. Moreover, the labels themselves are unknown in this domain, requiring them to be learned from pedestrian data. We are inspired by methods [51] that learn networks to reproduce preexisting embeddings using a student-teacher framework to circumvent this problem. PT-net creates a stationary latent space embedding of trajectories to create a pedestrian behavior dictionary. This dictionary groups trajectories with similar social behaviors into homogeneous clusters corresponding to interpretable behavior that can be readily human-labelled by observing a small sampling of trajectories within each cluster.

Unsupervised clustering to discover social behaviors avoids the shortcomings of manually defining presumed behaviors. For example, consider the cluster that corresponds to *leader-follower behavior* where two pedestrians travel approximately the same path, one in front of the other, separated by a distance. This behavior cannot be easily manually defined because the inter-person distance is randomly distributed and varies among environments. Our approach supports the discovery of diverse behaviors within a specific environment without the explicit definition of these properties, while having the advantage of being lightweight, unsupervised, and relying on very basic networks to achieve useful characterizations and accurate predictions of pedestrian behavior.

We show the utility of our method on the ETH [30] and UCY [23] datasets. Mapping semantic pedestrian behaviors from the dictionary to specific locations in scenes results in a *pedestrian behavior map* that can answer key questions about an environment. While trajectory prediction is not our main goal, we additionally demonstrate that our social behavior dictionary simplifies the task of pedestrian trajectory prediction and maintains comparable performance with current SOTA that use much larger, more complex networks.

Our main contributions are threefold: **(1)** introduction of PT-net: an unsupervised method for computing a semantically meaningful *pedestrian behavior dictionary* using a novel t-SNE imitator network; **(2)** construction of interpretable *pedestrian behavior maps* to characterize environment usage patterns in terms of pedestrian behavior; **(3)** competitive pedestrian trajectory prediction results with a much simpler network than current SOTA.

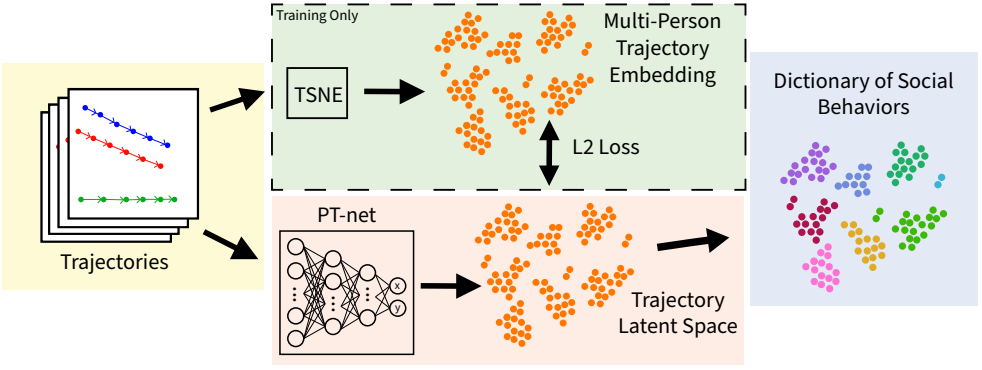

Figure 2: To make the social behavior dictionary, velocity and proximity-based features are computed from groups of trajectories.The t-SNE algorithm creates a 2D trajectory latent space embedding from these features, and PT-net learns to reproduce this embedding by directly predicting corresponding feature coordinates.This learned coordinate embedding separates distinct pedestrian social behaviors into clusters, forming a pedestrian behavior dictionary. During inference, PT-net uses the processed trajectories to get the embedding coordinates and matches them to the closest social behavior cluster.

2 Related Work

Prior work has clustered groups of similar trajectories [17, 39] based on similar motion patterns [7, 10, 20] and group shapes [12], similar probabilistic distributions [52], or using deep clustering [3, 38]. However, most of these clustering methods do not provide interpretable labels for their behavior and rely on hand crafted similarity metrics, with the exception of the deep clustering methods. Recent work in unsupervised guidance for pre-training provides a solution to these problems by building networks to learn good representations of data in a self-supervised manner. For example, [13, 14] learns a representation that matches visual bag-of-words output; [48] trains a network to match 2D t-SNE output, creating a latent space for texture recognition; [8, 9] train networks to generate clusters for unsupervised feature learning; and [39] trains a 3D convnet to match the k-means clustering of an embedding space. We follow this trend of using unsupervised methods to train clustering networks to build PT-net, a t-SNE-imitator network for projecting pedestrian paths to an interpretable latent space. Our approach is more lightweight than previous work and produces clusters corresponding to distinct pedestrian social behaviors.

