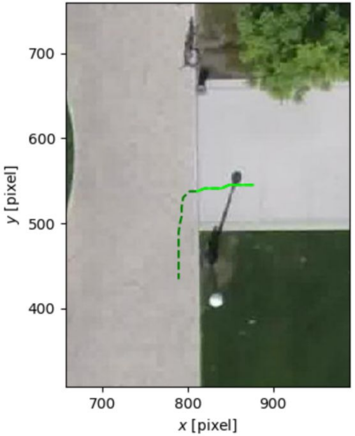
Human intent prediction

Paper: Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks

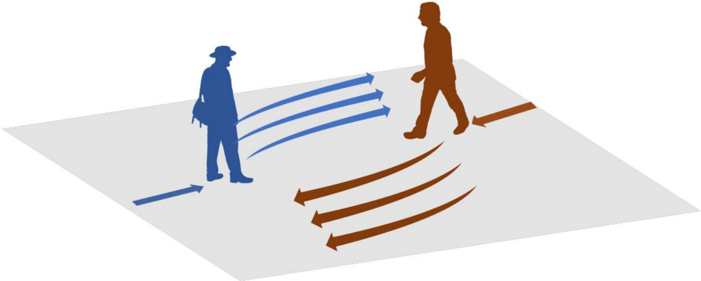
Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, Alexandre Alahi

Presented by Ajay Jain

Reading group | Feb 25th, 2018



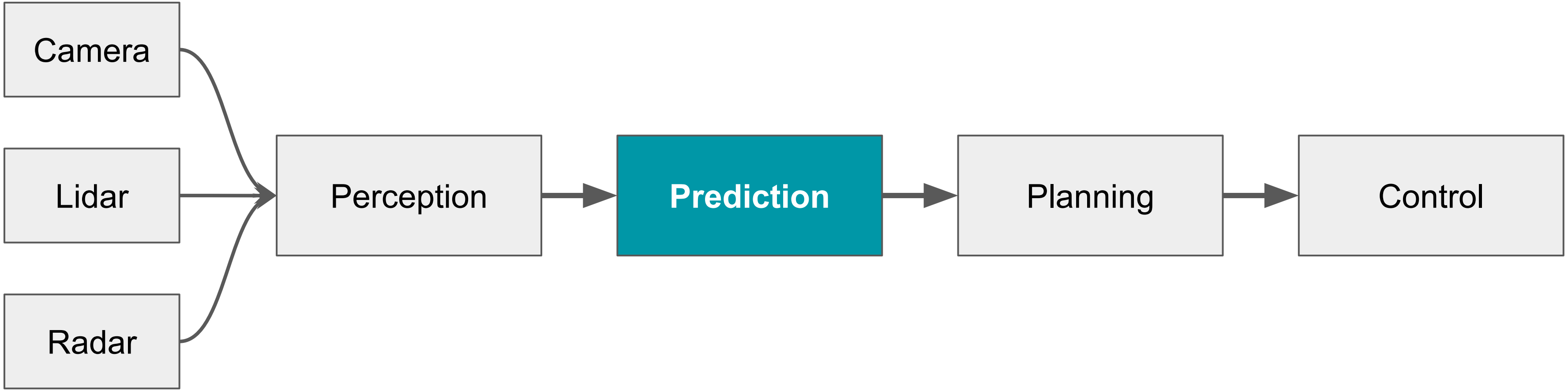
How important is the local map? [2]



Where will these people go? [1]

How do people interact?

Why predict where people and cars will go?



High level self-driving vehicle stack

Guiding questions

What is the purpose of predicting where people and vehicles will go?

Do we care more about predicting the most plausible behaviors, or error minimizing behaviors?

How to evaluate plausible behavior prediction?

What do humans look for when predicting where others will go?

