Real-Time Certified Probabilistic Pedestrian Forecasting
Author's Response

Henry O. Jacobs, Owen K. Hughes,
Matt Johnson-Roberson, and Ram Vasudevan

May 8, 2017

We would like to begin by thanking the reviewers and the editor for their careful reading and review of our paper. The comments provided were incredibly helpful and insightful, and have strengthened the paper considerably. Based on the recommendations of the reviewers and the editor, we have made several modifications to the paper. The changes made are summarized in detail below and a highlighted manuscript is attached that indicates additions in red. **This letter is largely self contained, so references here are to the bibliography in this file**.

Reviewer 1

1. **This paper presents a novel real-time probabilistic forecasting method for pedestrian trajectories via observing the historical trajectories in a particular scene. The problem is well motivated, formulated and handled. The reported experimental evaluations show pivotal improvement against the current state-of-the-art.**

   Thank you for the positive assessment.

2. **With the current form of the results it is hard to compare the improvements in terms of accuracy gained. This can be overcome by including results such as physical distance between predicted paths and ground truth observations. Such matrices are provided in the baseline models [1] [2] [3] with accuracy values in meters.**

   Thank you for the suggestion. The objective of this paper was to develop a predictor that would be effective for use in the autonomous vehicle context. In this instance, to ensure safety it is critical that the generated set of predictions contains the ground-truth observed trajectory while including as few false positive detections as possible. Since the ROC curve plots the true positive rate (i.e. the proportion of the ground truth trajectory that is correctly identified) against the false positive rate (i.e. the proportion of the generated set that does not correspond to the ground truth trajectory), it evaluates this aforementioned safety criteria for autonomous vehicles exactly.

   On the other hand the Modified Hausdorff Distance (MHD) from [4] measures the average geometric distance between the ground-truth observed trajectory and the generated set of predictions. Though popular in evaluating the geometric proximity of the ground truth trajectory and a predicted set, it is not the most appropriate way to evaluate the utility of an algorithm in the autonomous vehicle context. Specifically consider the instance of a generated

set of predictions which does not contain the ground-truth observed trajectory but is close to the ground-truth trajectory geometrically. If this generated set of predictions was used in the autonomous vehicle context, it would not preclude certain portions of the space where the person was occupying. Notice that In this instance the true positive rate would be zero meaning that the generated set of predictions was not useful. Whereas the MHD would describe this generated set of predictions as informative.

Nevertheless, we addressed this concern by adding a comparison to the MHD which is summarized in the Figure 1 below and included in the resubmitted version of our manuscript. We outperform [1] and [5] in the intermediate term, but at short times [5] puts more weight on the direction of the observed trajectory, and so their narrow distribution is closer to the trajectory. At large scales, [1] outperforms us because they know the endpoint of the trajectory. If you look at Figure 3 and 4, you'll see that while they are geometrically closer to the observed trajectory, they don't include it. We have included the following passage to explain both metrics in the revised submission:

#4.5

> In our analysis, we sought a metric that evaluated the utility of prediction algorithms in the autonomous vehicle context. In this instance it is critical that the generated set of predictions contains the ground-truth observed trajectory while including as few false positive detections as possible. ROC curves which plot the true positive rate (i.e. the proportion of the ground truth trajectory that is correctly identified) against the false positive rate (i.e. the proportion of the generated set that does not correspond to the ground truth trajectory), evaluate this aforementioned safety criteria for autonomous vehicles exactly. The Area Under the Curve (AUC) is a standard measure of the quality of a predictor. The closer that this AUC is to one, the better the prediction algorithm. Figure 2 shows the analysis of the AUC of each algorithm versus time. In addition, we evaluated the Modified Hausdorff Distance (MHD) from the ground truth trajectory to a sample from the predictions at each time to provide a geometric measure of how accurate the predictions are. Figure 1 shows MHD plotted against time. Though popular in evaluating the geometric proximity of the ground truth trajectory and a predicted set, it is not the most appropriate way to evaluate the utility of an algorithm in the autonomous vehicle context. Specifically consider the instance of a generated set of predictions which does not contain the ground-truth observed trajectory but is close to the ground-truth trajectory geometrically. If this generated set of predictions was used in the autonomous vehicle context, it would not preclude certain portions of the space where the person was occupying. Notice that In this instance the true positive rate would be zero meaning that the generated set of predictions was not useful. Whereas the MHD would describe this generated set of predictions as informative.