Early work in trajectory prediction focused on scene-level reasoning using LSTMs [34], intra-trajectory attention [11] with transformers [15], and scene reasoning with trajectory waypoints [22]. Recent work incorporates social information to impact model performance using hand-defined social relationships, such as social pooling based on occupancy maps [1, 26, 46] and relative pedestrian distances [16], or trajectory categorization and social pooling based on similar motion trends [6, 20, 36]. Due to the limitations of hand-crafted features, work began to focus more on learned social features. Social attention with GANs [32], CVAEs [28, 45], and transformers [40] allowed networks to focus on input information specifically pertaining to learned social groups. Concurrently, social interactions were modeled using graphs, and graph networks learned edge features characterizing the social interactions between pedestrians in a scene for both fixed [25, 29, 41] and varying [33, 37, 44] group sizes. Some methods combine social attention and social graph networks [5, 19, 49] to improve learned social features, or propose new loss functions [27] to allow networks to

implicitly learn to predict trajectories that are more socially compliant. While these methods produce reasonable results, they do not provide interpretable reasoning behind their trajectory predictions or explainable social behaviors.

3 Methods

3.1 PT-net for a Trajectory Latent Space

We devise PT-net to act as a t-SNE imitator by using a student-teacher framework to obtain a stable embedding from multi-pedestrian trajectories, which are particularly well-suited for t-SNE-embedding because of their low dimensionality. Given trajectories $X_t^i = \{x_t^i, y_t^i\}$ for each person i in a scene over t time steps, we split the trajectories into overlapping segments of length T where each segment is offset by Δt, and each pedestrian is present for the duration of the time steps in each trajectory segment. This provides environment specific pedestrian behavior examples that are limited by features such as sidewalks, entryways, and roads. The trajectory segments are augmented by rotation ($\theta = 30, 45, 60$ degrees) to insert synthetic variation into the data to make the learned latent space more generalizable. This allows us to learn a more complete social behavior dictionary without the need for more training data.

The relative velocity, v^i, for each pedestrian is computed as follows:

$$v^i = [x_t^i - x_{t-1}^i, y_t^i - y_{t-1}^i] \tag{1}$$

Subsequently, the distance, d^i, between person i at time t and the $N-1$ nearest pedestrians in the scene at time $t-1$, are computed for each time step in the segment as follows:

$$d^i = [x_t^i - x_{t-1}^j, y_t^i - y_{t-1}^j] \quad \forall\, j \in N \text{ where } j \neq i \tag{2}$$

where N is the total number of people in the trajectory segment. We choose raw distances to form d^i, as opposed to absolute or squared distances, to preserve direction in our proximity estimations. These two vectors, v^i and d^i, are computed for each person in the trajectory and concatenated to form

$$D = [\alpha v^i \mid d^i] \quad \forall\, i \in N \tag{3}$$

where α is a scaling factor included to combat the difference in scale between the relative velocities v^i and the proximity-based features d^i and $\mid$ denotes the concatenation operation. This process is repeated for each multi-person trajectory segment from the raw data.

From there, we split the data into groups containing equal numbers of pedestrians and use t-SNE to create a trajectory embedding for each group from its flattened D representation. The t-SNE embedding output is clustered using k-means, where k is chosen through visual inspection of the t-SNE embedding manifold. The cluster assignments of each point in t-SNE space are paired with their corresponding raw trajectories. Sampling small numbers of these points per cluster and comparing the associated raw trajectories reveals a taxonomy of semantically meaningful behaviors like leader-follower, walking in pairs, or standing around in small groups, as shown in Figure 4. See Figure 2 for an overview of this process.

This trajectory embedding learned directly from the pedestrian velocities using t-SNE is useful, but t-SNE embeddings are irreproducible, which is undesirable in the event that it is necessary to predict the social behavior clusters of previously unseen pedestrian trajectories or add to the number of social clusters by processing new data. To combat this, PT-net, a feed-forward MLP network, learns to mimic the resulting t-SNE embeddings. Each cluster in the pedestrian behavior embedding corresponds to one unique social behavior. This effectively quantizes a continuum of behaviors to reduce the dimensionality of an infinite set of social behaviors, thus providing a tractable lexicon for high level scene reasoning.