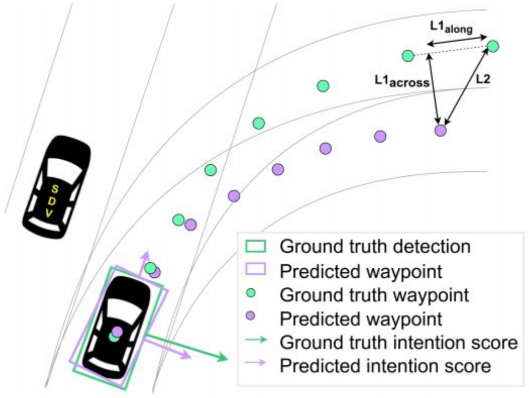
Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks

- Priorities:
  - **Interpersonal reasoning**: Jointly model dependencies
  - **Social acceptability** of prediction
  - **Multimodal** future predictions

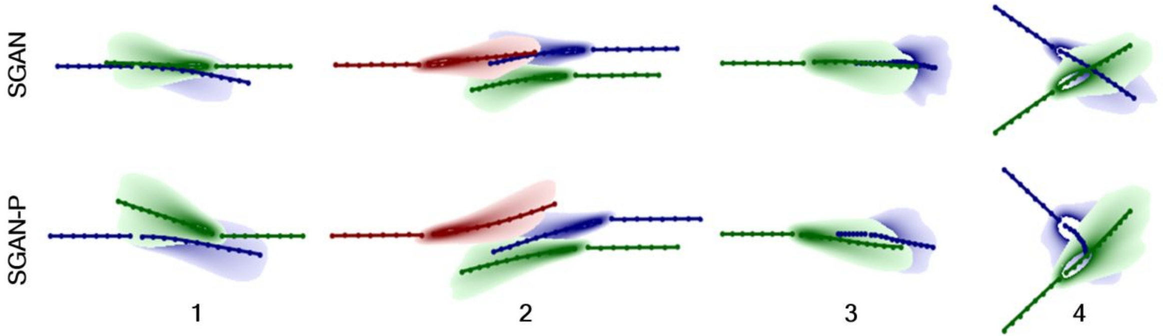
- Criticisms of prior work:
  - Don’t model interactions between *all* pedestrians
  - Learn an “average behavior” by minimizing euclidean distance (L2)

- Solution
  - “Pooling” across all people in the scene
  - Don’t use L2. Instead, use an adversarial loss and a variety loss

Predicting trajectories: output parameterization



Vehicle trajectory prediction [3]