3. **How the model can incorporate the interactions among pedestrians (i.e group motion) and the influences from neighbouring pedestrians (for instances such as collision avoidance)?**

   Thank you for this insightful question. Since the focus of this paper was on developing a method to generate predictions at real-time for the autonomous vehicle context, we did not consider the effect of incorporating interactions between predictions. However, social forces are a natural avenue for future work, and so we included the following paragraph in the paper:

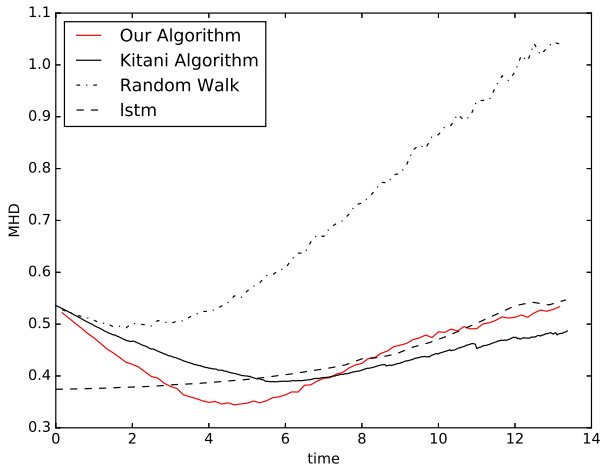

Figure 1: A comparison of the Modified Hausdorff Distance from the ground truth of the pedestrian to a 1000 point samples from each distribution. The method from [5] does well at short time scales since it places a highly confident distribution at the given initial position of the pedestrian, but the method developed in this paper outperforms all others at intermediate times. At longer timescales the MHD to the trajectory of most algorithms converges. [1], which requires the end point of each trajectory, outperforms all other algorithms which assume that the end point of the trajectory is unknown.

#4.10

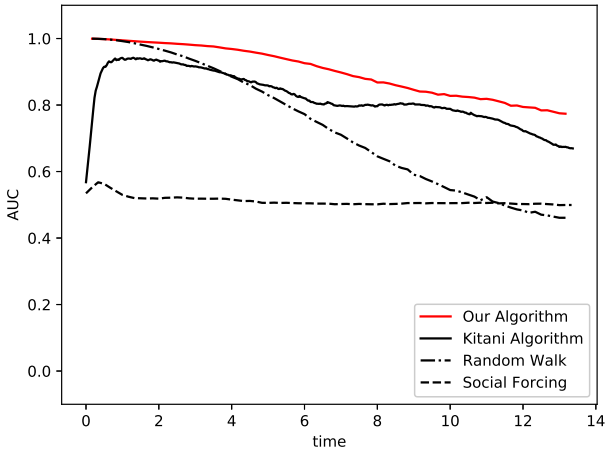

Figure 2: A comparison of the AUC of the various algorithms. Note that the initial dip in the performance of [1] is due to their confidence in their initial estimate. We sampled the S-LSTM [5] model 100 times in order to organically extract a less concentrated probability distribution from their algorithm, but their confidence combined with the over-reliance on social forces at moderate-to-large timescales lowered their performance.

#4.9

Though including social forces into the model is outside of the scope of this paper, its incorporation could begin as follows. In the current motion model with a single vector field, the acceleration of the ith agent is given by $\ddot{x}_i = s^2 DX(x_i) \cdot X(x_i)$. Incorporating a social force F_i acting on the ith agent can be done by instead asserting $\ddot{x}_i = s^2 DX(x) \cdot X(x) + F_i$. Usually $F_i = \sum_j \nabla U(x_j - x_i)$ where U is an interaction potential [6]. The algorithms for generating real-time prediction would then generalize to this new vector-field definition.

