Evaluation Challenges in Social Robot Navigation

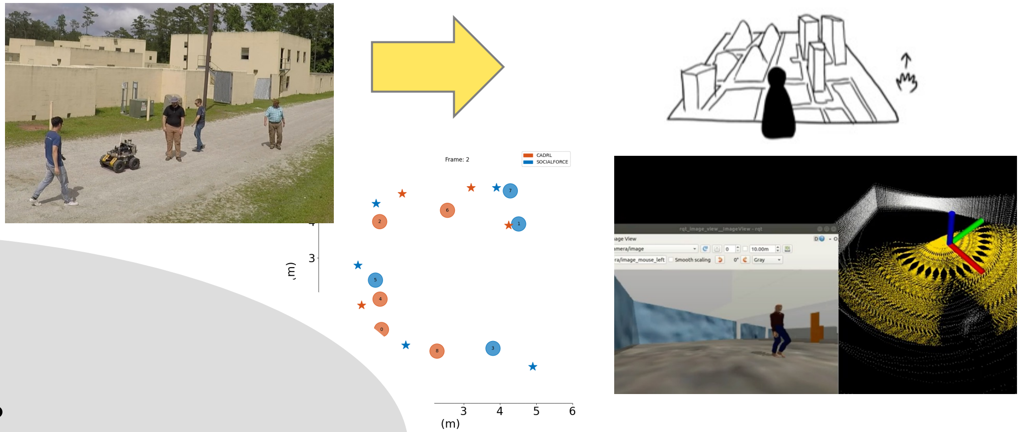
Jean Oh
The Robotics Institute
Carnegie Mellon University

roBot Intelligence Group (BIG)

Jean Oh

Social interaction

Imagination



Robot Intelligence

Creativity

Language

"Measure what is measurable, and make measurable what is not so."

– Galileo Galilei

What is good performance metric for a cleaning robot?

- Total amount of dust collected
- Area covered

All but perfect metrics can be exploited

When you design performance metric

- Be careful what you wish for
- What you get is what you ask for

Optimal solution is only optimal according to some objective function

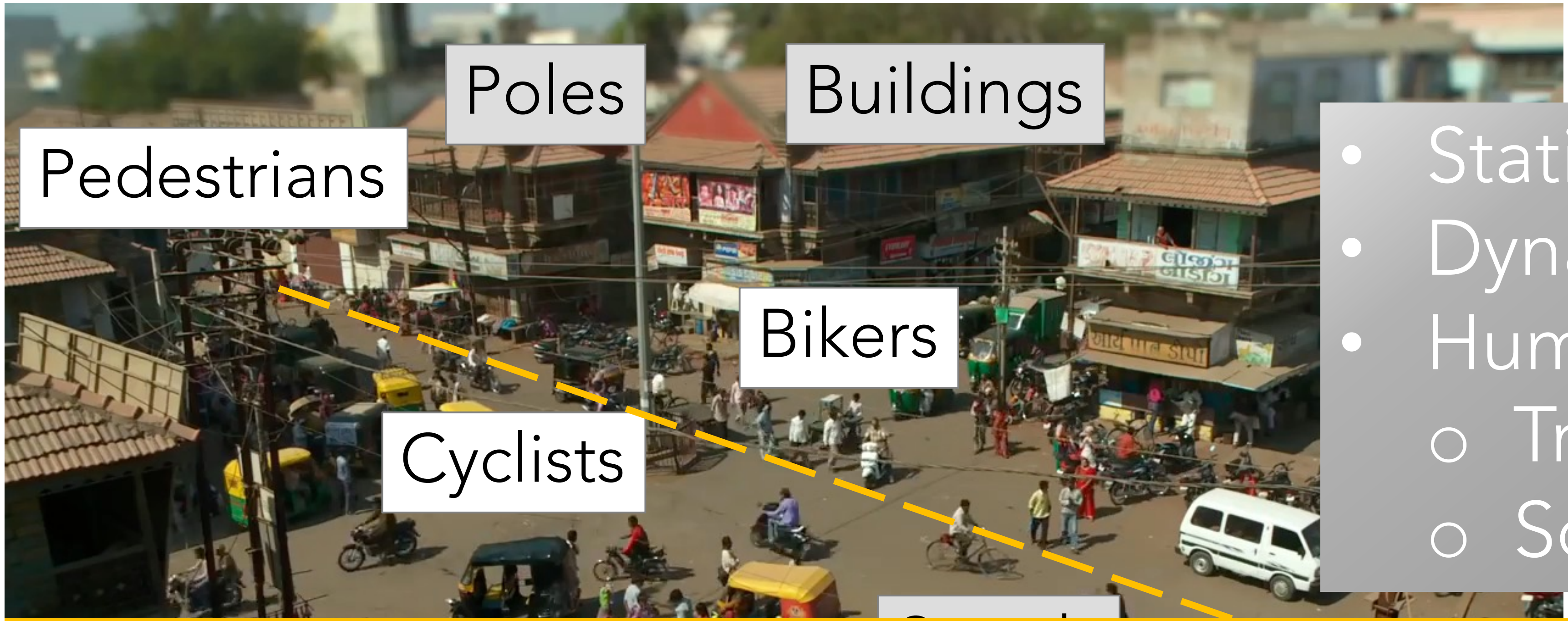
Autonomous navigation in a real world

- Static obstacles

- Main objective is to detect & avoid collision

https://www.youtube.com/watch?v=KnPiP9PkLAs

Autonomous navigation in a real world



- Static obstacles
- Dynamic obstacles
- Humanmade rules
  - Traffic rules
  - Social norms

- Main objective is to detect, track, & avoid collision
- From passive reaction to proactive coordination

https://www.youtube.com/watch?v=KnPiP9PkLAs

Safe & Seamless Close-proximity Operation of Manned and Unmanned Aircraft in Shared Space

Jay Patrikar, Ian Higgins, Sourish Ghosh, Jimin Sun, Jasmine Aloor, Joao Dantas, Brady Moon, Parv Kapoor, Ingrid Navarro, Benjamin Stoler, Rohan Baijal, Milad Hamidi

PIs: Sebastian Scherer (basti@cmu.edu)
Jean Oh (jeanoh@cmu.edu)

The Robotics Institute, Carnegie Mellon University

Social Robot Navigation

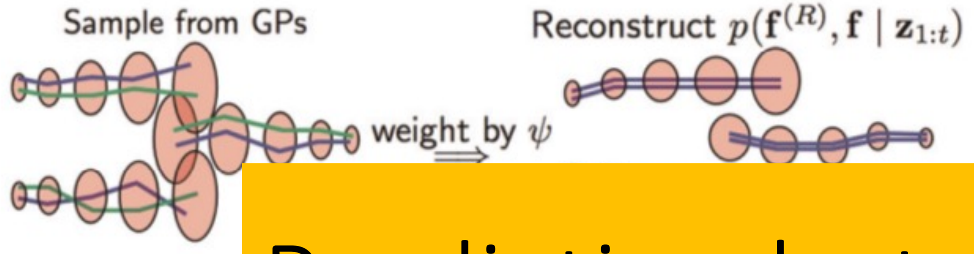
Research Question:

How can we make an autonomous vehicle navigate seamlessly with other vehicles in a complex environment?