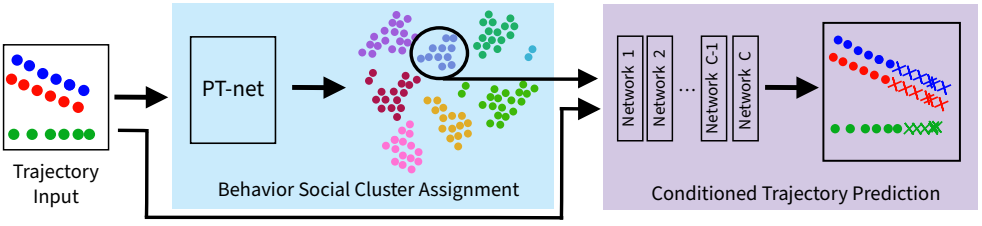

Figure 3: For pedestrian trajectory prediction, PT-net predicts the social behavior cluster assignment corresponding to the behavior of the pedestrians in the scene. This assignment dictates which of the MLPs will be used to predict the future trajectories of the pedestrians. We train one MLP per cluster in the pedestrian behavior dictionary and deterministically condition the prediction upon the social behaviors of the pedestrians in the scene.

3.2 Utilizing the Pedestrian Behavior Dictionary

Once the pedestrian behavior dictionary is created, it can be used to answer questions about human behavior and environment utilization. A particular environment can be characterized by enumerating the social behaviors that occur and computing the frequency and locations of these behaviors. That is, pedestrian behavior histograms and pedestrian behavior maps allow characterization of space usage and discovery of pedestrian behaviors (see Figures 5 and 6). Analysis of space in this manner is directly applicable to social science issues such as public space assessment [18, 35, 42, 43] and pedestrian behavior analysis [2, 21, 31]. Key questions in pedestrian behavior can be answered such as: what is the average distance in leader-follower relationships? Is this distance culturally dependent? What is the radius of movement when people are congregating during conversation or waiting?

Using a set of trajectories X as input, we also predict future trajectories $\hat{X}$ deterministically conditioned on the predicted, semantically meaningful, social behavior cluster, c_i, from PT-net as shown in Figure 3. First, the pedestrians in the scene are clustered into groups of at most N based on relative distance. If the number of pedestrians in the scene exceeds N, then combinations of pedestrians are created based on proximity such that each group contains at most N people and each pedestrian is included in at least one group. Then, the velocity and proximity input D from Equation 3 is assembled for each group in the current scene and flattened. PT-net directly uses this flattened representation to predict the social behavior cluster from the pedestrian behavior dictionary. Based on this cluster assignment, a specific feed-forward MLP network is chosen from an ensemble to make the pedestrian trajectory prediction. We train one MLP to predict future trajectories per cluster in the pedestrian behavior dictionary. This light-weight approach is possible because the dictionary effectively limits the possible state space of the trajectory prediction problem to a manageable range.

4 Results

4.1 Datasets and Training

PT-net is tested against SOTA methods using the ETH [30] and UCY [23] datasets. The trajectories are normalized based on the scene sizes and centered on the origin. Note that each of the datasets occur on sidewalked areas. This limits the types of behaviors that pedestrians can exhibit. For example, it's highly unlikely that a pedestrian will walk in circles or meander directionlessly because the social convention is to walk parallel or perpendicular to the buildings. Subsequently, the trajectories are augmented with several rotations to introduce