4. **Presentation and organisation The motivation behind choosing start/end points based clustering is not clear. Does clustering based on start/end points help to identify different motion models? Isn't it better to cluster considering the entire trajectory?**

   Thank you for this relevant question. As stated in [7], the choice of clustering technique is not nearly as important as the distance metric. Our success with distance functions that took a whole trajectory was limited on the data we tested and we found more success using a custom distance metric that only considered the start and end points of a trajectory as we describe in Comment 6 to Reviewer 1. We have several hypotheses as to why this is true. For one, the spatial scale of the data we test against showcases important features (e.g. curb cuts and storefronts), which lead to clustered start points and end points. Similarly, we tested against high quality data in which trajectories with irreconcilable occlusions were rare and thus could be omitted.

5. **Do you cluster different scenes provided in campus dataset [5] (i.e Valley, Fangron) separately or all together? When motion models are learnt, are they learnt in scene specific manner or altogether?**

   Thank you for noticing that this is unclear. Our algorithm trains on each scene independently, so clustering, vector field learning, and learning the potential functions are all done for each scene separately. To make this more clear, we added the following text to the paper:

#4.3

We did a 2-fold cross validation by using 20% of the data for testing and the remainder for training within each scene. We learned separate collections of vector fields and model parameters for each fold on all of the four scenes on which we tested.

6. **The motivation behind choosing the affinity propagation clustering algorithm is not clearly stated. What is the distance measure used (i.e Euclidian) ? Not specifying such information makes reproduction of this work not feasible.**

   This was an oversight on our part, thank you for noticing it. The distance used is a custom distance function. Let one trajectory A start at $(x_{A,\text{start}}, y_{A,\text{start}})$ and end at $(x_{A,\text{end}}, y_{A,\text{end}})$, and another trajectory B start at $(x_{B,\text{start}}, y_{B,\text{start}})$ and end at $(x_{B,\text{end}}, y_{B,\text{end}})$. We define the points $\mathbf{a}_1 = (x_{A,\text{start}}, y_{A,\text{start}}, x_{A,\text{end}}, y_{A,\text{end}})$, $\mathbf{a}_2 = (x_{A,\text{end}}, y_{A,\text{end}}, x_{A,\text{start}}, y_{A,\text{start}})$, and $\mathbf{b} = (x_{B,\text{start}}, y_{B,\text{start}}, x_{B,\text{end}}, y_{B,\text{end}})$. We then define our distance measure as $d(A, B) := \min\{d_e(\mathbf{a}_1, \mathbf{b}), d_e(\mathbf{a}_2, \mathbf{b})\}$, for the euclidean distance d_e in $\mathbb{R}^4$. This is designed such that transposing endpoints of a trajectory doesn't matter, and so a trajectory that starts at point A and ends at point B is close to a trajectory that starts at point B and ends at point A. We have included the following clarification:

#4.2

We then cluster in $\mathbb{R}^4$ using Affinity Propagation [8] and a custom distance measure defined on the endpoints of trajectories. Let one trajectory A start at $(x_{A,\text{start}}, y_{A,\text{start}})$ and end at $(x_{A,\text{end}}, y_{A,\text{end}})$, and another trajectory B start at $(x_{B,\text{start}}, y_{B,\text{start}})$ and end at $(x_{B,\text{end}}, y_{B,\text{end}})$. We define the points $\mathbf{a}_1 = (x_{A,\text{start}}, y_{A,\text{start}}, x_{A,\text{end}}, y_{A,\text{end}})$, $\mathbf{a}_2 = (x_{A,\text{end}}, y_{A,\text{end}}, x_{A,\text{start}}, y_{A,\text{start}})$, and $\mathbf{b} = (x_{B,\text{start}}, y_{B,\text{start}}, x_{B,\text{end}}, y_{B,\text{end}})$. We then define our distance measure as $d(A, B) := \min\{d_e(\mathbf{a}_1, \mathbf{b}), d_e(\mathbf{a}_2, \mathbf{b})\}$, for the euclidean distance d_e in $\mathbb{R}^4$. This function measures the distance between the endpoints irrespective of their ordering, which means that a trajectory that starts from point A and ends at point B will be close to a trajectory that starts from point B and ends at point A. The scale of the datasets we tested on had large enough spatial scale that clustering based on endpoints captured people moving from destination to destination, e.g. from a storefront to the sidewalk at the edge of a scene. On this dataset, distance functions that utilize the entire trajectory such as DTW, LCSS, and PF from [7] and TRACLUS from [9] did not identify pedestrian intent as well as our method did. While it is unclear why TRACLUS underperformed on our data, PF, DTW, and LCSS appeared to put trajectories that were very similar for periods of time together even though they had very different intents.

