A2X: An Agent and Environment Interaction Benchmark for Multimodal Human Trajectory Prediction

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] These can be found through the link to the repository.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We provide the minimum and maximum error in addition to the mean.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

A Appendix

To ensure accessibility:

1. We have provided the link to access the dataset and its metadata in the abstract.
2. The dataset is in the same data format used by prior works [14, 8], and it is divided into two main components: A2A and A2E. Each component has training/validation/testing splits, which can be combined for evaluation on the entire **A2X** dataset. The subdirectories within A2A and A2E simply organize the underlying files used for training/testing. The files within the subdirectories exist as triples. For a file named "data0":
   - "data0.txt" contains the position data needed for extracting scenarios for training/testing.
   - "data0.png" represents the environment as a binary image at twice the scale of the position data.
   - "data0.hom" contains the homography matrix needed to align the environment data with the position data.

   Both the position data and homography matrix are comma-delimited, and the columns of the position data correspond to (1) the time in frames, (2) a unique pedestrian ID, (3) the x-coordinate of the pedestrian's position in meters, and (4) the y-coordinate of the position in meters.
3. The dataset repository also contains the code used to train/test Social GAN [5], PECNet [11], and Trajectron++ [15].
4. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License and is intended for non-commercial academic use.
5. The dataset will be permanently hosted on GitHub for long-term preservation, and upon acceptance, the GitHub repository will be archived and assigned a DOI with Zenodo, a data archiving tool.

A.1 Existing Methods for Human Trajectory Prediction

Social LSTM [1] and Social Attention [18] propose a deterministic model which predict a sequence of future trajectories given a sequence of observed trajectories. Inherently, however, forecasting trajectories is accompanied by the uncertainty in the future, and it raises the question of the limitation of those uni-modal models which predict only one sequence of future trajectories given a sequence of observed trajectories. Many studies [5, 11, 15, 20, 6, 10] assume the multi-modalities in the future human behavior and predict its distribution to learn the uncertainty. In this paper, we focus on three State-Of-The-Art methodologies to demonstrate our benchmark dataset; SocialGAN [5], PECNet [11], and Trajectron++ [15].

SocialGAN [5] adopts GAN [4] framework to forecast all possible future trajectories. The generator creates samples similar with the data distribution while the discriminator is trained to distinguish if the samples are the ground-truth or the generated data. Furthermore, they tackle the problem of collision between pedestrians present in one scene by introducing pooling mechanism. The global pooling of each person's feature in one scene captures the human-human interaction, which prevents the collision between neighboring pedestrians.

PECNet [11] solves trajectory prediction problem with two steps. First, they predict the future goal position based on the observed trajectories by modeling the distribution of the goal positions with Variational Autoencoder (VAE) [7]. The second step is to predict each step of future trajectories by interpolating the observed trajectories and the estimated future goal position.

Trajectron++ [15] proposes a graph structured recurrent model based on conditional VAE [16] to predict the future trajectories. In the training time, they encode both past and future trajectories to obtain the latent factor z from the posterior distribution while in the inference time, it is sampled from the prior distribution only based on the past trajectories. By taking advantage of the graph structure, they introduce edge encoding to model the interaction between nodes (agents) in one scene. Moreover, they incorporate the heterogeneous data by encoding the local map to avoid the obstacle collision.

We investigate these three models as the representatives of the various state-of-the-art works. We choose them because PECNet shows outstanding performance on the long-term trajectory while the short-term trajectory is most well predicted in Trajectron++. We expect SocialGAN, as one of the earliest and most frequently referred models, to be a bound around existing models with respect to PECNet and Trajectron++.

A.2 Existing Datasets for Human Trajectory Prediction

ETH [12] and UCY [9] are commonly used datasets that contain five outdoor scenes (ETH and Hotel from ETH, Univ, Zara1 and Zara2 from UCY), with jointly more than 1600 pedestrian trajectories annotated every 0.4 seconds. They include collision avoidance and group movement.

Stanford drone dataset (SDD) [13] consists of eight outdoor scenes in Stanford campus collected from a drone. The dataset contains more than 19,000 targets including not only pedestrians, but also bicyclists, skateboarders, cars and buses. The coordinates of the trajectories are in the image coordinate system from the bird's eye view, instead of physical word coordinate system.

Stanford crowd dataset (CFF) [2] consists of pedestrian trajectories collected within a train station building of size 25m $\times$ 100m for 12 $\times$ 2 hours by a set of distributed cameras. The dataset is quite noisy, due to the detection, tracking and localization error, and the difficulty to measure the accurate positions of the non navigable areas.

L-CAS 3D Point Cloud People Dataset (LCAS) [19] consists of 28,002 scan frames collected within a university building by a 3D LiDAR mounted on a robot that is either stationary or moving. A scan frame contains around 30,000 3D points, based on which pedestrians are labeled with 3D bounding boxes and marked as either visible or partially visible.

WILDTRACK [3] collected with seven static HD cameras in a public square captures and annotates dense groups of pedestrians for about 60 minutes. The seven cameras' fields of view in large part overlap, allowing precise joint calibrations of image sequences, which may further ensure high-precision trajectory data.

There are other datasets that combines existing repositories. For instance, TrajNet++ [8] combines ETH/UCY, CFF, LCAS, and Wildtrack datasets, as well as a synthetic dataset generated by ORCA [17].