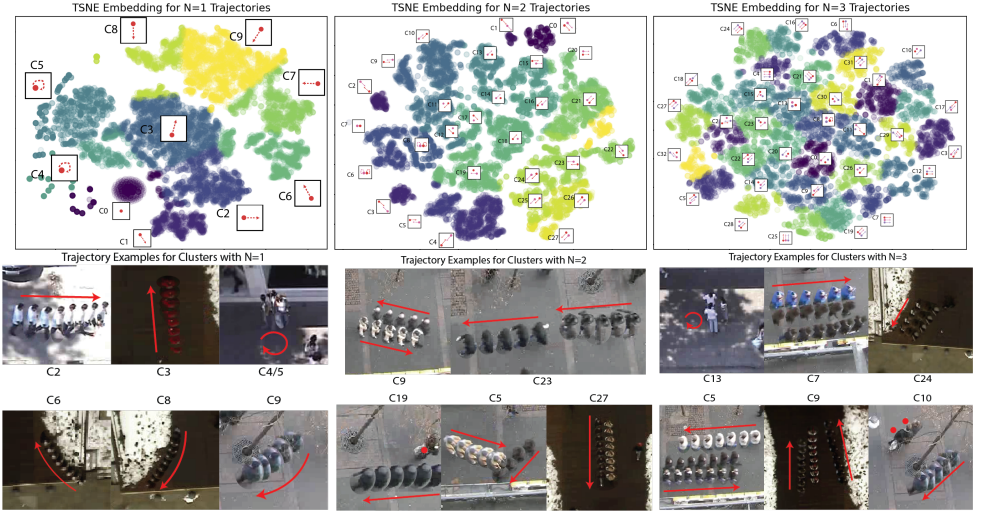

Figure 4: Visualizing the Social Behavior Dictionary. **(Top)** Each dot represents the t-SNE embedding of a 3.2 second (8 frame) trajectory. The colored clusters denote distinct social behaviors for N=1,2,3 people (corresponding colors across the graphs do not denote related behaviors). Cluster diagrams near each cluster illustrate the pedestrian behavior. **(Bottom)** Trajectory diagrams showing a temporal integration of video to see the movement (for a selection of behaviors) in one frame. Red arrows show pedestrian direction.

more behavior variety with the training data available, allowing for the learning of a more complete social behavior dictionary.

To create the input data for t-SNE, we use sliding windows of size $T = 8$ timesteps and $\Delta t = 1$ to learn behaviors that would be relevant to existing work and to allow for more granular detection of social behavior changes. We choose α to be 15, and the maximum number of people in each trajectory used for our experiments is $N = 3$. PT-net is a collection of three, four layer MLPs, one for each N, with ReLU activations trained for 300 epochs.

For the trajectory prediction task, positions $X_t^i = \{x_t^i, y_t^i\}$ for each person i in a scene over $t = 8$ timesteps are input to the ensemble of MLPs which predict $\hat{X}_t^i = \{\hat{x}_t^i, \hat{y}_t^i\}$ positions over $t = 12$ timesteps into the future for each person. This choice of input and prediction horizon is standard in multiple SOTA methods [1, 16, 50]. The ensemble networks consist of four linear layers with ReLU activation functions and residual connections each trained for 1000 epochs on only the data corresponding to their respective cluster assignments. Because PT-net is a scene-specific model, we train on 80% (2880 trajectories) of the data from all scenes and test on the remaining 20% (720 trajectories) for each environment. While some applications require no pre-training on the scene, the observation of a scene before algorithm deployment is quite reasonable in numerous applications, such as IOT, smart buildings, and traffic monitoring, due to a ubiquitous fixed camera.

4.2 Interpretable Pedestrian Behavior Dictionary

The learned latent space for $N = 1, 2, 3$ pedestrians is shown in Figure 4. Drawings with representative example trajectories are superimposed on each graph next to their corresponding clusters. Distinct clusters appear for each N which indicates that our velocity and proximity-based trajectory processing is sufficient for learning distinct behaviors. Because we use velocity-based features, similar behaviors that are executed in different directions (ie. left-

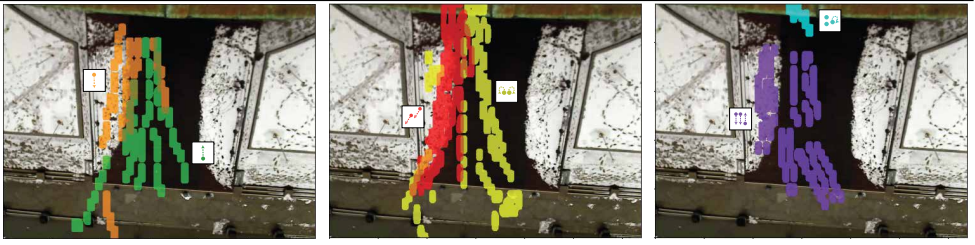

Figure 5: Pedestrian Behavior maps. Colored boxes indicate select social behaviors occurring over the entire data collection period for ETH. We infer a rich story about environment usage for varying numbers of pedestrians and social behavior clusters. **Left:** (Green) Person entering the building; (Orange) Person leaving the building. People leave the building in a more constrained path, indicating they are giving right-of-way to those entering. **Middle:** (Yellow) Two people standing still together; (Red) Two people leaving the building in a leader-follower formation. People standing still tend to congregate off to the sides or at an island in the middle, and people exiting fit into the gaps left behind. **Right:** (Purple) Two people walking side by side to exit the building passing one person entering the building; (Blue) Three people standing still. Notice there is a bottleneck around the door that prevents pedestrians from moving, but movement becomes easier further from the door.

right vs top-bottom) often form different clusters. We show composite images of multiple frames of movement from the ETH and UCY datasets beneath each latent space graph to illustrate example trajectories from a selection of clusters. The red arrows in each image denote the directions of movement for the pedestrians.