7. **Adequacy of Citation I believe literature review section can be improved using a sub section on social force models, which is extensively applied for pedestrian trajectory forecasting. Helbing and Molnar, 1995; Koppula and Saxena, 2013; Pellegrini et al., 2010; Yamaguchi et al., 2011; Xu et al., 2012; Wang et al. 2008; Hospedales et al. 2009; Emonet et al. 2011; Yi et al. 2015.**

   We agree that this adds to the paper, thank you for the suggestion. We added the following section to our literature review to reflect the relevant work:

#1.1

On the other hand, many authors have approached pedestrian forecasting by deriving their motion model from interactions between pedestrians. Early work by [6] and [10] describe pedestrians interactions using physically motivated methods. Several models such as [11], [12], and [13] derive their motion models from [6] who incorporates collision avoidance through an interaction potential. However this method suffers from not planning for other pedestrian' positions at future times. [14], [15] , and [16] all take optical flow as input. [16] and [14] both use variants of Hierarchical Dirichlet Processes on discretized optical flow to determine temporal motifs (i.e. classes of motion within the scene), where [15] substitutes a Markov model. These models are not agent based, and the lack of an explicit motion model limits their predictive power. [17] predict trajectories by introducing and sampling Anticipatory Temporal Conditional Random Fields which incorporate learned affordances for humans and objects based on observed objectives within the scene. [18] , [19], and [20] create agent-based models based on Gaussian Processes, though they suffer issues when trained on discretized trajectories. [5] uses Long Short-Term Memory (lstm) to learn pedestrian motion models without making assumptions about the manner in which agents will interact and is the state-of-the-art algorithm for socially based models while also having a quick run time.

8. **Furthermore a comprehensive review on trajectory clustering algorithms is required. This should be used to motivate the reason of choosing affinity propagation algorithm in section 4. Morris and Trivedi 2009; Giannotti et al. 2007; Lee et al. 2007; Ester et al. 1996**

   Thank you for this suggestion and for the list of relevant titles. We included a note about different algorithms as well as a justification for why we chose our metric which is described in Comment 6 to Reviewer 1.

Reviewer 2

1. **Keep the good work going. Interesting approach.**

   Thank you for the kind comments.

2. **Need to be clearer and dive a little bit deeper into the technical approach.**

   Thank you for this criticism. We hope we adequately addressed it by strengthening the description of our method for learning the motion model, as well as quantifying the way that we tested our data as is described in our responses to the suggestions by other reviewers.

3. **Need to use and apply more deep learning approaches such as social LSTM or other machine methods.**

   Thank you for the suggestion. We hope that we have addressed it by including a review of literature of social motivated models, including S-LSTM, as well as a section on how social factors can be integrated into our model. In addition, we compared our approach against S-LSTM in addition to the method proposed in [1]. We included the visualizations in Figures 3 and 4 below, also which are also updated in the paper.

#4.8

Reviewer 5

1. **It would be more interesting if the authors show whether the prediction algorithm would transfer the learned "knowledge" of forecasting in novel scenes.**

   Thank you for your comment. Though this was not the focus of this paper, the low-dimensional parameterization of the vector fields make transfer learning more amenable. As it is, this is as an immediate avenue for future work which we describe in the following addition to the text:

   > We also hope to do transfer learning with this model using scene segmentation from [21], as well as the semantic context descriptors and routing scores from [22] to show how vector fields can be transferred to novel scenes. It appears that the low-order parameterization of our model, and unit-length vector field assumption make it particularly amenable to the methods developed in [22].