Existing human trajectory datasets have limitations in the sense of embodying interactions. They either do not contain A2E interactions [3], or exhibit limited A2A interactions at small scale in simple environments. We speculate that many self-centered pedestrians are prone to avoid or mitigate, consciously or unconsciously, the influence of the environments and other pedestrians during their navigation. In this work, we are proposing datasets that augment A2E and A2A interactions, which may bring benefits for enhancing learning models by encoding more complex dynamics in trajectories.

A.3 Existing Benchmarks for Human Trajectory Prediction

Average Displacement Error (ADE) and Final Displacement Error(FDE) suggested by SocialL-STM [1] is most commonly used evaluation metrics by most of the trajectory forecasting works. ADE is the average L2 distance between the ground truth and the predicted trajectories across all future steps in a given prediction window. FDE is the L2 distance between the ground truth final destination and the predicted final destination at the end of the future steps in a given prediction window.

However, many trajectory forecasting models assume multi-modality in the future behavior, which makes their model generate more than one prediction of the sequence of future trajectories given one sequence of past trajectories. The current strategy used in the prior works is reporting the minimum ADE / FDE results across randomly sampled k predictions where $k = 20$ in most cases.

In order to evaluate the multi-modal models, Trajectron [6] introduces Negative LogLikelihood (NLL) which is used also in [15, 8]. Given a future time step to predict, they compute the average NLL of the GT trajectory under a distribution generated by a kernel density estimate on trajectory prediction samples.

Trajnet++ [8] tackles the issue that various human trajectory prediction models demonstrate their methods in a different subset of benchmark datasets. To evaluate them on the same set of trajectory data, Trajnet++ introduces their own benchmark. Especially they focus on generating data with sufficient human interaction in order to evaluate the capacity of each model in predicting plausible trajectories without collisions with other pedestrians. To measure the collisions, they suggest new metrics; Collision1 and Collision2. Collision1 is to compute the collision rate between primary pedestrian and the neighbors in the predicted future scene. Collision2 is to compute the collision rate between the primary pedestrian's prediction and the neighbors in the ground-truth future.

In this paper, we further investigate the effects of environment for realistic future trajectory prediction. We suggest our benchmark by varying the environments as well as by varying the number of agents. We also provide the new evaluation metric to check the performance of prior works in this new environment conditioned benchmark. Moreover, we propose to report min / mean / max of ADE / FDE so that we can evaluate the multi-modal models.

Dataset	Split	Total # of Scenarios	Min. # of Interacting Agents	Mean # of Interacting Agents	Max. # of Interacting Agents
A2A	Train	15336	2.25	9.35	12.72
A2A	Val.	15628	2.30	9.36	12.78
A2A	Test	7734	2.42	9.27	12.42
A2E	Train	397	8.09	29.76	54.08
A2E	Val.	128	8.27	30.19	54.47
A2E	Test	109	9.00	30.06	53.57

Table 1: This table reports the basic statistics of the training/validation/testing splits of the A2A and A2E datasets.

A.4 A2A and A2E Dataset Statistics

Table 1 shows a set of statistics for the A2A and A2E Datasets. For each split of each dataset, we report the total number of scenarios, the minimum number of potentially interacting agents present in a scenario averaged across all scenarios in the split, the mean number of interacting agents averaged across the scenarios, and the maximum number of interacting agents averaged across the scenarios. The average scenario in A2A has a considerably lower number of potentially interacting agents than the average scenario in A2E. This explains why training on A2E has a tendency to increase the collision-free likelihood (Tab. 1), i.e., models must learn collision avoidance under more challenging circumstances with more agents.

A.5 Additional Results on A2A and A2E Combined

Tables 2 and 3 report the remaining test results on the full **A2X** datasets (i.e., A2A combined with A2E) corresponding to Tables 1 and 2 in the Main Text. The results are consistent with the observations made in the Section 5. Since A2A consists of more scenarios than A2E, Tables 2 and 3 show similar results to the testing on A2A from Tables 1 and 2 in the Main Text.

Test	Model	Train	Accuracy Metrics		Realism Metrics						Decidab.
			ADE ↓ min / mean / max	FDE ↓ min / mean / max	Length	Speed mean / max	Accel. mean / max	ACFL	ECFL	%Diff. ↓	MVE ↓
A2A + A2E	GT	N/A	0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00	4.50	1.02 / 1.32	0.28 / 1.01	0.95	1.00	0	0.00
	SGAN	A2A	**0.36** / 0.76 / 1.49	**0.61** / 1.60 / 3.32	**4.29**	**0.98** / 1.44	0.09 / **0.55**	0.30	**0.97**	**48**	0.90
		A2E	2.09 / 2.35 / 2.69	3.80 / 4.42 / 5.28	3.22	0.73 / 1.38	**0.12** / 0.39	**0.57**	0.97	51	**0.70**
		Both	0.36 / **0.73** / **1.34**	0.63 / **1.53** / **2.97**	4.19	0.95 / **1.33**	0.06 / 0.33	0.33	0.97	51	0.83
	PECN	A2A	**0.62** / **0.65** / **0.68**	**1.12** / **1.27** / **1.44**	4.55	1.03 / 2.12	**0.37** / **3.37**	0.57	0.98	**60**	**0.07**
		A2E	1.20 / 1.22 / 1.25	1.75 / 1.92 / 2.11	**4.57**	**1.04** / 4.06	1.09 / 8.47	**0.59**	**0.98**	165	0.09
		Both	0.71 / 0.73 / 0.76	1.40 / 1.54 / 1.69	4.83	1.10 / 2.60	0.48 / 4.50	0.57	0.98	90	0.10
	T++	A2A	**0.22** / 0.67 / 1.88	**0.41** / 1.53 / 4.24	**4.45**	**1.01** / 2.37	**0.37** / 3.17	0.22	**0.96**	47	**1.08**
		A2E	0.53 / 1.01 / 1.70	1.07 / 2.19 / **3.76**	4.30	0.98 / **1.78**	0.29 / **2.13**	**0.24**	0.98	**46**	1.32
		Both	0.22 / **0.63** / **1.72**	0.42 / **1.45** / 3.93	4.42	1.00 / 2.25	0.34 / 2.91	0.23	0.97	47	1.11