For N larger than 3, it is still possible to find social behavior clusters, but their interpretability becomes more challenging as their number increases significantly with each subsequent increase of N and behaviors become more complex. Behavior separation is smaller for high dimensions (see $N = 3$ in Figure 4); however, the embedding space still groups similar pedestrian behaviors in a sufficient manner for the downstream tasks of predicting future pedestrian trajectories and interpreting behavior patterns in a space.

4.3 Pedestrian Behavior Maps

We use the pedestrian behavior dictionary to characterize space usage and social behavior patterns by creating pedestrian behavior maps for each social behavior cluster in an environment. Clusters exhibiting the same behaviors with different pedestrian permutations are combined into the same behavior maps. For example in $N = 2$, clusters $C = 2$ and $C = 3$ show the same semantic and directional behavior, but with the pedestrian order switched. The behavior maps of different clusters are superimposed to analyze the inter-pedestrian and inter-social group interactions that take place.

Figure 5 shows a selection of behavior maps in the ETH environment depicting the location that a particular behavior was exhibited over the duration of the entire dataset. Figure 5 (Left) shows the map for one pedestrian entering the building (green) juxtaposed with one pedestrian leaving the building (orange). There are more people entering the building from the right, thus forcing the people exiting to stay to the left and give right of way to those entering. Figure 5 (Middle) overlays the behavior map of two people standing still together (yellow) and two people leaving the building leader-follower (red). People tend to stand still off to the sides, or at an island in the middle of the walkway. The people leaving the building are forced to travel in the gaps between these congregators. Figure 5 (Right) combines the

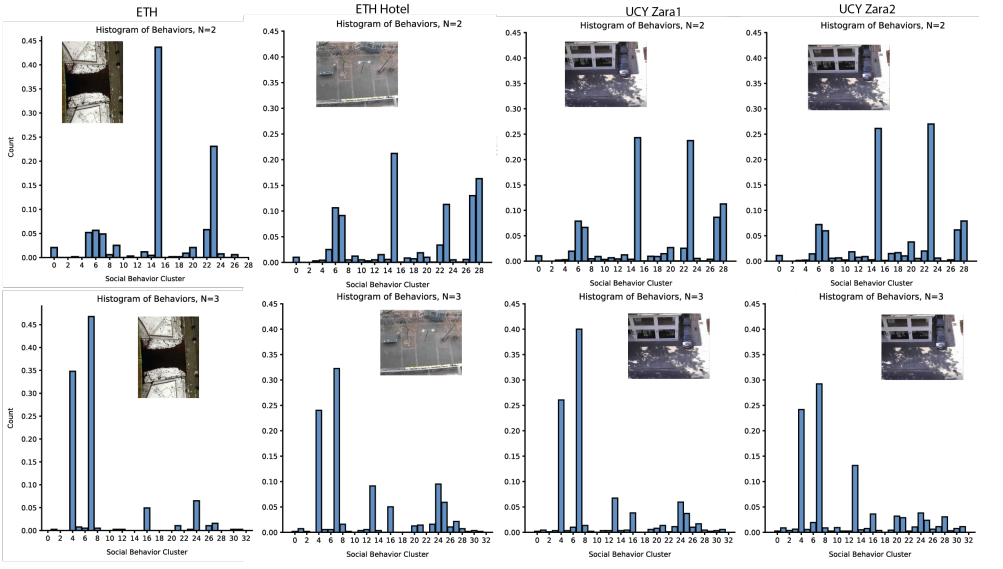

Figure 6: The predicted histogram of behaviors in each environment (ETH, ETH Hotel, UCY Zara1, UCY Zara2) for $N = 2$ (top) and $N = 3$ (bottom) people. Pedestrians utilize a different distribution of behaviors in ETH as opposed to all other environments. Because ETH depicts people walking in and out of a building, it is a much more constrained space than the open sidewalks in the other environments. Even between Zara1 and Zara 2, which take place in the same environment at different times, there is still variation due to differing numbers of pedestrians and different pedestrian behavior patterns as the day progresses.

behavior maps of two people walking side by side to exit the building passing one person entering the building (purple) and that of three people standing still together (blue). The behavior map for three people standing still is localized around the entrance to the building, implying there is a bottleneck around the door preventing people from moving freely. This bottleneck dissipates further away from the door where there is more free space to move. This signals that there may be a need for additional entryways for this building to improve pedestrian traffic flow.