2. **More statistically result is expected rather than analyzing "four different scenes", and comparing run-time by "averaged across several agents and scenes" (how many?).**

   Thank you for noticing this oversight on our part. We have modified the paper in the passages listed below to clarify the list of scenes we ran in comparison, and how many agents total were run. The set of agents used during evaluation were the same that were used to determine run times. Note that we attempted to compare on all of the scenes in the Stanford Drone Dataset, but the difficulty in getting the other models to converge in a reasonable amount of time prevented us from doing so.

#4.4

> Our analysis was conducted on the Coupa, Bookstore, Death Circle, and Gates scenes from the dataset from [3], with a total of 142 trajectories analyzed.

#4.6

> The run time per frame for each algorithm was generated using the mean run time for 400 frames, averaged across all of the trajectory data used in the quality analysis.

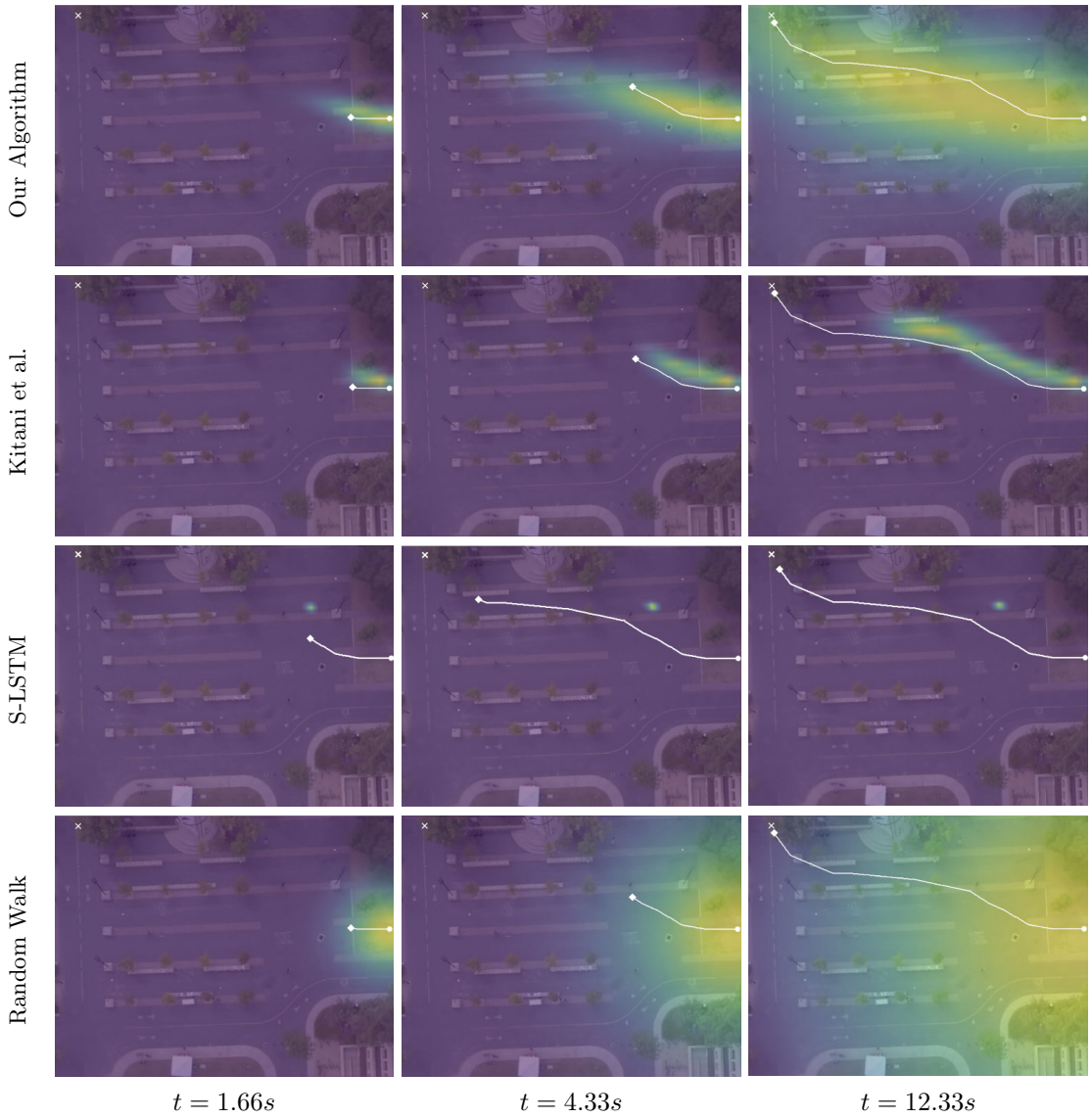

Figure 3: An illustration of the predictions generated by the various algorithms. In this figure, the dot is the start point of the test trajectory, the diamond is the position at time t, and the X is the end of the trajectory. The likelihood of detection is depicted using the virdis color palette. Notice that the Random Walk is imprecise, while the predictions generated by the algorithm in [7] suffer from the inability of their motion model to adequately match the speed of the agent. The algorithm in [13] is confident and close to the trajectory at small times, but their lack of a motion model causes their prediction to compress into a point at intermediate time scales.