Table 2: This table showcases the evaluation results of Social GAN (SGAN), PECNet (PECN), and Trajectron++ (T++) after training on either A2A, A2E, or both A2A and A2E and testing on A2A and A2E combined. For every metric in a testing set, the best value has been made bold for each model.

Test	Model	Train	Accuracy Metrics		Realism Metrics						Decidab.
			ADE ↓ min = mean = max	FDE ↓ min = mean = max	Length	Speed mean / max	Accel. mean / max	ACFL	ECFL	%Diff. ↓	MVE ↓
A2A + A2E	GT	N/A	0.00	0.00	4.50	1.02 / 1.32	0.29 / 1.04	0.95	1.00	0	0.00
	SGAN	A2A	0.90	1.98	**4.31**	**0.98** / 1.21	0.16 / **0.41**	0.69	**0.99**	**37**	0.00
		A2E	2.50	4.84	**3.78**	0.86 / **1.32**	**0.20** / 0.36	**0.79**	0.97	40	0.00
		Both	**0.86**	**1.86**	4.26	0.97 / 1.16	0.11 / 0.23	0.70	0.99	40	0.00
	PECN	A2A	**0.64**	**1.26**	**4.48**	**1.02** / **1.55**	**0.33** / **1.76**	0.66	**0.98**	**58**	0.00
		A2E	1.24	1.98	4.37	0.99 / 3.17	0.99 / 6.18	**0.68**	0.98	165	0.00
		Both	0.74	1.52	4.73	1.07 / 2.11	0.43 / 3.11	0.64	0.98	88	0.00
	T++	A2A	0.81	1.83	**4.53**	**1.03** / **1.31**	0.44 / 0.99	**0.65**	**0.99**	**26**	0.00
		A2E	1.02	2.21	4.56	1.04 / 1.32	**0.42** / 0.97	0.63	0.98	30	0.00
		Both	**0.80**	**1.82**	4.54	1.03 / 1.31	0.44 / **1.00**	0.65	0.99	26	0.00

Table 3: This table reports the results of MMC on each of the 9 trained models. On average, MMC produces predictions that are consistently better than the worse case prediction prior to MMC. Only one value is reported for ADE and FDE, because the minimum, mean, and maximum are equal when $k = 1$. The MVE is always 0 when $k = 1$.

A.6 Visualizations of Agent-to-Environment Interaction Scenarios

The A2E data is composed of bottleneck and pathfinding scenarios. Figure 1 illustrates the full trajectories of all agents until they have passed the bottleneck to the left of the room. When either the bottleneck shrinks or the agent density per square meter increases, congestion increases at the bottleneck, slowing the agents down and producing long-term A2A and A2E interactions. Figure 2 showcases more isolated A2E interactions, where a single agent is moving to a random destination in a large, complex environments. The non-navigable regions of the environment cause the A2E interactions, which would not exist if there was a straight path to the destination.

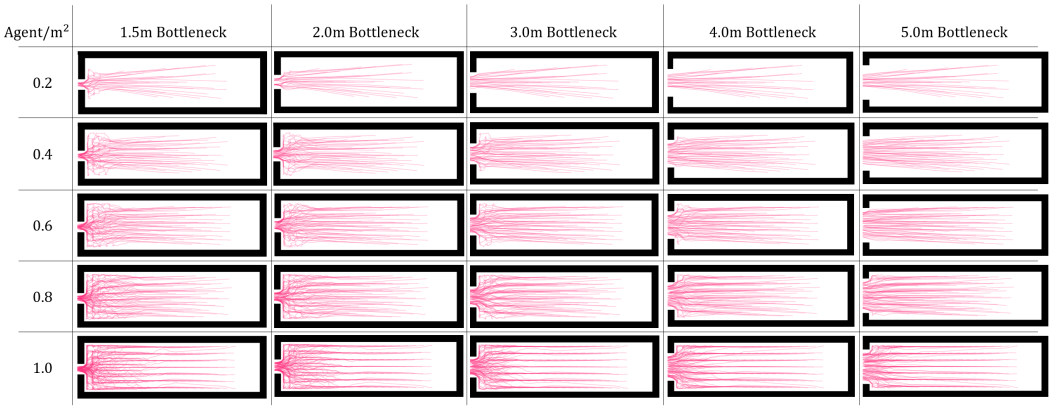

Figure 1: The above images correspond to bottleneck scenarios in A2E. The bottleneck scenarios have 5 different agent density levels with 5 different bottleneck sizes. All agents are spawned at the right side of the room and move left to exit through the bottleneck. The red lines represent the full trajectories of agents.