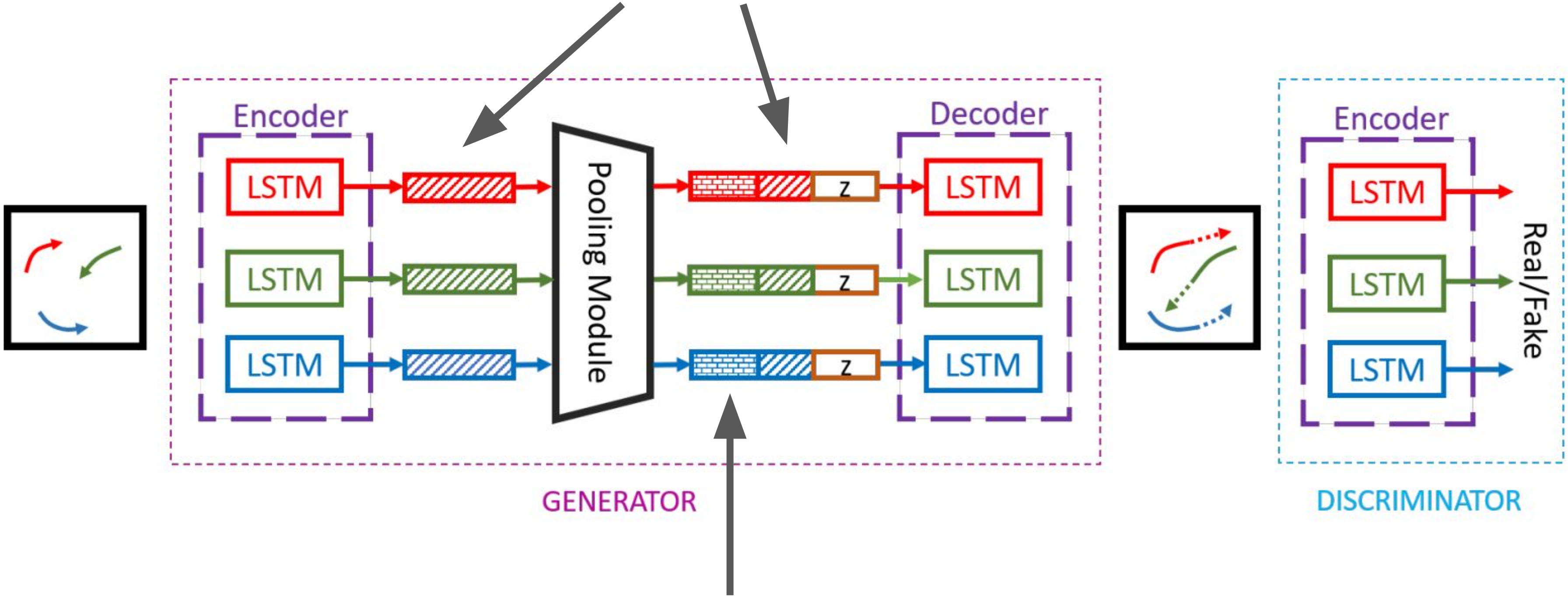
Social GAN trajectory prediction

Trajectory distributions are aggregated from 300 samples

$$\mathcal{L}_{\mathrm{MSE}} = ||Y_i - \hat{Y}_i||_2$$

Architecture

Per-pedestrian context (not pooled)



Pooled multi-pedestrian context

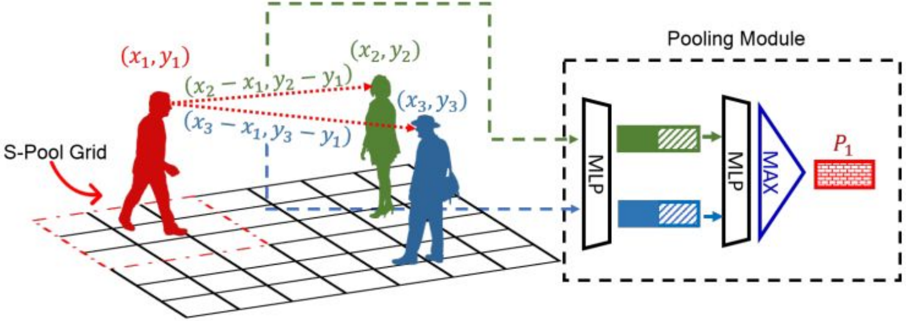
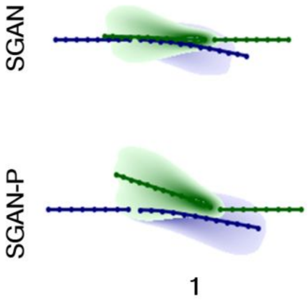


Figure 3: Comparison between our pooling mechanism (red dotted arrows) and Social Pooling [1] (red dashed grid) for the red person. Our method computes relative positions between the red and all other people; these positions are concatenated with each person's hidden state, processed independently by an MLP, then pooled elementwise to compute red person's pooling vector P_1. Social pooling only considers people inside the grid, and cannot model interactions between all pairs of people.



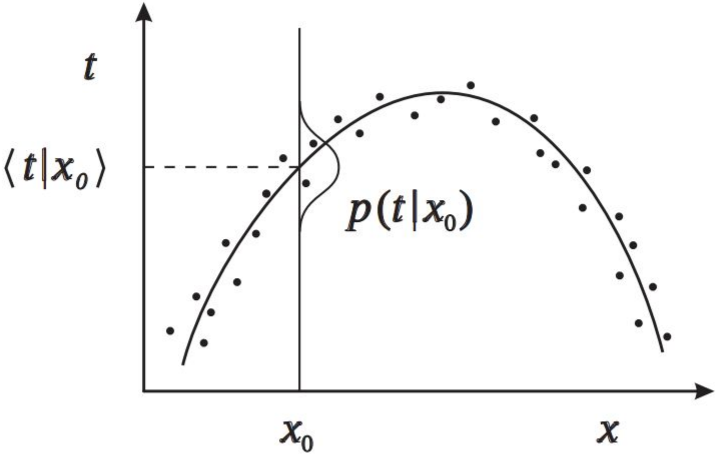
Sum-of-squares error minimized by conditional average