Historically, there have been no avenues to discover these social patterns outside of manual observation of a space, as done frequently in behavioral mapping applications in urban planning and design [9, 24, 43], which can be time consuming and have the potential for increased biases and inaccuracies. Our automated behavioral mapping allows for better quality results with a higher throughput, enabling interpretable social behavior analyses of environments for social scientists.

4.4 Pedestrian Behavior Histograms

Environment usage and social behavior characterization can also be done by observing the distribution of behaviors in each space. Figure 6 shows the predicted pedestrian behavior histograms for ETH, ETH Hotel, UCY Zara1, and UCY Zara2 (along the column dimension) for $N = 2, 3$ people (along the row dimension). The histograms for ETH are significantly different from those of the other environments due to the difference in space usage (entryway to a building vs regular sidewalks). Pedestrians in ETH mainly exhibit horizontal leader-follower and walking side-by-side behaviors (N2:C15,23 and N3:C4,7) that allow them to

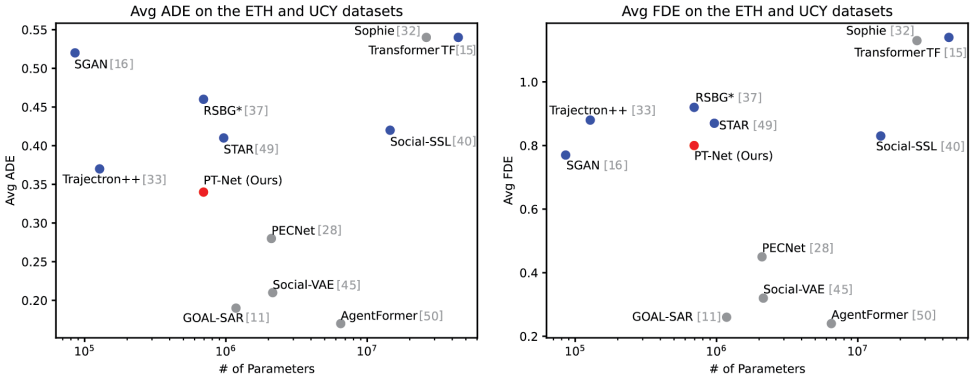

Figure 7: Comparison of the ADE/FDE (lower is better) of PT-net and SOTA for the ETH and UCY datasets. Our method (red) outperforms comparable SOTA shown in blue. We show 53% decrease in ADE over SGAN despite similar FDE results. Trajectron++ performs similarly to our method, but does not provide intepretable social behavior labels. The SOTA methods in grey are multi-pass methods that choose the best of multiple (k=20) sampled predictions when reporting statistics, which is unrealistic in many real world applications. (Note: * denotes an unavailable parameter number which has been set equal to ours for ADE/FDE comparison.) These results are summarized in tabular form in Figure 9.

enter or exit the building, while the other environments allow for more diversified movement. ETH Hotel, UCY Zara1, and UCY Zara2 have higher concentrations of pedestrians exhibiting the group congregating (N2:C6,7 and N3:C13,24), walking side-by-side upwards (N2:C28 and N3:C25), and walking leader-follower or side-by-side diagonally downward (N2:C27 and N3:C16) behaviors than ETH. The spatial structure of ETH Hotel predisposes it towards side-by-side vertical behavior because it has a train stop at the top of the frame. However, UCY Zara1 and UCY Zara2 are more primarily dominated by horizontal leader-follower or side-by-side behaviors because the sidewalk in front of the building is a more popular avenue than the alley at the edge of the frame.

Even with this similarity, UCY Zara1 and UCY Zara2 have much more pedestrian congregation than ETH, showing that pedestrians in the ETH environment are more purposed in their movement or that loitering is not accepted in the space. Additionally, between UCY Zara1 and UCY Zara2, which take place in the same environment at different times, there is variation in pedestrian quantity showing a preference for which time of day they prefer to be more active in the space. Knowing the social behavior distribution in a space has potential for behavior anomaly detection. Specifically, once a behavior baseline has been established, the distributions of different observation periods can be compared for outliers.

4.5 Pedestrian Trajectory Prediction

Figure 7 and figure 9 show a comparison of PT-net and SOTA methods, where we show improved performance on the ETH [30] and UCY [23] datasets. We provide metrics for the average displacement error (ADE) and final displacement error (FDE) of our method in red. Some SOTA methods report ADE_{20} and FDE_{20}, which require running inference for k=20 times and choosing the best prediction from these multiple passes. These multi-pass methods (shown in grey in Figure 7) are impractical in many real world applications, where no ground truth is available to determine the best output and acceptable computation latency may be low. Additionally, they all have significantly more parameters than our approach.