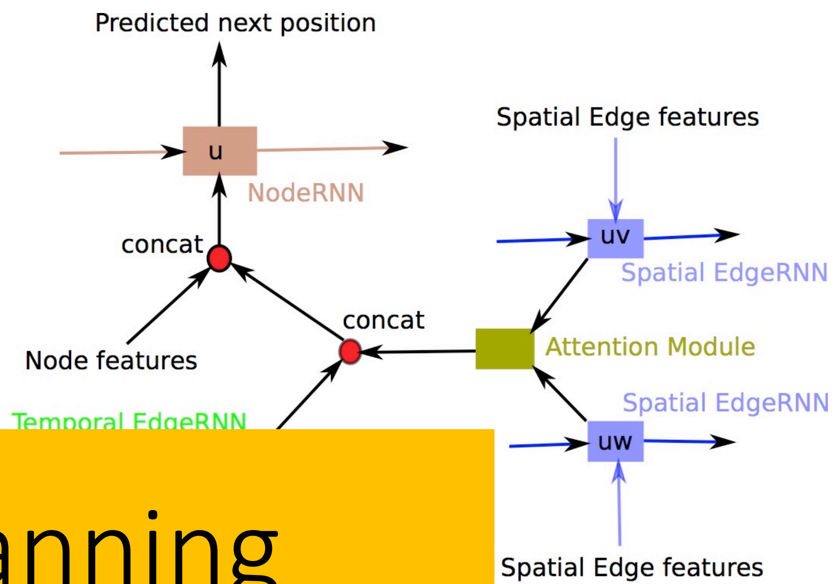
Prediction vs. Navigation vs. Social Navigation

	Goal-oriented objectives	Social objectives	Environmental context / physical constraints
Static Navigation	Yes	Safety	Traversability, static obstacles
Trajectory prediction	No	Naturalness	Dynamic obstacles
Social Navigation	Yes	Safety / Norm / Comfort / Naturalness	Static + Dynamic obstacles

Pedestrian prediction



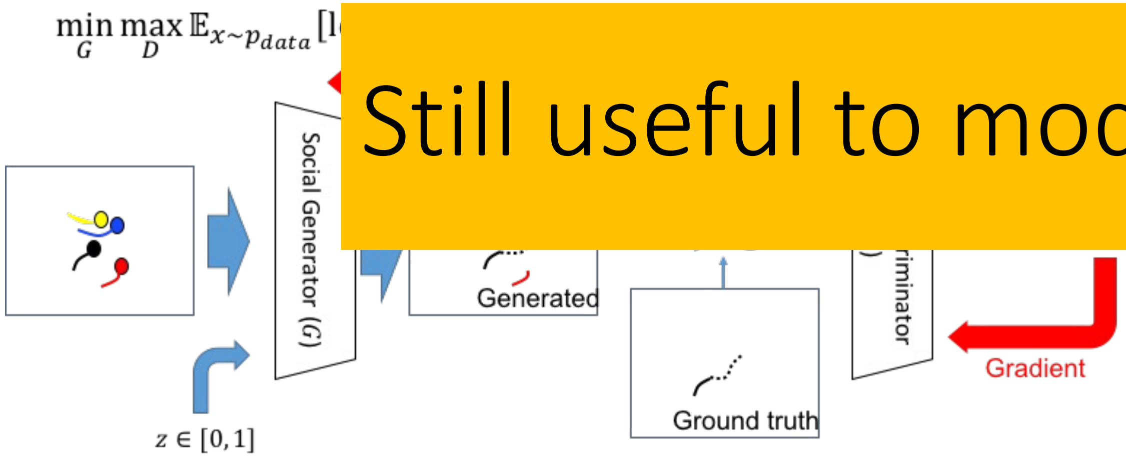
Interacting Gaus
IGP with learned



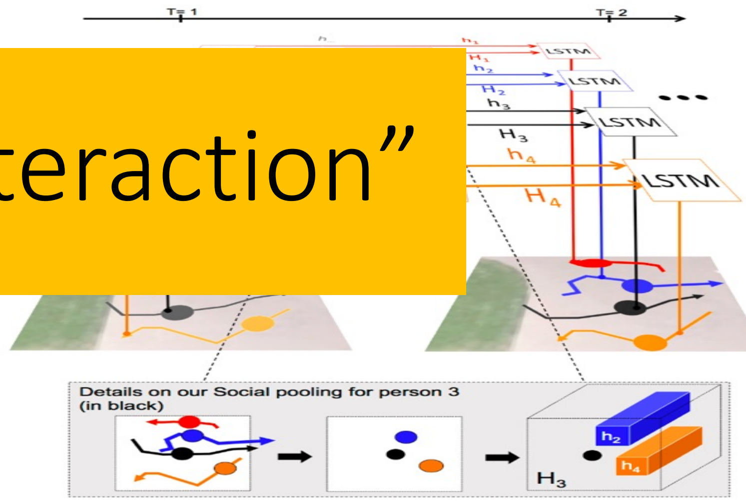
nula et al., 2018)

Prediction but not planning

Still useful to model "interaction"



Social GAN (Gupta et al., 2018)

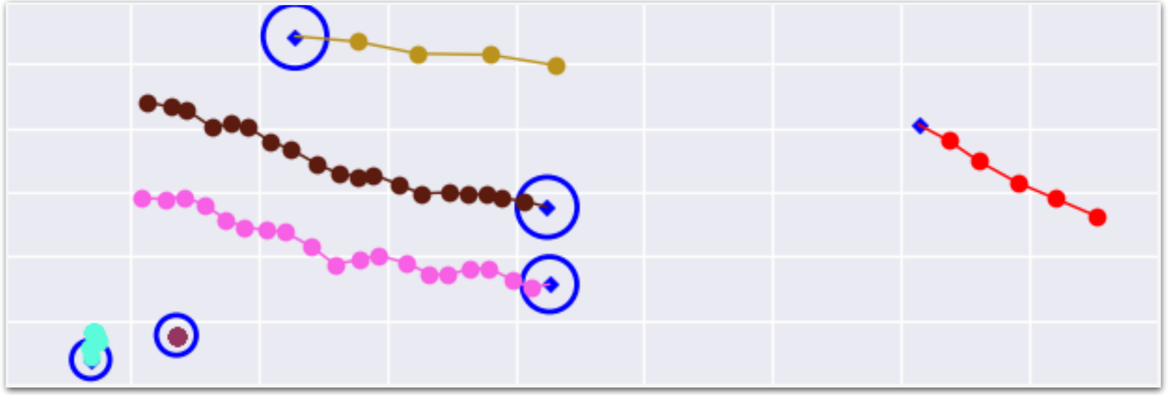


Social LSTM (Alahi et al., 2016)