The function that minimizes squared error is the expectation of the true conditional distribution (conditioned on the observation)

From *Mixture Density Networks* (1994)

The “conditional average” may be infeasible

No multimodality!

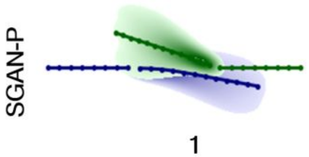


$$\langle Q \mid \mathrm{x}\rangle \equiv \int Q(\mathrm{t})\, p(\mathrm{t} \mid \mathrm{x})\, d\mathrm{t}$$

$$\hat{y}(x) = E[y|x]$$

MACHINE INTELLIGENCE COMMUNITY

Variety loss



Social GAN trajectory prediction
Trajectory distributions are
aggregated from 300 samples

$$\mathcal{L}_{\text{MSE}} = ||Y_i - \hat{Y}_i||_2$$
$$\mathcal{L}_{\text{variety}} = \min_k ||Y_i - \hat{Y}_i^{(k)}||_2$$

Social GAN loss

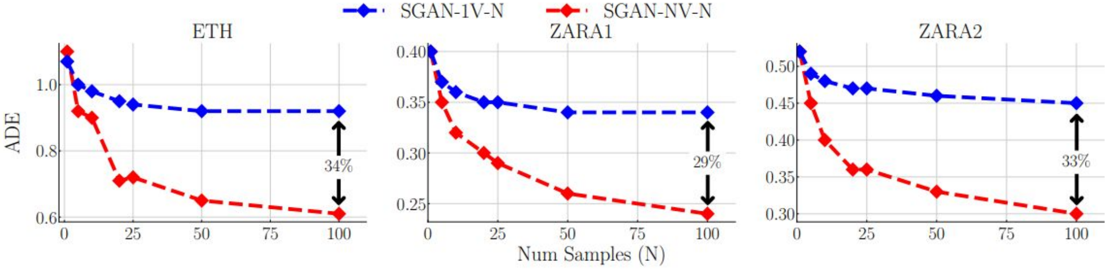


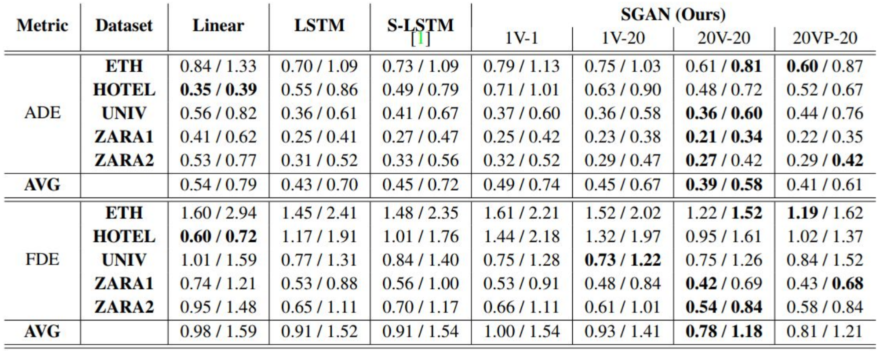
Figure 4: Effect of variety loss. For SGAN-1V-N we train a single model, drawing one sample for each sequence during training and N samples during testing. For SGAN-NV-N we train several models with our variety loss, using N samples during both training and testing. Training with the variety loss significantly improves accuracy.