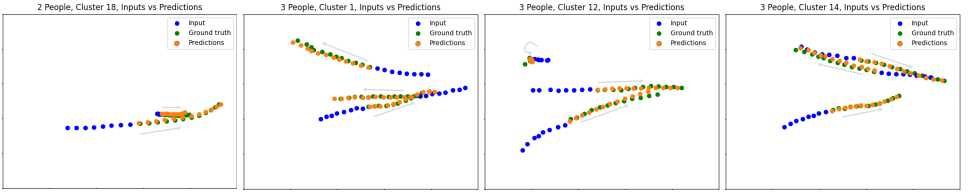

Figure 8: Our framework provides accurate trajectory prediction by conditioning on learned social behaviors. Input trajectories (blue), ground truth future trajectories (green), and predictions (orange) are shown above. Grey arrows show relative pedestrian directions.

The SOTA methods comparable to our approach (shown in blue in Figure 7) do not use the multi-pass strategy, and are referred to as unimodal or deterministic in the literature. Observe that our method outperforms most of these as shown in Figure 7. Social GAN [16] is the only one that has a slightly better FDE, but note that it has a 53% higher ADE (lower is better). Additionally, our method provides some improvement compared to Trajectron++ in terms of ADE and FDE while also adding the functionality of interpretable social behavior labels. Figure 8 shows example trajectory predictions for our method. The input trajectories are in blue, the ground truth future trajectories are in green, and the predictions are in orange. We predict plausible future trajectories for a myriad of socially complex scenarios using simple MLP networks.

Method	Avg of ADE and FDE
TransformerTF [15]	0.84
RSBG* [37]	0.69
SGAN [16]	0.65
STAR [49]	0.64
Social-SSL [40]	0.63
Trajectron++ [33]	0.62
PT-Net (Ours)	**0.57**

Figure 9: Our method outperforms all other comparable SOTA (on deterministic/unimodal prediction) in Average FDE and ADE on the ETH and UCY datasets.

5 Conclusion

In this paper, we propose PT-net, a lightweight, unsupervised method for learning an interpretable pedestrian behavior dictionary for a given environment through trajectory clustering. Unsupervised methods remove the need for costly dataset labeling while allowing for the discovery of a dictionary containing diverse behaviors. With this dictionary, it is possible to characterize space usage and social behavior patterns to answer key questions in social science fields, like psychology and urban planning, using behavior maps and histograms to visualize the distribution of behaviors. While, the completeness of this distribution is limited by the diversity of the observed behaviors from the pedestrian datasets used during training, this can be partially abated through data augmentation. We also demonstrate comparable performance to SOTA in trajectory prediction on the ETH and UCY datasets with a much simpler network. Decreasing the complexity and size of trajectory prediction methods is important for mobile computing and applications with limited computational resources.

Acknowledgements

This work was supported by grant NSF NRT-FW-HTF: Socially Cognizant Robotics for a Technology Enhanced Society (SOCRATES), No. 2021628.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] Gianluca Antonini, Santiago Venegas Martinez, Michel Bierlaire, and Jean Philippe Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69(2):159–180, 2006.

[3] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Learning pedestrian group representations for multi-modal trajectory prediction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 270–289. Springer, 2022.

[4] Souhaila Bendjedidi, Yassine Bada, and Rim Meziani. Urban plaza design process using space syntax analysis: El-houria plaza, biskra, algeria. *International Review for Spatial Planning and Sustainable Development*, 7(2):125–142, 2019.

[5] Alessia Bertugli, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. Ac-vrnn: Attentive conditional-vrnn for multi-future trajectory prediction. *Computer Vision and Image Understanding*, 210:103245, 2021.

[6] Niccoló Bisagno, Bo Zhang, and Nicola Conci. Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[7] Niccoló Bisagno, Cristiano Saltori, Bo Zhang, Francesco GB De Natale, and Nicola Conci. Embedding group and obstacle information in lstm networks for human trajectory prediction in crowded scenes. *Computer Vision and Image Understanding*, 203: 103126, 2021.

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

[10] Guangyi Chen, Junlong Li, Nuoxing Zhou, Liangliang Ren, and Jiwen Lu. Personalized trajectory prediction via distribution discrimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15580–15589, 2021.