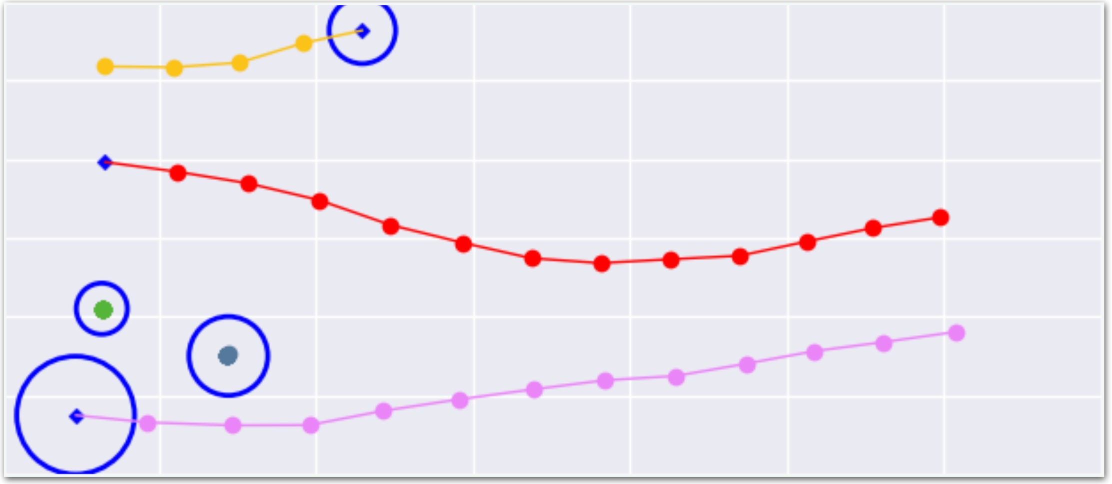
Survey article [Rudenko et al., 2020]

Social Attention: qualitative results

[Vemula et al., 2018]



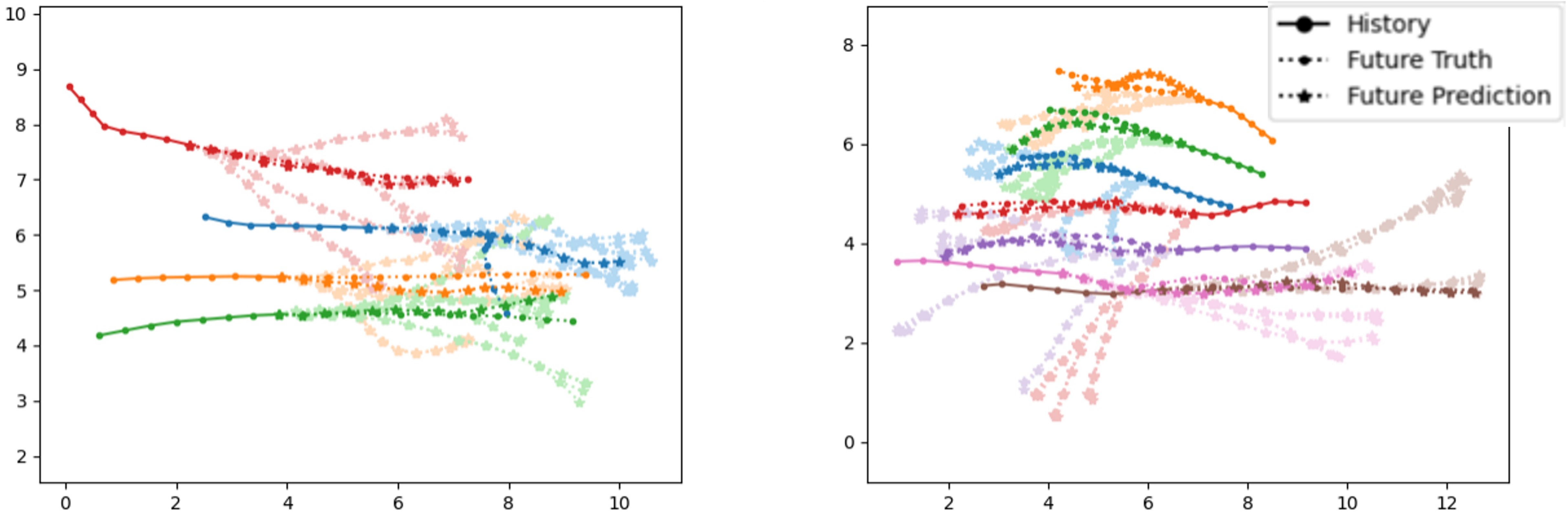
Learns to give equal relative importance to pedestrians far away to exert any influence



Learns to give high importance to agents with whom there might be a future collision, irrespective of their current proximity

A. Vemula, K. Muelling, J. Oh. Social Attention: Modeling Attention in Human Crowds. In Proc. of IEEE Conference on Robotics and Automation (ICRA), 2018 (Best Paper Award in Cognitive Robotics)

SPEC: Qualitative Analysis



D. Zhao and J. Oh. "Noticing Motion Patterns: Temporal CNN with a Novel Convolution Operator for Human Trajectory Prediction." IEEE Robotics and Automation Letters (RA-L), Special Issue on Long-Term Human Motion Prediction (2020).

Evaluation is challenging

Settings:

- Datasets: Recorded pedestrians
- Physical robot testing
- Simulation

Metrics: (Rudenko et al., 2020)

- Geometric metric
  - Average Displacement Error (ADE)
  - Final Displacement Error (FDE)
  - Modified Hausdorff Distance
- Probabilistic metrics
  - Negative log likelihood
  - Negative log loss
  - Prediction probability
  - mADE, mFDE
  - Cumulative probability