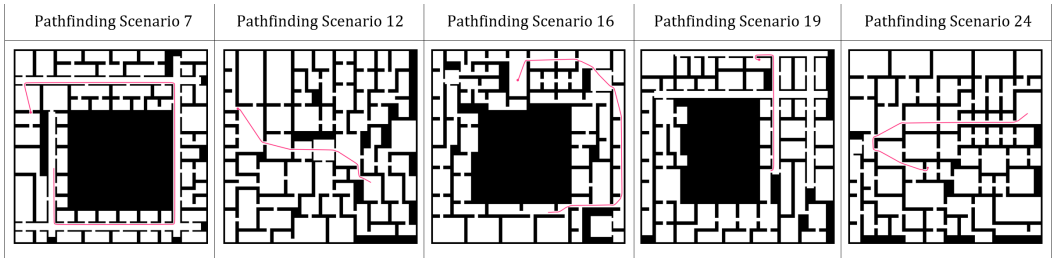

Figure 2: The above images correspond to pathfinding scenarios in A2E, in which a single agent moves toward a random destination. The red line represents the full trajectory of the agent. The small circle at one end of the red line indicates the destination.

A.7 Qualitative Evaluation

Models Social GAN [5], PECNet [11], and Trajectron++ [15] were each trained on either A2A, A2E, or both A2A and A2E, resulting in a total of 9 trained models. Figures 3 through 8 visualize the $k = 20$ predictions per agent (in blue) for each model on the same scenario. The ground truth (in magenta) is the same for each of the nine models per figure. Each row of the figures corresponds to a particular model, and each column corresponds to a particular training set. We observe that Social GAN has a tendency to overfit when trained on A2E as indicated by the leftward bias of predicted trajectories on A2A scenarios. All of the A2E training scenarios feature movement from right to left, but other models such as PECNet and Trajectron++ do not suffer from the same overfitting. PECNet shows the highest decidability among the three models by a significant margin despite predicting 20 paths per agent. This results in a lower number of instances where agents collide with walls. However, this appears to result in cases where the model has high certainty in the incorrect movement of an agent. Trajectron++ seems to strike a balance between Social GAN and PECNet. Its predictions diverge like Social GAN and unlike PECNet, but similar to PECNet, it does not overfit as severely as Social GAN when trained on A2E.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016.

[2] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-Aware Large-Scale Crowd Forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2203–2210, 2014.

[3] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret. Wildtrack: A Multi-Camera HD Dataset for Dense Unscripted Pedestrian Detection.

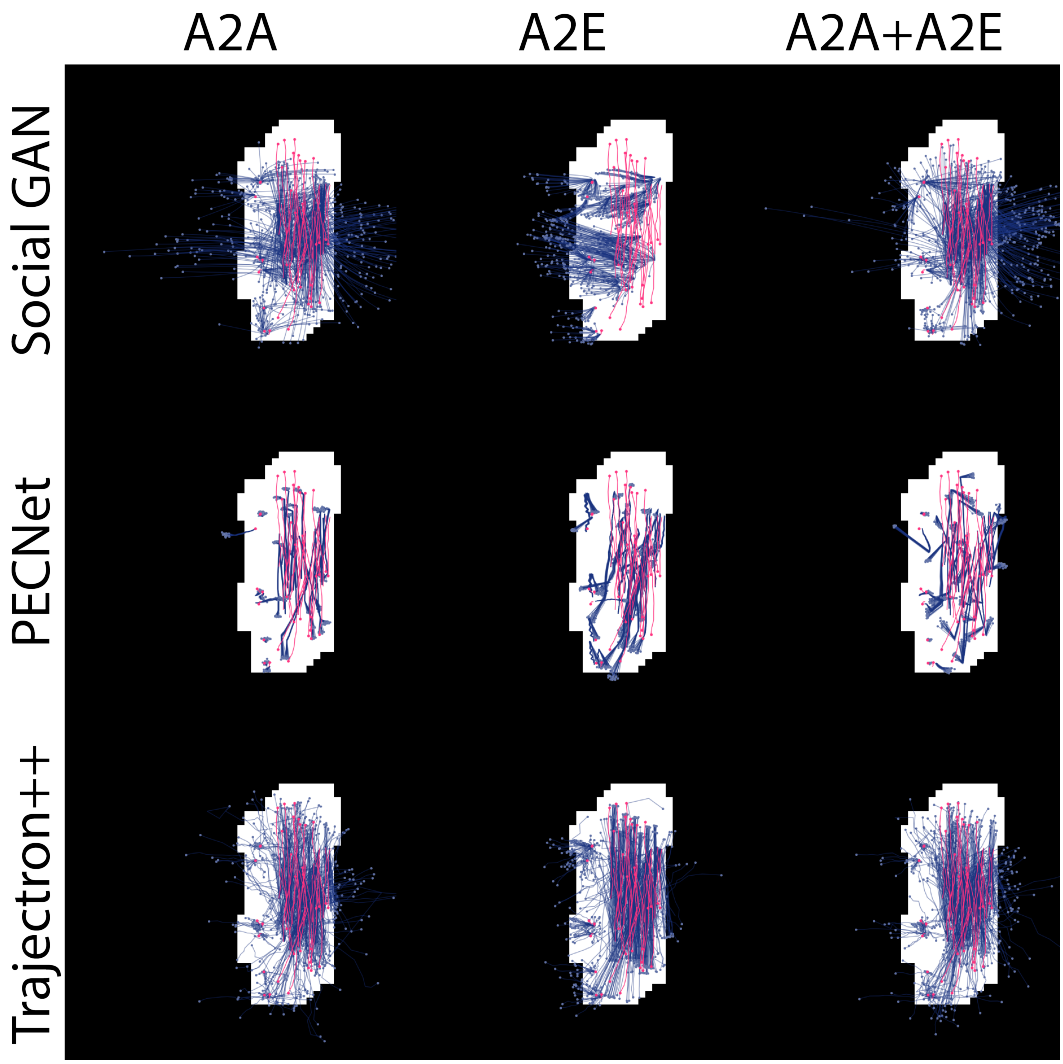

Figure 3: The above table of images shows the predictions (blue) and ground truth (magenta) for 9 models tested on the same scenario. The model and training dataset for a particular image is given by the row and column it belongs to respectively.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5039, 2018.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, page 2672–2680, 2014.