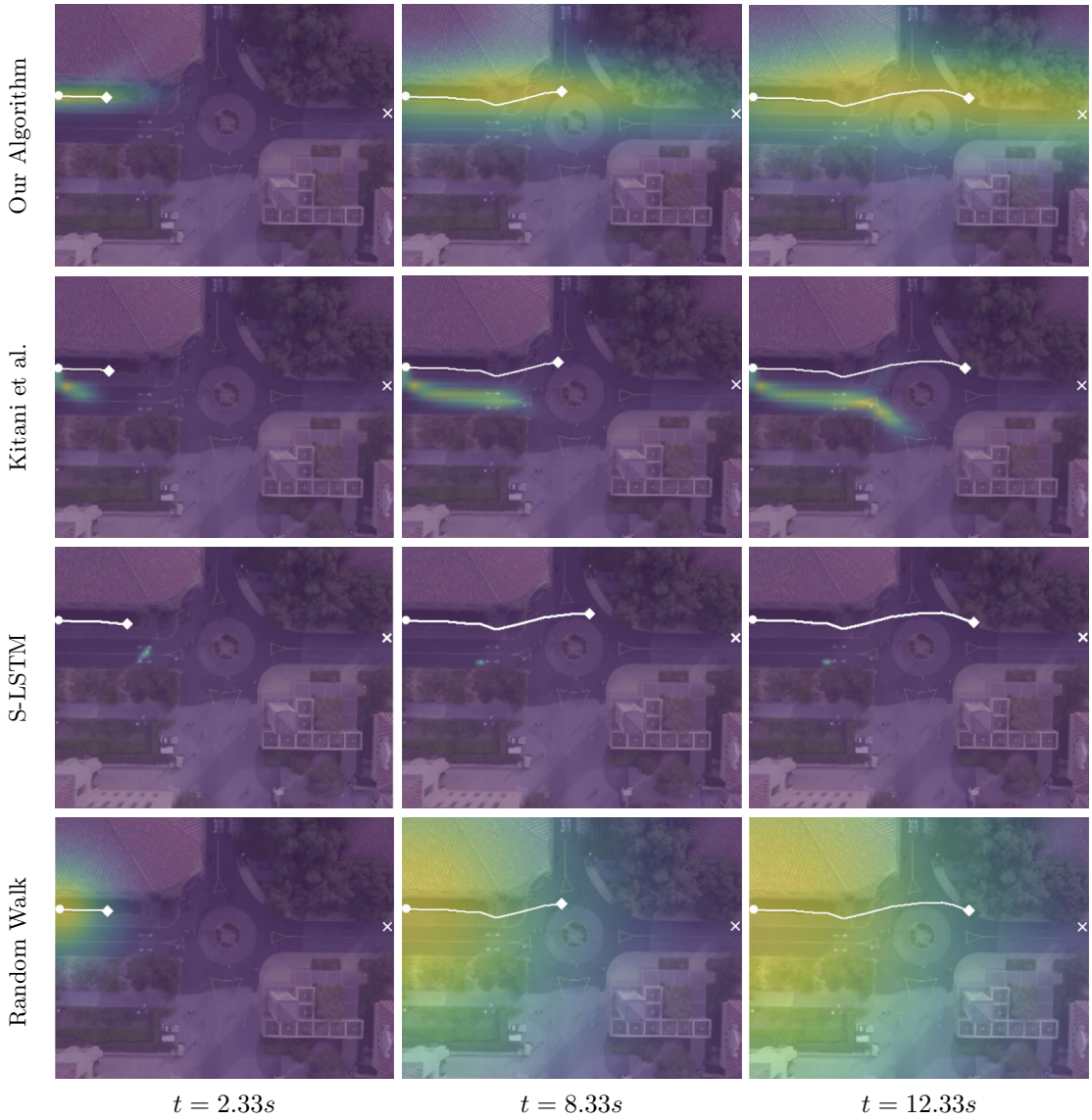

Figure 4: An illustration of the predictions generated by the various algorithms. In this figure, the dot is the start point of the test trajectory, the diamond is the position at time t, and the X is the end of the trajectory. The likelihood of detection is depicted using the virdis color palette. Notice that the Random Walk is imprecise, while the predictions generated by the algorithm in [7] are unable to match the speed of the agent and choose the wrong direction to follow the agent around the circle. The algorithm in [13] is confident and close to the trajectory at small times, but their lack of a motion model causes their prediction to compress into a point at intermediate time scales.

3. **Besides, it is better to explain the meaning of ROC and AUC in section IV.**

   This is a necessary clarification, thank you for the suggestion. We have included a description of ROC and AUC, as well as a justification for why we chose to evaluate the methods using this criterion in our response to Comment 2 for Reviewer 1.

References

[1] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, *Activity Forecasting*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 201–214.

[2] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, "Intent-aware long-term prediction of pedestrian motion," *Proceedings of the International Conference on Robotics and Automation (ICRA)*, May 2016.

[3] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, *Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes*. Cham: Springer International Publishing, 2016, pp. 549–565. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46484-8_33

[4] M.-P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1. IEEE, 1994, pp. 566–568.

[5] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 961–971.

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