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Linear * [1]	1.33 / 2.94	0.39 / 0.72	0.82 / 1.59	0.62 / 1.21	0.77 / 1.48	0.79 / 1.59
SR-LSTM-2 * [30]	0.63 / 1.25	0.37 / 0.74	0.51 / 1.10	0.41 / 0.90	0.32 / 0.70	0.45 / 0.94
S-LSTM [1]	1.09 / 2.35	0.79 / 1.76	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.72 / 1.54
S-GAN-P [6]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
SoPhie [23]	0.70 / 1.43	0.76 / 1.67	0.54 / 1.24	0.30 / 0.63	0.38 / 0.78	0.54 / 1.15
CGNS [13]	**0.62** / 1.40	0.70 / 0.93	0.48 / 1.22	0.32 / 0.59	0.35 / 0.71	0.49 / 0.97
PIF [14]	0.73 / 1.65	**0.30 / 0.59**	0.60 / 1.27	0.38 / 0.81	0.31 / 0.68	0.46 / 1.00
STSGN [29]	0.75 / 1.63	0.63 / 1.01	0.48 / 1.08	0.30 / 0.65	**0.26** / 0.57	0.48 / 0.99
GAT [10]	0.68 / 1.29	0.68 / 1.40	0.57 / 1.29	**0.29** / 0.60	0.37 / 0.75	0.52 / 1.07
Social-BiGAT [10]	0.69 / 1.29	0.49 / 1.01	0.55 / 1.32	0.30 / 0.62	0.36 / 0.75	0.48 / 1.00
Social-STGCNN	0.64 / **1.11**	0.49 / 0.85	**0.44 / 0.79**	0.34 / **0.53**	0.30 / **0.48**	**0.44 / 0.75**

Table 2. ADE/FDE from (Mohamed et al., 2020)

Model	ETH	Hotel	Univ.	Zara1	Zara2	Ave.
Linear	1.33 / 2.94	0.39 / 0.72	0.82 / 1.59	0.62 / 1.21	0.77 / 1.48	0.79 / 1.59
S-LSTM[7]	1.09 / 2.35	0.79 / 1.76	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.72 / 1.54
SGAN(20VP20)[8]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
STSGN[17]	0.75 / 1.63	0.63 / 1.01	0.48 / 1.08	**0.30** / 0.65	**0.26** / 0.57	0.48 / 0.99
S-BiGAT[14]	0.69 / 1.29	0.49 / 1.01	0.55 / 1.32	**0.30** / 0.62	0.36 / 0.75	0.48 / 1.00
S-STGCNN[12]	0.64 / **1.11**	0.49 / 0.85	**0.44 / 0.79**	0.34 / **0.53**	0.30 / **0.48**	0.44 / **0.75**
Social-PEC	**0.61 / 1.11**	**0.31 / 0.52**	0.47 / 0.82	0.43 / 0.77	0.35 / 0.60	**0.43** / 0.76

Table 1. ADE/FDE from (Zhao & Oh, 2020)

Evaluation is challenging

Settings:
- Datasets: Recorded pedestrians
- Physical robot testing
- Simul

Metric
- Geom
  - A
  - Fi
  - M
- Proba
  - Negative log likelihood
  - Negative log loss
  - Prediction probability
  - mADE, mFDE
  - Cumulative probability

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Linear * [1]	1.33 / 2.94	0.39 / 0.72	0.82 / 1.59	0.62 / 1.21	0.77 / 1.48	0.79 / 1.59
SR-LSTM-2 * [30]	0.63 / 1.25	0.37 / 0.74	0.51 / 1.10	0.41 / 0.90	0.32 / 0.70	0.45 / 0.94
S-LSTM [1]	1.09 / 2.35	0.79 / 1.76	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.72 / 1.54
S-GAN-P [6]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
SoPhie [23]	0.70 / 1.43	0.76 / 1.67	0.54 / 1.24	0.30 / 0.63	0.38 / 0.78	0.54 / 1.15
						0.97
						1.00
						0.99
						1.07
						1.00
Social-STGCNN	0.64 / **1.11**	0.49 / 0.85	**0.44 / 0.79**	0.34 / **0.53**	0.30 / **0.48**	**0.44 / 0.75**

020)

Performance is reaching saturation, but have we solved the real problem?

						Ave.
						9 / 1.59
S-LSTM[7]						2 / 1.54
SGAN(20VP20)[8]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
STSGN[17]	0.75 / 1.63	0.63 / 1.01	0.48 / 1.08	**0.30** / 0.65	**0.26** / 0.57	0.48 / 0.99
S-BiGAT[14]	0.69 / 1.29	0.49 / 1.01	0.55 / 1.32	**0.30** / 0.62	0.36 / 0.75	0.48 / 1.00
S-STGCNN[12]	0.64 / **1.11**	0.49 / 0.85	**0.44 / 0.79**	0.34 / **0.53**	0.30 / **0.48**	0.44 / **0.75**
Social-PEC	**0.61 / 1.11**	**0.31 / 0.52**	0.47 / 0.82	0.43 / 0.77	0.35 / 0.60	**0.43** / 0.76

Table 1. ADE/FDE from (Zhao & Oh, 2020)

Target problem

Prediction vs. Navigation vs. Social Navigation

	Goal-oriented objectives	Social objectives	Environmental context / physical constraints
Static Navigation	Yes	Safety	Traversability, static obstacles
Trajectory prediction	No	Naturalness	Dynamic obstacles
Social Navigation	Yes	Safety / Norm / Comfort / Naturalness	Static + Dynamic obstacles