[5] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018.

[6] B. Ivanovic and M. Pavone. The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2375–2384, 2019.

[7] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations (ICLR), Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[8] P. Kothari, S. Kreiss, and A. Alahi. Human Trajectory Forecasting in Crowds: A Deep Learning Perspective. *IEEE Transactions on Intelligent Transportation Systems*, 2021. doi: 10.1109/TITS.2021.3069362.

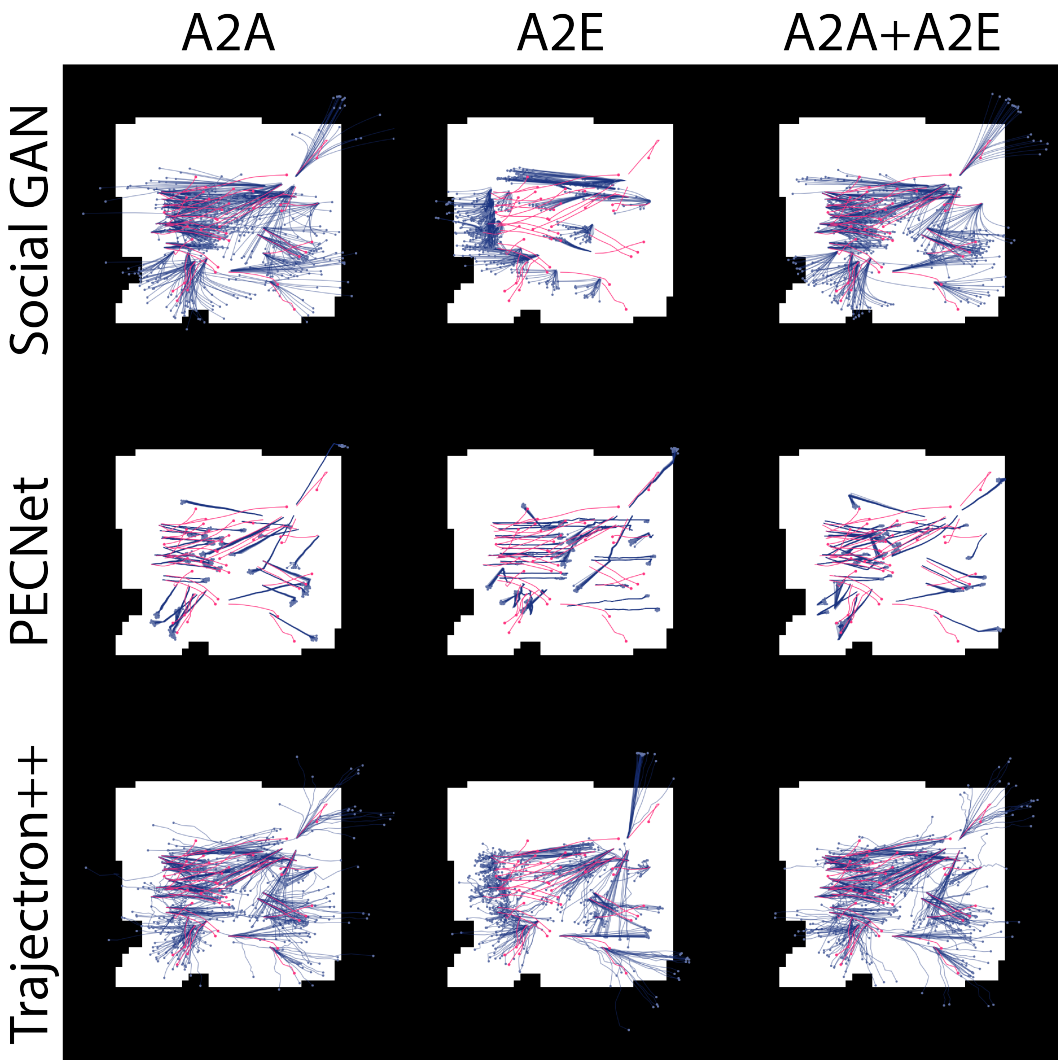

Figure 4: The above table of images shows the predictions (blue) and ground truth (magenta) for 9 models tested on the same scenario. The model and training dataset for a particular image is given by the row and column it belongs to respectively.

[9] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by Example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[10] K. Mangalam, Y. An, H. Girase, and J. Malik. From Goals, Waypoints Paths To Long Term Human Trajectory Forecasting. *arXiv preprint arXiv:2012.01526*, 2020.

[11] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon. It is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction. *arXiv preprint arXiv:2004.02025*, 2020.

[12] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll Never Walk Alone: Modeling Social Behavior for Multi-Target Tracking. In *2009 IEEE 12th International Conference on Computer Vision (CVPR)*, pages 261–268. IEEE, 2009.