Results

Can control high-level actions by choosing appropriate noise z

Pooling does not reduce error, but has a qualitative effect

Variety loss helps if many samples are allowed



Metric	Dataset	Linear	LSTM	S-LSTM [1]	SGAN (Ours)			
					1V-1	1V-20	20V-20	20VP-20
ADE	**ETH**	0.84 / 1.33	0.70 / 1.09	0.73 / 1.09	0.79 / 1.13	0.75 / 1.03	0.61 / **0.81**	**0.60** / 0.87
	HOTEL	**0.35 / 0.39**	0.55 / 0.86	0.49 / 0.79	0.71 / 1.01	0.63 / 0.90	0.48 / 0.72	0.52 / 0.67
	UNIV	0.56 / 0.82	0.36 / 0.61	0.41 / 0.67	0.37 / 0.60	0.36 / 0.58	**0.36 / 0.60**	0.44 / 0.76
	ZARA1	0.41 / 0.62	0.25 / 0.41	0.27 / 0.47	0.25 / 0.42	0.23 / 0.38	**0.21 / 0.34**	0.22 / 0.35
	ZARA2	0.53 / 0.77	0.31 / 0.52	0.33 / 0.56	0.32 / 0.52	0.29 / 0.47	**0.27** / 0.42	0.29 / **0.42**
AVG		0.54 / 0.79	0.43 / 0.70	0.45 / 0.72	0.49 / 0.74	0.45 / 0.67	**0.39 / 0.58**	0.41 / 0.61
FDE	**ETH**	1.60 / 2.94	1.45 / 2.41	1.48 / 2.35	1.61 / 2.21	1.52 / 2.02	1.22 / **1.52**	**1.19** / 1.62
	HOTEL	**0.60 / 0.72**	1.17 / 1.91	1.01 / 1.76	1.44 / 2.18	1.32 / 1.97	0.95 / 1.61	1.02 / 1.37
	UNIV	1.01 / 1.59	0.77 / 1.31	0.84 / 1.40	0.75 / 1.28	**0.73 / 1.22**	0.75 / 1.26	0.84 / 1.52
	ZARA1	0.74 / 1.21	0.53 / 0.88	0.56 / 1.00	0.53 / 0.91	0.48 / 0.84	**0.42** / 0.69	0.43 / **0.68**
	ZARA2	0.95 / 1.48	0.65 / 1.11	0.70 / 1.17	0.66 / 1.11	0.61 / 1.01	**0.54 / 0.84**	0.58 / 0.84
AVG		0.98 / 1.59	0.91 / 1.52	0.91 / 1.54	1.00 / 1.54	0.93 / 1.41	**0.78 / 1.18**	0.81 / 1.21

Table 1: Quantitative results of all methods across datasets. We report two error metrics Average Displacement Error (ADE) and Final Displacement Error (FDE) for $t_{pred} = 8$ and $t_{pred} = 12$ (8 / 12) in meters. Our method consistently outperforms state-of-the-art S-LSTM method and is especially good for long term predictions (lower is better).

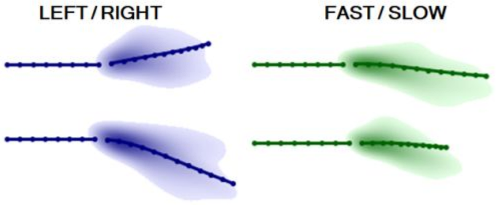


Figure 7: Latent Space Exploration. Certain directions in the latent manifold are associated with direction (left) and speed (right). Observing the same past but varying the input z along different directions causes the model to predict trajectories going either right/left or fast/slow on average.

References

[1] Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. https://arxiv.org/pdf/1803.10892.pdf. Code: https://github.com/agrimgupta92/sgan

[2] An Evaluation of Trajectory Prediction Approaches and Notes on the TrajNet Benchmark. https://arxiv.org/pdf/1805.07663.pdf

[3] Mixture Density Networks. https://publications.aston.ac.uk/373/1/NCRG_94_004.pdf

[4] IntentNet: Learning to Predict Intention from Raw Sensor Data. http://www.cs.toronto.edu/~wenjie/papers/intentnet_corl18.pdf