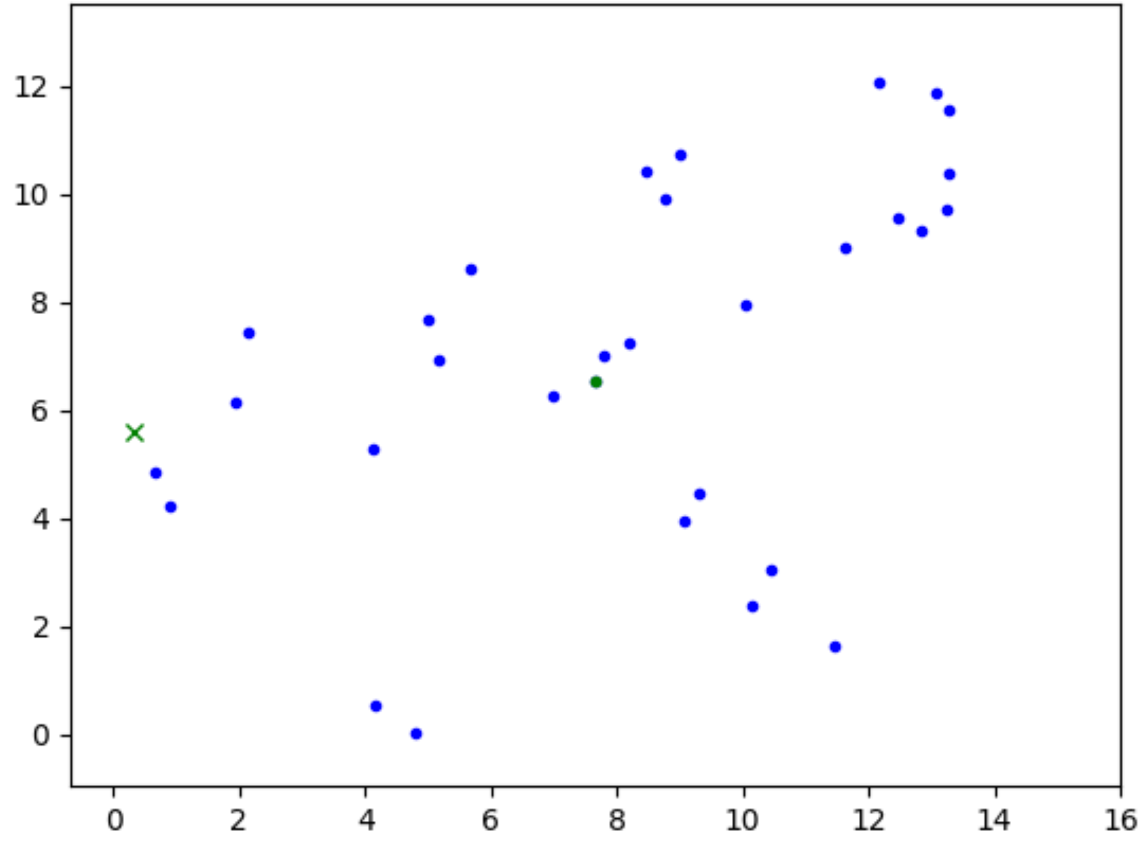
Social Navigation

	Social Objectives
Reinforcement learning (Berg et al., 2011, Chen et al., 2019)	Safety / Comfort
Inverse reinforcement learning (Vasquez et al., 2014)	Naturalness
Generative approach (Tsai & Oh, 2020)	Safety / Comfort Naturalness

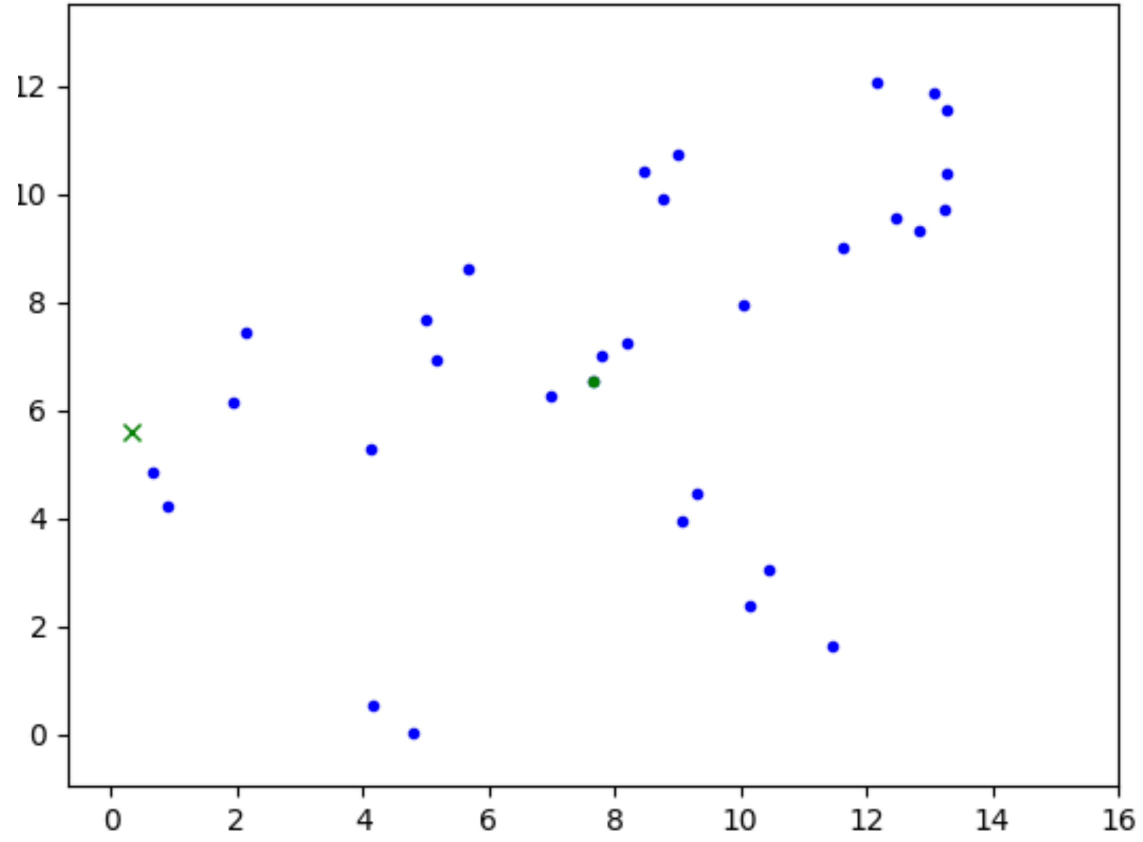
Interpretability / explainability

: observed trajectory
* : predicted trajectory
x : goal point
: pedestrians

Intention Only (f_{intent})



Social Aware ($f_{intent} + f_{social}$)



Evaluation is challenging

	G-S-LSTM	NaviGAN-R	human
S-score	0.40	0.38	0.44
Comfort%	81%	**97%**	96%
Arrival%	91%	85%	100%

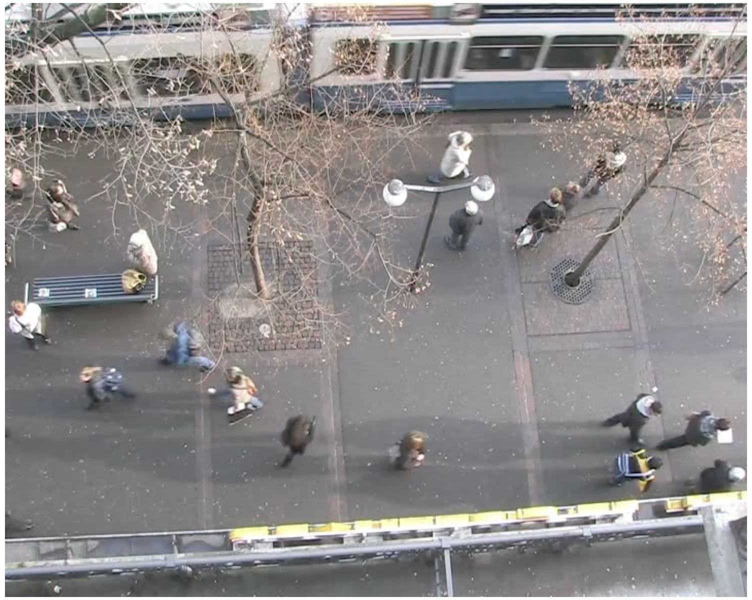
How can we improve evaluation?