[13] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning Social Etiquette: Human Trajectory Understanding in Crowded Scenes. In *European Conference on Computer Vision (ECCV)*, pages 549–565. Springer, 2016.

[14] A. Sadeghian, V. Kosaraju, A. Gupta, S. Savarese, and A. Alahi. Trajnet: Towards a benchmark for human trajectory prediction. *arXiv preprint*, 2018.

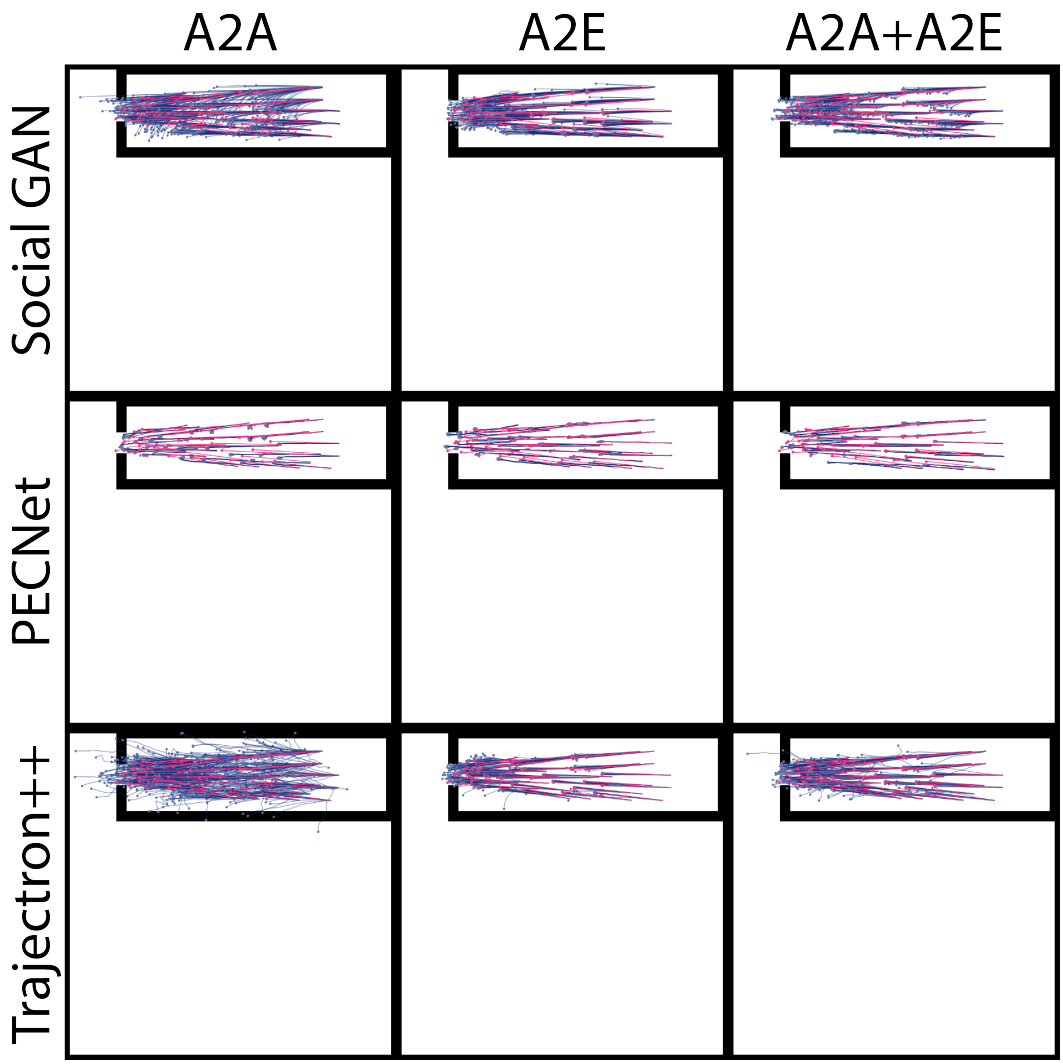

Figure 5: The above table of images shows the predictions (blue) and ground truth (magenta) for 9 models tested on the same scenario. The model and training dataset for a particular image is given by the row and column it belongs to respectively.

[15] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision (ECCV)*, pages 683–700. Springer, 2020.

[16] K. Sohn, H. Lee, and X. Yan. Learning Structured Output Representation using Deep Conditional Generative Models. In *Neural Information Processing Systems (NIPS)*, 2015.

[17] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha. Reciprocal n-Body Collision Avoidance. In *Robotics Research*, pages 3–19. Springer, 2011.

[18] A. Vemula, K. Muelling, and J. Oh. Social Attention: Modeling Attention in Human Crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4601–4607, 2018.

[19] Z. Yan, T. Duckett, and N. Bellotto. Online Learning for Human Classification in 3D LiDAR-based Tracking. In *In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, Canada, September 2017.

[20] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu. Multi-Agent Tensor Fusion for Contextual Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12118–12126, June 2019.