[7] B. Morris and M. Trivedi, "Learning trajectory patterns by clustering: Experimental studies and comparative evaluation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 312–319.

[8] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[9] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007, pp. 593–604.

[10] Y. Xu and H.-J. Huang, "Simulation of exit choosing in pedestrian evacuation with consideration of the direction visual field," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 991–1000, 2012.

[11] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1345–1352.

[12] S. Yi, H. Li, and X. Wang, "Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance," *IEEE transactions on image processing*, vol. 25, no. 9, pp. 4354–4368, 2016.

[13] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 261–268.

[14] R. Emonet, J. Varadarajan, and J.-M. Odobez, "Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3233–3240.

[15] T. Hospedales, S. Gong, and T. Xiang, "A markov clustering topic model for mining behaviour in video," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1165–1172.

[16] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[17] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[18] M. Tay and C. Laugier, "Modelling smooth paths using gaussian processes," in *Field and Service Robotics*. Springer, 2008, pp. 381–390.

[19] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 283–298, 2008.

[20] P. Trautman, J. Ma, R. M. Murray, and A. Krause, "Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 335–356, 2015.

[21] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Computer Vision and Pattern Recognition*, 2014.

[22] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," in *Proc. of European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, October 2016. [Online]. Available: http://arxiv.org/abs/1603.06987