$$r_t = \begin{cases} \text{[illegible]} & \\ -0.1 + \dfrac{d_o}{2} & , if\ d_o \leq \boxed{0.2} \\ 1 & , if\ d_g \leq \boxed{0.5} \\ 0 & , otherwise \end{cases}$$

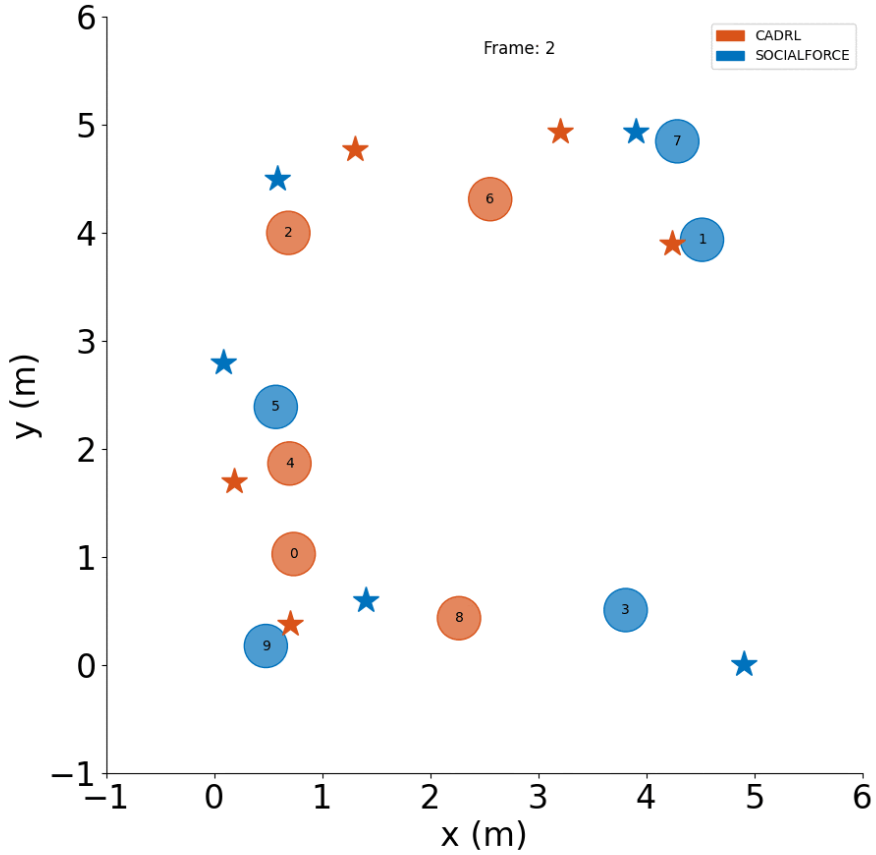
Social score (S-score) [Chen et al., 2019]

T.-E. Tsai and J. Oh A Generative Approach for Socially Compliant Navigation, In: IEEE Conference on Robotics and Automation (ICRA). 2020.

Can we simulate human pedestrians and generate edge cases? (ongoing work)



ETH Hotel dataset



Baselines: Statistics from human data

Metric	ETH	HOTEL	UNIV	UNIV	UNIV	ZARA1	ZARA2
Mean Population Density	0.3126444648	0.3123697026	0.3048187304	0.4764722944	0.4112925254	0.3537553013	0.3835216581
Total Collisions	44	142	12	4007	1150	54	809
Mean Agent Speeds (m/s)	2.3802701	1.158193835	1.469599044	0.7261555953	0.7913439566	1.157796617	1.136960778
Mean Agent Accel (m/s^2)	0.5525833067	0.2738044999	0.1150437337	0.1048869258	0.1405354981	0.07878554306	0.0858139069
Mean agent Jerk (m/s^3)	1.051525397	0.5071377015	0.4104586122	0.3827528888	0.5144874065	0.2963972969	0.3065193646
Mean Agent Energy (m^2/s^2)	0.09254848463	0.1157354273	0.007811867109	0.0435155658	0.06333488574	0.01932419826	0.02004629748
Mean Time Present	5.625555556	5.800514139	8.911864407	20.09542169	16.10414747	13.52702703	18.6627451
Mean Extra Time to Goal	0.3197769019	0.8302389987	0.6435179882	3.638110707	3.672329463	0.8873987434	1.636151288
Mean Path Ef	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.9483896619
Mean Closest	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.408500446
Mean Furthes	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	14.52586934
Mean Path Irr	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.2264788333
Mean Social	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.03590621796
% of Agents Average Speeds Slower than 0.5 m/s	4.722222222	21.59383033	1.694915254	31.80722892	13.82488479	1.351351351	8.823529412
Mean % of Time Congested below 0.5 m/s	6.066721466	21.61246192	2.847147994	34.11464698	22.18143786	4.029115252	9.09340225
Mean Frechet Distance	0.83	0.35	1.645091595	1.62	1.52	0.97	1.13
Mean Dynamic Time Warping	8.06	3.16	20.7926389	63.91	45.89	20.68	41.67
Mean Personal Space	77.42155907	67.71609848	50.94757022	18.45461162	32.71878798	80.85441768	37.01399992
			examples	students 001	students003		

1. Statistics vary across different datasets

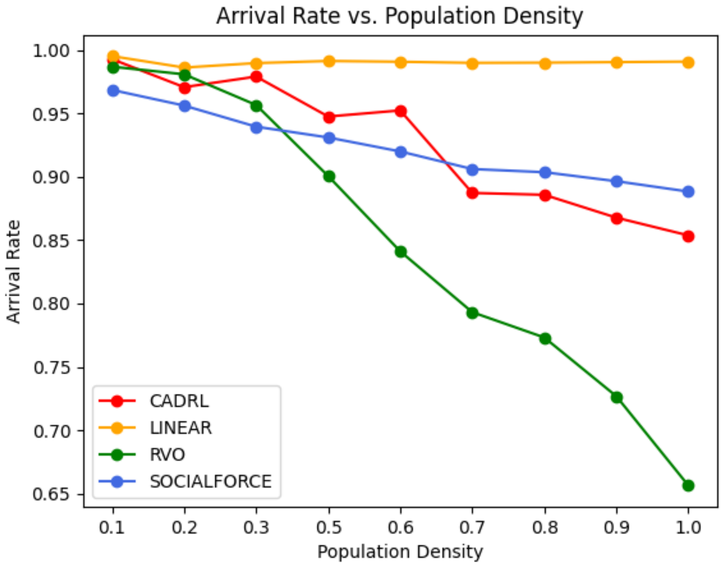
Evaluation of the algorithms in self-play setting (example: UNIV dataset)