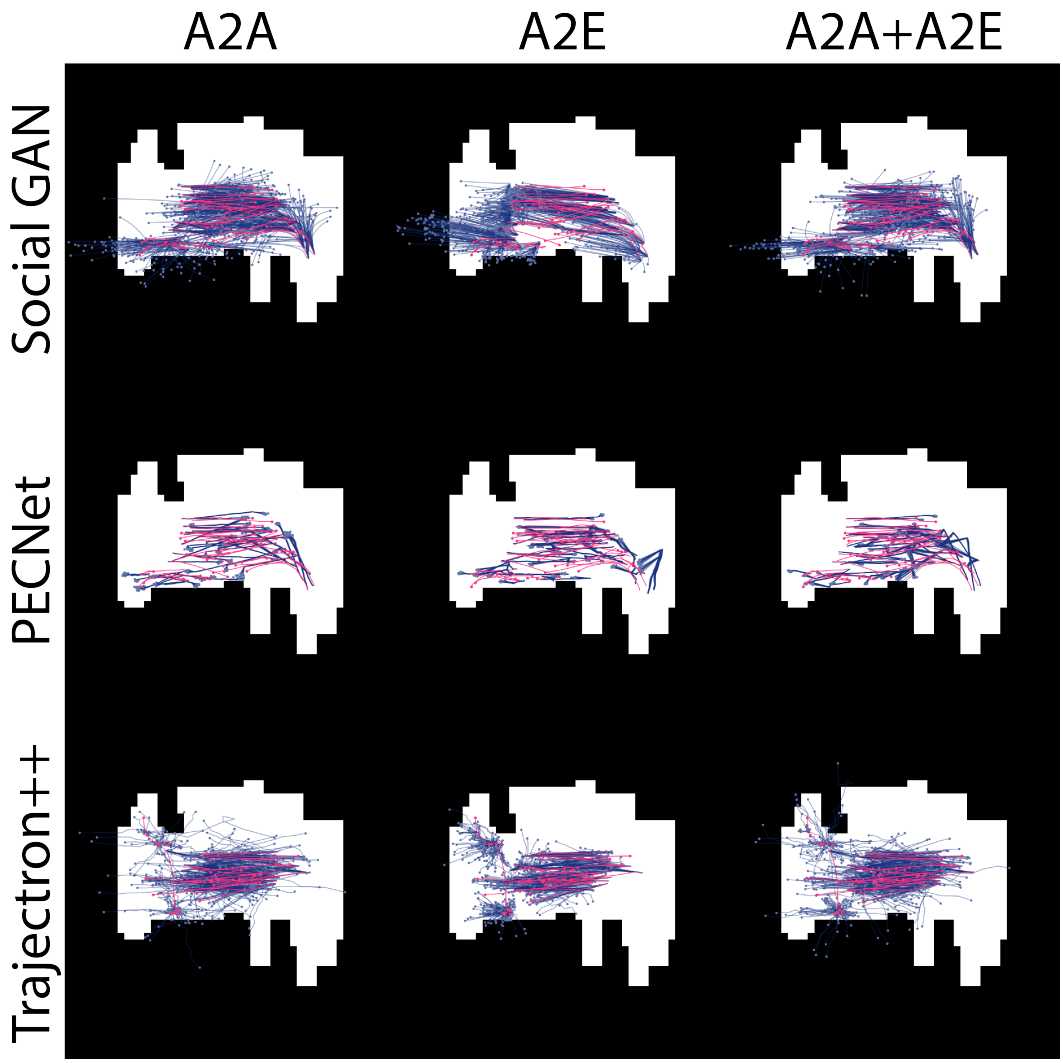

Figure 6: The above table of images shows the predictions (blue) and ground truth (magenta) for 9 models tested on the same scenario. The model and training dataset for a particular image is given by the row and column it belongs to respectively.