[11] Luigi Filippo Chiara, Pasquale Coscia, Sourav Das, Simone Calderara, Rita Cucchiara, and Lamberto Ballan. Goal-driven self-attentive recurrent networks for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2518–2527, 2022.

[12] Weina Ge, Robert T. Collins, and R. Barry Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):1003–1016, 2012. doi: 10.1109/TPAMI.2011.176.

[13] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6928–6938, 2020.

[14] Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Online bag-of-visual-words generation for unsupervised representation learning. *arXiv preprint arXiv:2012.11552*, 2020.

[15] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10335–10342. IEEE, 2021.

[16] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018.

[17] Yutao Han, Rina Tse, and Mark Campbell. Pedestrian motion model using non-parametric trajectory clustering and discrete transition points. *IEEE Robotics and Automation Letters*, 4(3):2614–2621, 2019.

[18] Jordi Honey-Rosés, Isabelle Anguelovski, Vincent K Chireh, Carolyn Daher, Cecil Konijnendijk van den Bosch, Jill S Litt, Vrushti Mawani, Michael K McCall, Arturo Orellana, Emilia Oscilowicz, et al. The impact of covid-19 on public space: an early review of the emerging questions–design, perceptions and inequities. *Cities & health*, 5(sup1):S263–S279, 2021.

[19] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[21] Parth Kothari, Brian Sifringer, and Alexandre Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15556–15566, 2021.

[22] Mihee Lee, Samuel S Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. Muse-vae: Multi-scale vae for environment-aware long term trajectory prediction. *arXiv preprint arXiv:2201.07189*, 2022.

[23] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[24] Maryam Lesan and Morten Gjerde. Sidewalk design in multi-cultural settings: A study of street furniture layout and design. *Urban Design International*, 26:21–41, 2021.

[25] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020.

[26] Matteo Lisotto, Pasquale Coscia, and Lamberto Ballan. Social and scene-aware trajectory prediction in crowded spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[27] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15118–15129, 2021.

[28] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European Conference on Computer Vision*, pages 759–776. Springer, 2020.

[29] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020.

[30] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, pages 261–268. IEEE, 2009.

[31] Th Robin, Gianluca Antonini, Michel Bierlaire, and Javier Cruz. Specification, estimation and validation of a pedestrian walking behavior model. *Transportation Research Part B: Methodological*, 43(1):36–56, 2009.

[32] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[33] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700. Springer, 2020.

[34] Nasim Shafiee, Taskin Padir, and Ehsan Elhamifar. Introvert: Human trajectory prediction via conditional 3d attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16815–16825, 2021.

[35] Mario L Small and Laura Adler. The role of space in the formation of social ties. *Annual Review of Sociology*, 45:111–132, 2019.

[36] Yue Song, Niccoló Bisagno, Syed Zohaib Hassan, and Nicola Conci. Ag-gan: An attentive group-aware gan for pedestrian trajectory prediction. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8703–8710. IEEE, 2021.

[37] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 660–669, 2020.

[38] Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13250–13259, 2021.

[39] Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. Unsupervised learning of video representations via dense trajectory clustering. In *European Conference on Computer Vision*, pages 404–421. Springer, 2020.

[40] Li-Wu Tsao, Yan-Kai Wang, Hao-Siang Lin, Hong-Han Shuai, Lai-Kuan Wong, and Wen-Huang Cheng. Social-ssl: Self-supervised cross-sequence representation learning based on transformers for multi-agent trajectory prediction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 234–250. Springer, 2022.

[41] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pages 4601–4607. IEEE, 2018.

[42] William H Whyte. The design of spaces. In *The city reader*, pages 631–639. Routledge, 2015.

[43] William Hollingsworth Whyte et al. The social life of small urban spaces. 1980.

[44] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6507, 2022.

[45] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 511–528. Springer, 2022.

[46] Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1194. IEEE, 2018.

[47] Hao Xue, Du Q Huynh, and Mark Reynolds. Poppl: Pedestrian trajectory prediction by lstm with automatic route class clustering. *IEEE transactions on neural networks and learning systems*, 32(1):77–90, 2020.

[48] Jia Xue, Hang Zhang, and Kristin Dana. Deep texture manifold for ground terrain recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2018.

[49] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision*, pages 507–523. Springer, 2020.

[50] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.

[51] Hang Zhang, Jia Xue, and Kristin Dana. Deep ten: Texture encoding network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 708–717, 2017.

[52] Zhang Zhang, Kaiqi Huang, Tieniu Tan, Peipei Yang, and Jun Li. Red-sfa: Relation discovery based slow feature analysis for trajectory clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–760, 2016.