Metric	UNIV	CADRL	LINEAR	RVO	SOCIALFORCE	CVMCVM	SLSLSTM	SOCIALGAN	SPEC	STGCNN
Mean Population Density	0.3975278501	0.3472515248	0.3509337397	**0.3847244492**	0.3692341188	0.3622348168	0.3556616657	0.339333059	0.3325776914	0.3294938716
Total Collisions	1723	6	1468	2204.666667	416.3333333	1695.333333	**1697.333333**	1267.666667	1516.666667	1149
Mean Agent Speeds (m/s)	0.995699532	0.6723677736	0.6829691502	0.5006556012	0.4072565826	0.6830343372	0.6827701575	**0.6830438697**	0.6829667187	0.6828380588
Mean Agent Accel (m/s^2)	0.1201553859	0.612437318	0.6092288804	0.5840025709	0.6400772189	0.6172825503	0.5698783774	**0.5405915469**	0.5860519668	0.5868453704
Mean agent Jerk (m/s^3)	0.4358996359	4.113602076	**3.724377352**	3.803070873	4.788470119	3.771397745	3.930844878	4.245000709	3.907776868	4.160032972
Mean Agent Energy (m^2/s^2)	0.03822077288	0.1100470263	0.0843338024	0.1846045926	0.2654254698	-0.03826908955	-0.01105743823	0.03078228072	0.03219459618	**0.03623843857**
Mean Time Present	15.03714452	12.38890477	11.44557957	18.7640627	20.67558483	11.25442614	**12.99689523**	18.15877684	18.77065965	17.46653853
Mean Extra Time to Goal	2.317986053	1.038723276	0.02159775092	2.947601783	1.834487335	0.02126524006	1.801867991	1.555981342	**2.145439465**	1.875514367
Mean Path Efficiency	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**0.8888192605**	0.9114520544
Mean Closest Proxim	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**1.510674469**	1.4974031
Mean Furthest Proxi	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	20.11347914	20.29266275
Mean Path Irregularit	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7.122485047	2.737197199
Mean Social Score @	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.5479867549	-0.3772797719
% of Agents Average	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
Mean % of Time Con	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.5958214641	0.5859886833
Mean Frechet Distance	1.34338933	1.23246997	0.1398323245	0.8646358617	0.567550283	0.1410523182	0.2525811906	1.179284761	0.7659049205	**1.309322665**
Mean Dynamic Time Warping	39.99682084	135.7402589	7.999784067	131.1768809	75.67122213	7.871672719	**14.58666894**	151.976061	92.6619339	165.0198372
Mean Personal Space	50.94757022									

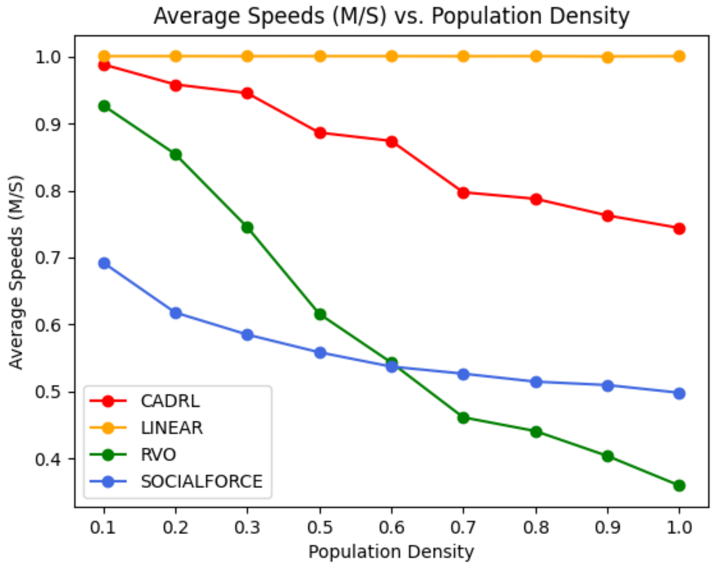
2. No single algorithm is a clear winner

- Green indicates the algorithm closest to the human data

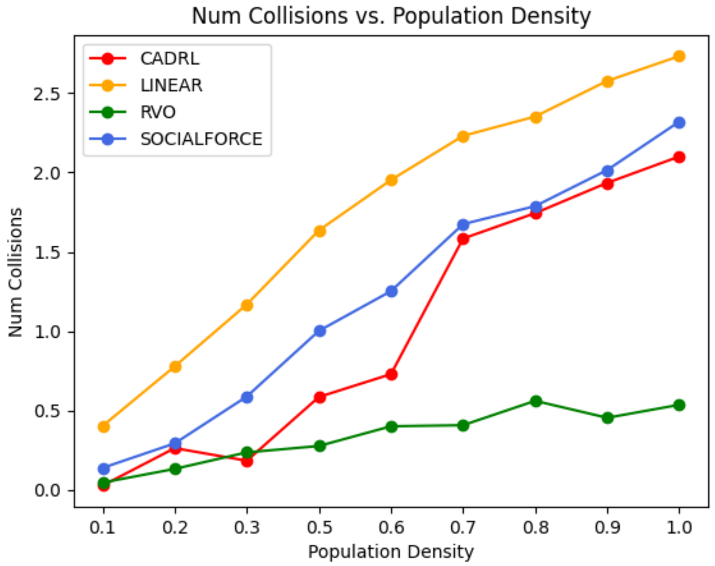
3. In self-play, algorithms exhibit unique trends