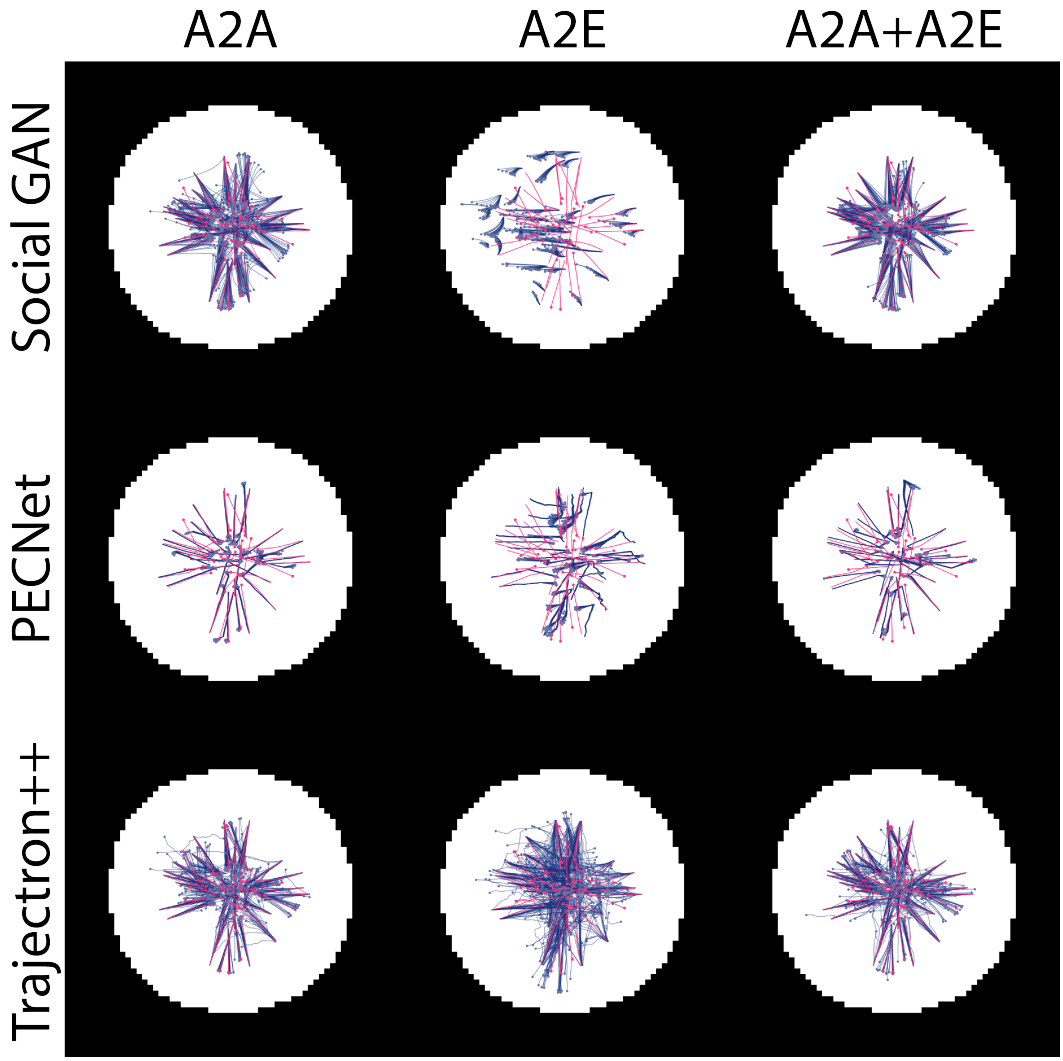

Figure 7: The above table of images shows the predictions (blue) and ground truth (magenta) for 9 models tested on the same scenario. The model and training dataset for a particular image is given by the row and column it belongs to respectively.

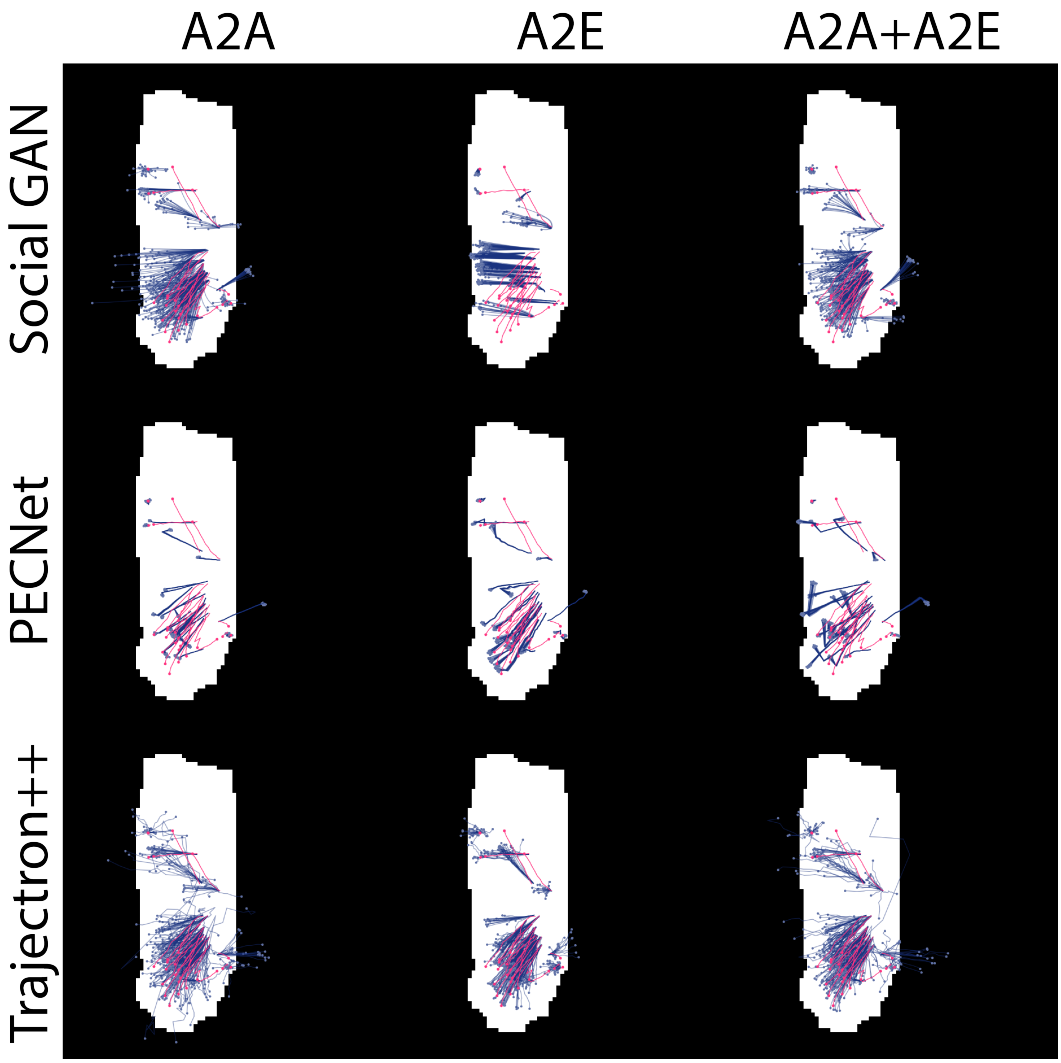

Figure 8: The above table of images shows the predictions (blue) and ground truth (magenta) for 9 models tested on the same scenario. The model and training dataset for a particular image is given by the row and column it belongs to respectively.