Arrival rate



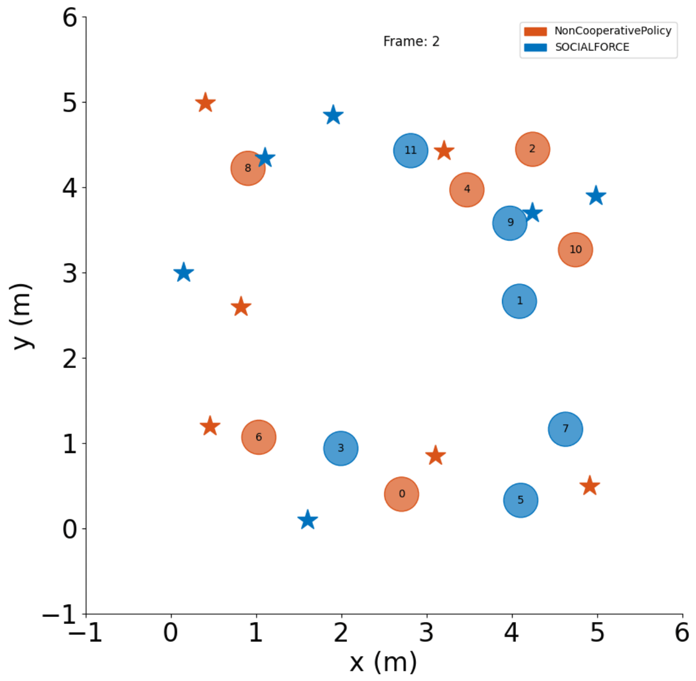
Arrival speed



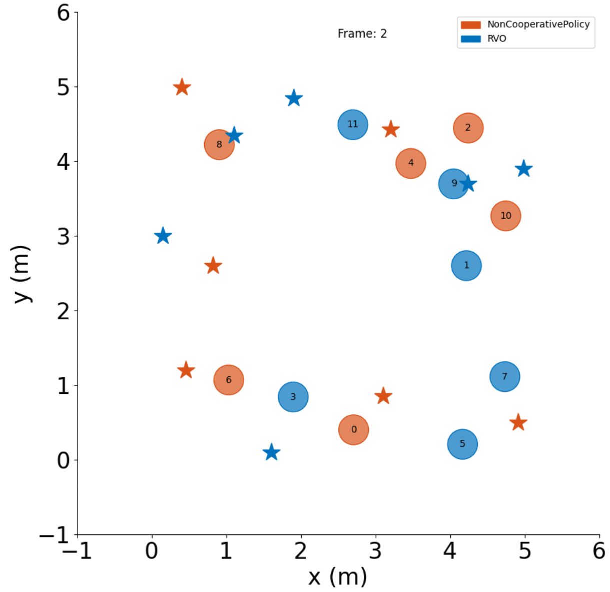
Number of collisions

4. In tournament, algorithms exploit/get exploited

Noncooperative vs Social force



Noncooperative vs RVO

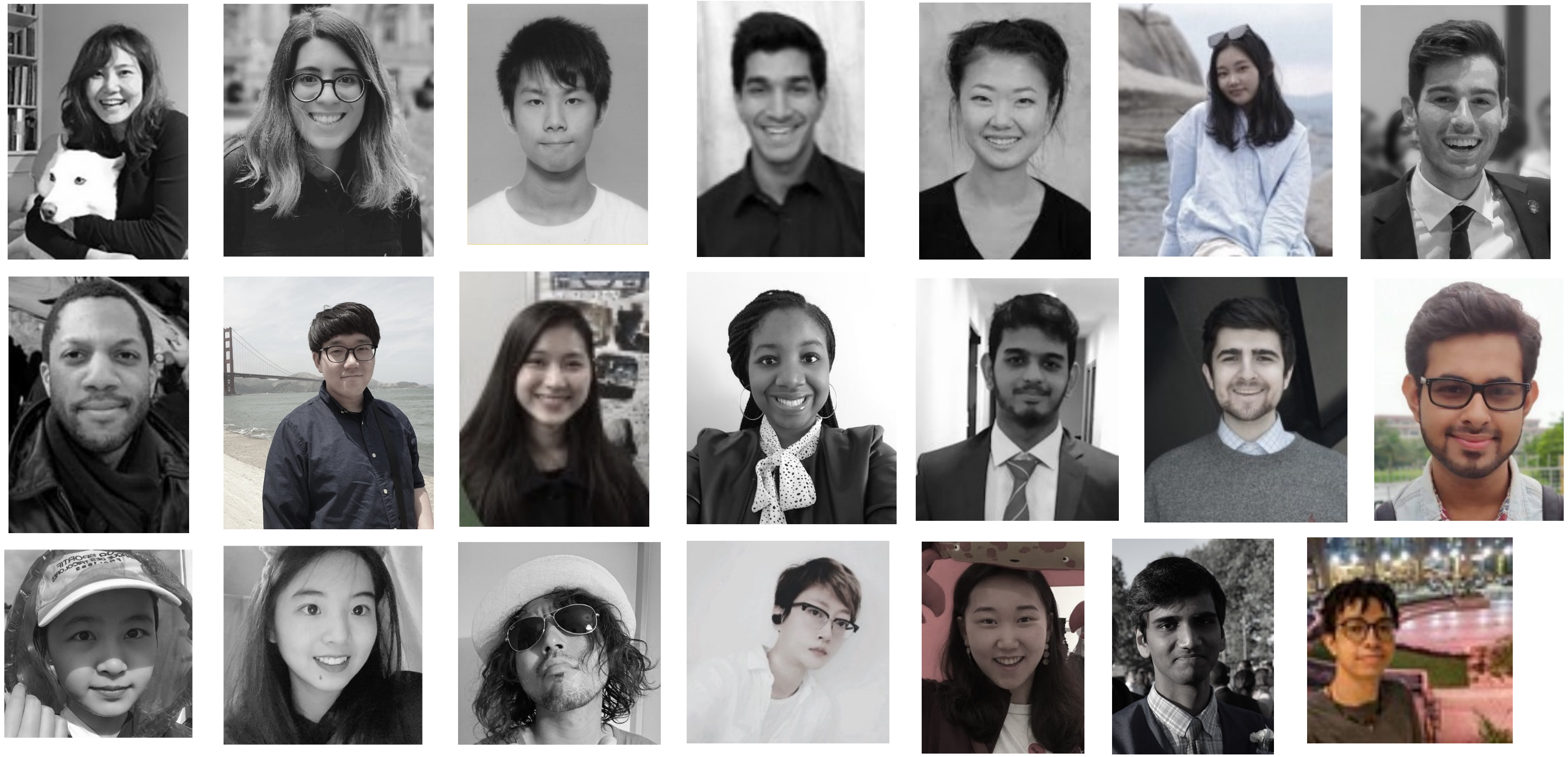


Future directions on social robot navigation

- Statistics vary across different datasets. Can we define a similarity metric using statistical properties?
- No single algorithm is a clear winner. Can we generate an agent population that is statistically similar?
- In self-play, algorithms exhibit unique trends. Are algorithms including enough variance?
- In tournament, algorithms exploit/get exploited. Can we design high-level strategy of mixture algorithm in a repeated game setting?

roBot Intelligence Group (BIG)

Jean Oh



1 postdocs
7 PhD students
5 MS students
4 undergrads
2 visiting researchers

9 Countries
50:50 female:male ratio

Ingrid Navarro, Sam Shum, Tanmay Shankar, Xuning Yang, Emily Byun, Peter Schaldenbrand, Jonathan Francis, Meghdeep Jana, Jimin Sun, Abby Yao, Nariaki Kitamura, Ben Stoler, Mayank Mali, Zhixuan Liu, Shaunak Halbe, Zhanxin Wu, Beverley-Claire Okogwu, Yuning Wu, Soonmin Hwang, Almutwakel Hassan

Thank you for your attention. Questions?

Mavrogiannis, C., Baldini, F., Wang, A., Zhao, D., Steinfeld, A., Trautman, P., & Oh, J. (2021). "Core Challenges of Social Robot Navigation: A Survey." arXiv preprint arXiv:2103.05668.